Anonymous ACL submission

## Abstract

This document contains the report of the second project of the course of Language Understanding System. The goal is to develop a simple dialog system within Rasa framework in the movie domain, which is able to interact with the user answering to pertinent questions (*e.g.* director names, actors name, release year etc etc). All necessary steps will be explained, starting from the pre-processing of the data to fit the requests of Rasa, to the configuration of the database and the query system to access information. The preferred language for the both was Python (v2.7), whereas the RDBMS for the database was MySql.

## 1 Introduction

It has been almost a decade since the trend of bots' development had a sharp increase and many products were sold on the market. Even though many of them had a lot of success, one of the main issue that many both were not able to face was to recognize the context and this derives from the fact that many of them used state machine to solve this task. With the spreading of new techniques, such as machine learning algorithms, a lot of research has been done on this issue, but has not translated into actual developers tool.

Rasa is a framework which provides a new approach to conversational softwares: instead of taking hard-coded rules it exploits the fact that if on one hand understanding when the bot is wrong is easy, on the other hand understanding*why* it is wrong can be very tricky. Following this way, it is possible to decide everything the bot can do or say, even the training can be done either in a supervised way (if data are available) or with an interactive learning starting from scratch.

## 2 Data Analysis

The provided dataset is the same of the previous project, namely the Microsoft NL-SPARQL dataset and it was already split into two parts: one for training and one for testing. In addition, each of these groups of data was divided into main data for the Movie Domain and additional features. The latter contains labels for each sentence.

To summarize, what was given were five files:

- **NL-SPARQL.train.data** containing a two columns set of data, tab separated, where the first column represent the words and the second tags,
- **NLSPARQL.train.utt.labels.txt** containing a single column set of data, identifying the labels for each sentence,
- **NL-SPARQL.test.data** containing a two columns set of data, tab separated, where the first column represent words and the second tags,
- **NLSPARQL.test.utt.labels.txt** containing a single column set of data, identifying the labels for each sentence,
- **moviedb.sql**: the database of the movie domain