Implementation of task-oriented Spoken Dialog Systems (SDS) in the Italian language in the scenario: Cinema Ticket/Info Robot

Vannoli Marco matr.1860363*
Master's degree in artificial intelligence and robotics Department of computer engineering, automatic e management "Antonio Ruberti" University of Sapienza Roma
(Dated: February 19, 2020)

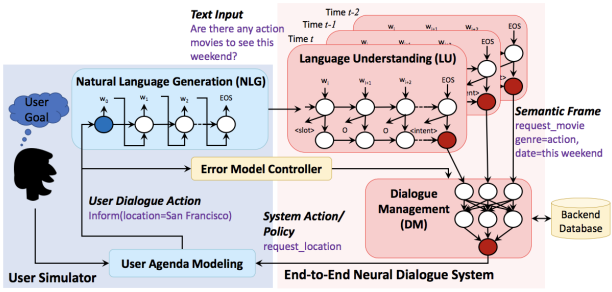
1. INTRODUCTION

The purpose of the paper presented here is to implement a system based on SDS systems on a well-defined scenario: the Cinema Ticket Robot (CTS agent). The Spoken Dialog Systems (SDS) are systems that involve speaking with a user in order to accompany them towards a specific goal, thus helping them to achieve a certain purpose. In fact, they are named "Task-Oriented" because they plan to direct the user towards a specific objective. In this case, as far as the chosen scenario is concerned, the aim will be to help the user towards the purchase of one or more tickets for a certain film in any cinema in Italy. Obviously, this request will include a phase in which the user will respond to the robot in such a way as to provide him with the information necessary to proceed with the purchase or booking. Thus the language chosen for this project is Italian and this brought to me the need to translate sentences from English to Italian in order to use some systems which work very well in the English domain. [1]

2. THE TASK-ORIENTED SPOKEN DIALOG SYSTEMS

A spoken dialogue systems SDS is designed of a set of modules which each of them perform different part of the process and brings to information needed to reach final goal. [1][2]

FIG. 1. **The architecture of Task-oriented SDS**



The first block that allows us to recognize the voice of the user is the automatic speech recognition (ASR). For this project, a library called speech-recognize for python has been used to recognize the user's voice and working it in the form of text. Returning to the SDS, it is a system composed of blocks each need to collect information together in order to made speak robot correct way: How the figure above shows, the SDS system is composed by three sub-task it to reach the entire purpose:

- **spoken language understanding**: sub-topic of natural-language processing in artificial intelligence that deals with machine reading comprehension [4]. It is composed, how it will be shown in the following sections, by two entity which are:
  - **Intent Detection** : can be seen as a text classification task in which some sentences relative to own domain are taken to predict an intent that the user can intend. In fact, for this elaborate, 11 intents chosen self have been selected in order to perform an intent classifier. These intents have been chose to have a dialogue with robot both to buy movie tickets and to get information about the films, on actors, on rating of them (stars) and the showing of them at various times of the year. (data created randomly without using a database). In the next sections, the structure of the model will show and relative details will be explained. [3][5]
  - **Information Extraction with slot filling** : the extraction of information means the phase in which the necessary information is extrapolated immediately from the sentence in conjunction with the intent. This phase sees the recognition of the entities that concerns a subproblem related to natural language processing i.e the named-entity recognizer. For example, from the Italian sentence "Vorrei sapere quali film di genere poliziesco verranno proiettati questa settimana" the information about of film genre is extrapolated, in this case, "poliziesco" ("detective gere") and also the time frame i.e "questa settimana" (this week). This phase is also called slot filling because the fields that correspond to the essential informa-

* vannolimarco@hotmail.it

tion to then reach the goal are automatically filled by the only first request that the user makes. [6]

- **dialogue management**: The dialogue management is the part where the intent and slot information is incorporated to extrapolate the action or phrase that the robot must say to respond to the user's initial request. Such answers, if the slots concerning the corresponding intent are all full then the robot will answer with the final answer otherwise it will go ahead with questions to the user to fill the missing slots. For example, from the phrase "I want to buy two tickets for the cinema tomorrow evening" the robot acquires the intention of wanting to buy two tickets for the cinema for tomorrow night and will ask questions regarding the title of the film, the exact time and the name of the cinema with the purpose to fill the slots of the missing fields and then proceed with the purchase. The state that therefore acquires the DM is the intent and from the latter acquires the fields related to this intent:

  For example, a state of belief could be:
  sentence = *"Vorrei acquistare due biglietti per un film per domani" ("I would like to buy two tickets for a movie tomorrow")*
  intent = request.booking_film
  slot_tickets = *due* (two)
  slot_date = *domani (tomorrow)*
  slot_film = ?
  slot_time = ?
  slot_location = ?

  The robot will then start asking the user questions to obtain the missing information i.e *"Quale film vuoi andare a vedere?" (Which movie do you want to go to?), "A che ora?" (at what time?).*

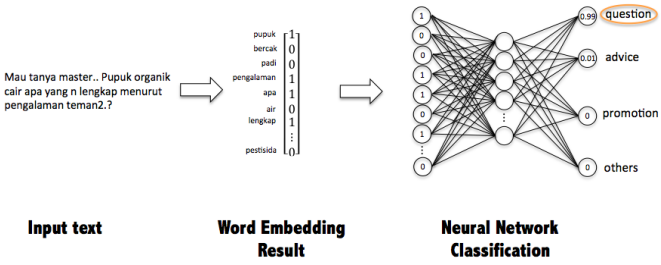
- **poken language generation**: This block aims to convey or requesting information in natural spoken language. The understanding part is usually performed in two steps:
  - The natural language sentences are generated from the systems internal data representation. This has different levels of complexity, from simple template-based approaches to complex planning systems that dynamically decide what information to actually give, in which order and with which words. Due to our fairly limited target domain i.e movie tickets/info, we used the simpler approach and hand-wrote the sentence that robot must say.
  - The second step is speech synthesis which is the artificial production of human speech. It transforms the generated sentence into a spoken utterance understandable to the user, and a text-to-speech system has been used in order to perform it. As text-to-speech, the library of **pyttsx3** provided for python has been used.

The next section will show how the structure for intent classifier and the slot filling has been build and the results about accuracy and loss will be shown.

3. INTENT CLASSIFIER

The intent classifier that has been implemented follows the line of a text classification problem, as the figure 2 show. [3][5]

FIG. 2. **Intent classifier process**



The targets that have been chosen for this project fall within the domain of cinema ticket sales and information and are shown in the table below:

Intents considered as labels	
Intents	Description
request.booking_film	Purchase request for cinema tickets
request.congratulations	Congratulations to the robot
request.help	User request for help
request.info_film	Request for information on a film
request.info_film_advi	Request to provide advice on films
request.info_film_gen	Request for information on a genre of a film
request.info_film_max_score	Request for information on the film with the highest rating
request.info_film_max_score_actor	Request for information on the film that saw the most awarded actor (oscar or not)
request.info_film_score	Request for information on the score of a film
request.info_films	Request for information on film programming
request.info_robot	Robot information request (purpose)

From a sentence, the robot is able to predict the intent of the user and continue to do in according to it. These intents have been chosen in order to aim the purchase of a ticket but also to provide information regarding the film. Obviously, the absence of a database doesn't allow us to provide true data but in a random way and therefore untruthful data with the truth.

A. data and preprocessing

The data file which allowed to perform the intent classifier is a CSV file produced by self in which are all sentence needed with own intent. The preprocessing has been executed as a simple text classifier or the sentences have been elaborated in this way:

1. All punctuation has been removed from each sentence,
2. Each word has been made into lower case and taken the corresponding lemma,
3. All the terms have been tokenized and vocabulary has been created with the words/lemma of the file,
4. Every single word in the sentences has been replaced with its value in the vocabulary,
5. Each tokenized sentence was padded with the maximum length of the longest sentence in the file,
6. The created a vocabulary of intent and rendered in one-hot encoding,
7. The dataset obtained was divided into training data and test data with 33% for testing data.

B. the model and training

The model of neural network is composed by one embedding layer with relative length of vocabulary, one Bidirectional LSTM layer, two layers composed by one dense layer with 'relu' function as activation function and one dropout layer, one BatchNormalization layer to normalize the activations of the previous layer at each batch and finally one dense layer with one neuron for intent label (11) and a softmax activation function. The model was trained with 200-500 epochs using Adam as optimizer. The hyperparameters which allowed to obtain good results about accuracy and loss are:

Hyperparameters set in the model	
Hyperparameter	Description
learning rate	0,001
dropout	0,10
batch size	16

C. results

The results about accuracy and loss are shown in the bottom figure (3,4):

FIG. 3. **Trend of accuracy train/test**

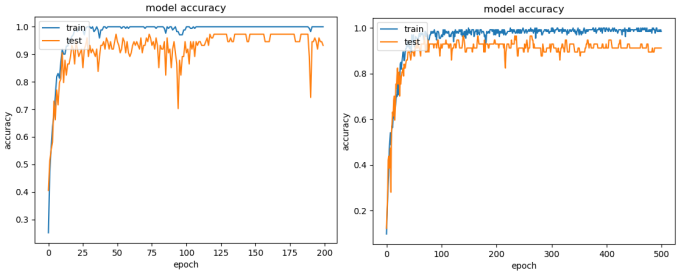
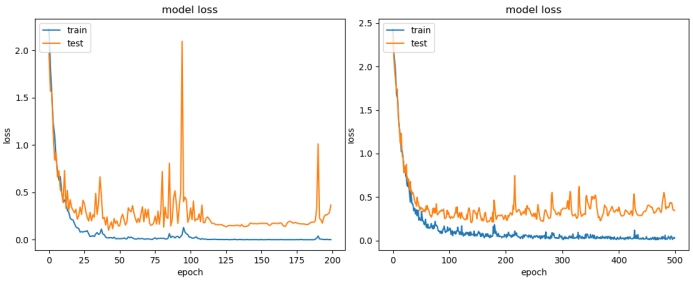


FIG. 4. **Trend of loss function train/test**



The accuracy for testing data is around 90-96% of accuracy and represents a good results. Of course, if you enhance the dataset by adding more sentences necessary to contextualize the intent for the corresponding domain, the robot will be more able to better understand the user's sentences (make the dataset larger).

4. SLOT FILLING

Python's spacy library was used for the slot filling process, which allowed the recognition of entity such as people, dates, times and other types of entities. This issue is called named-entity recognizer and represent a sub-task of natural language preprocessing. This operation involved a phase of translation of the sentences from Italian to English and then phase of retranslated from English them once the needed entities were obtained. The latter operation was carried out because spacy with English dataset (en_core_web_sm) is more functional now than the Italian one. Then a logic was inserted that allows to extrapolate the useful entity from the user's sentences ('Date', 'Time', 'Cardinal' etc ..) and then fill in the essential slots for the understanding process. A recognition of the genres has also been introduced to recognize the types of genres of the films and therefore to obtain information regarding the genres immediately. [6]

5. TESTING WITH ROBOT

The tests that have been carried out with the robot have included conversations touching all the intents and slots that have been taken into consideration by the implemented SDS system. The robot predicts that once he answers the user correctly, he starts the conversation again by asking the same if he needs to know more. Certainly, it is important to specify that the information provided by robots is random due to an absence of databases in which to be able to take the truthful information (in this case on the titles of the films, on the projection of cinemas in Italy and on the names of cinemas in Italy etc ..). The system will close if the user calls a specific word or "Chiudi" ("Close"). The example below presents a tested co-conversion type:

Gina : *Ciao, io sono Gina e ti aiuterò a prenotare dei biglietti per qualsiasi film in qualsiasi cinema in italia e avere informazioni generali sui film, come ti posso aiutare?*
You : *Vorrei acquistare due biglietti per un film per domani*
Gina : *Quale film vuoi andare a vedere?*
You : *Joker*
Gina : *In quale cinema?*
Gina : *Ripeti il nome del cinema...*
You : *Uci cinema Roma est*
Gina : *A che ora vuoi vedere il film? Orari disponibili : 10:00 , 11.30 , 12:00 , 19:00 , 21:30*
You : *alle 21:30*
Gina : *Procedo nell'acquisto con le seguenti informazioni che mi hai fornito :*
Film : Joker
Alle ore : 21:30
Il giorno : domani
Nel cinema : Uci cinema Roma est
INFO DEI BIGLIETTI:
Numero biglietti : Due
Sala : T
Posti : fila G, dal numero 13 scorrendo verso destra
Confermi?
You : *Sì*
Gina : *Grazie per la prenotazione.*
Gina : *Come ti posso aiutare ancora?*
You : *Desidero sapere che film ci sono questa settimana*
Gina : *li vuoi filtrare per genere?*
You : *Si*
Gina : *Per quale genere li vuoi filtrare? ti do dei consigli : Animazione, Poliziesco, Thriller ...*
You : *Animazione*
Gina : *I film con il genere animazione che sono proiettati questa settimana sono i seguenti: Il cavaliere oscuro, Ritorno al futuro*
Gina : *Come ti posso aiutare ancora?*
You : *Chiudi*

Gina is the name chosen for the robot because Gina was one of the most important female figures in Italian cinema. Of course, the phrases used for the intent have been formulated in such a way as to cover almost all possible formulations of the user's sentence but the robot might understand something different from the initial intent (it depends on the data with which it was trained the model).

6. CONCLUSION AND FINAL CONSIDERATIONS

In conclusion, the results show a good functionality of the system that manages to recognize the user's intent in the domain considered for this project. The project can be improved by taking into consideration more data for the model and implementing other systems that would give more accuracy in the recognition of the entities and therefore in the slot filling for the chosen domain. For future work, there is the desire to implement a network of systems capable of working at best with the Italian language and increasing precision in the cognitive field of the robot.

7. REFERENCE

[1] *Using Task-Oriented Spoken Dialogue Systems for Language Learning: Potential, Practical Applications and Challenges* Antoine Raux, Maxine Eskenazi Language Technologies Institute Carnegie Mellon University 5000 Forbes Avenue 15213 Pittsburgh, USA

[2] *Spoken Dialogue Systems* Jokinen, Kristiina,Mctear, Michael Google Deepmind London, UK paper at ICLR 2016

[3] *Intent Classification of Short-Text on Social Media* Purohit, Hemant,Dong, Guozhu,Shalin, Valerie,Thirunarayan, Krishnaprasad 2015/12/01 10.1109/SmartCity.2015.75

[4] *Natural Language Understanding* Sowa, John 2015/12/14 10.13140/RG.2.1.3409.6726

[5] *Utterance Intent Classification of a Spoken Dialogue System with Efficiently Untied Recursive Autoencoders* Tsuneo Kato,Ryosuke Sumitomo,Atsushi Nagai,Jianming Wu,Naoki Noda, Seiichi Yamamoto August 2017 Saarbrücken, Germany Association for Computational Linguistics

[6] *Stanford's Distantly-Supervised Slot-Filling System* Mihai Surdeanu, Sonal Gupta, John Bauer, David McClosky, Angel X. Chang, Valentin I. Spitkovsky, Christopher D. Manning Computer Science Department, Stanford University, Stanford, CA 94305