Language Understanding Systems — Final project - Dialog System within rasa framework in movie domain

Anonymous ACL submission

1 Introduction

The objective of the final project is to create a dialogue system with *rasa* framework in movie domain. In short words, create a question answering bot about movies. In addition to that, the bot is connected to a speech service in order to get questions from the users' voice and to give answers through a speech synthesizer.

2 Rasa data and files

Rasa needs different files for training:

- domain information
- stories for dialogue training
- data for NLU training

2.1 From NL-SPARQL dataset to NLU Data Format

The first file created for the project was the one containing data for the NLU training. The data format for NLU training is a JSON file containing the text, the intent and the entities, if there are any. Hence, the starting dataset, which is the same of the previous project, needed to be modified and to do so it was created a script in python that makes the conversion.

2.1.1 Entities

An **entity**, in the rasa format, is a set of one or more words referable to a specific concept. The concept is the entity and the value of the entity is the set of words. The original dataset is formed by many sentences. Each sentence is divided by words and each word is in one line with the related IOB-tag. The IOB-tags may start with one of these letters: I, O or B. ”O” stands for ”Outside the span”, while ”B” is ”Beginning of span” and ”I” is ”Inside of span”. The ”O”-ones are not important for this task, because they don't carry any important information for the classifier. ”B” and ”I” tags, though, indicate that the related word has an important meaning. For instance, the words refer to a movie or to an actor. The script collect all the words until the ”\n”, which signals the end of the sentence, and put all in the right place:

- the sentence in the text field
- the entities in the entities list

For each entity it computes the start and the end of the entity values.

2.1.2 Intents

The **intent** describes what your user probably meant to say.