IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

NLP and Text Mining
Chatbot project

Students:
Johan MEJIA
Diego CARREÑO
Tatiana MORENO

RESUME

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

https://leverageedu.com/blog/study-abroad/

Looking for information about studying abroad is not an easy task among the reasons are that there is a lot of information available, too many options and it can be difficult to filter (countries, cities, costs, languages, etc.)

We propose a chatbot that serves as an assistant to find information about studying abroad in a simpler way

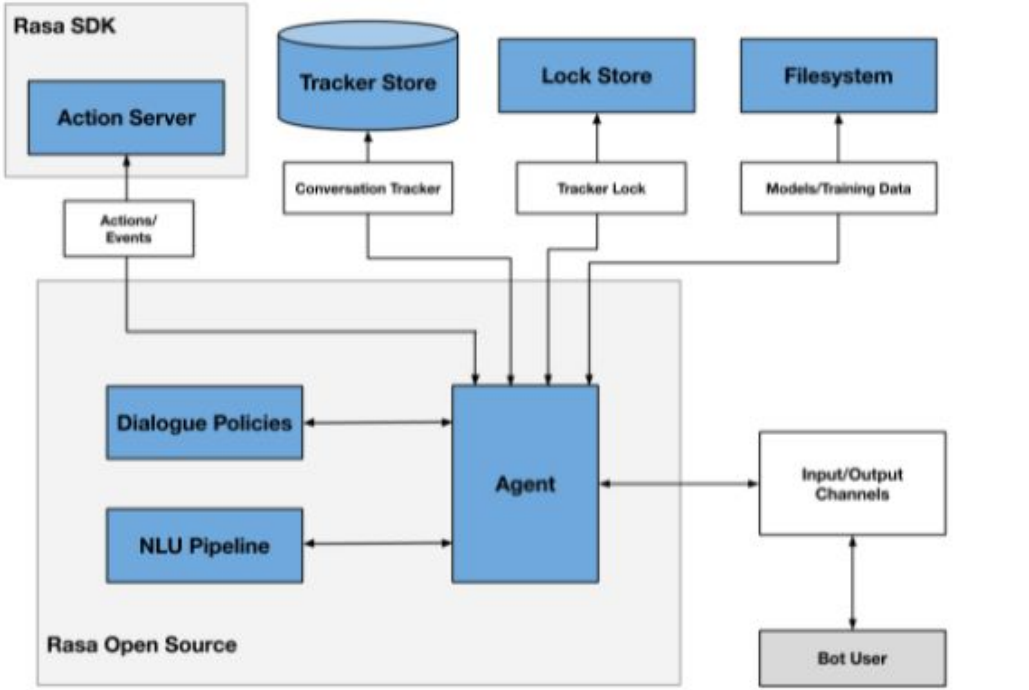
A database was created with information from universities in 6 countries and 15 cities.

Table 1. Countries and cities in database

Country	City
USA	New York
	California
France	Paris
	Brest
	Rennes
UK	London
	Edinburgh
	Manchester
Germany	Berlin
	Munich
Spain	Barcelona
	Madrid
	Valencia
Switzerland	Zürich
	Lausanne

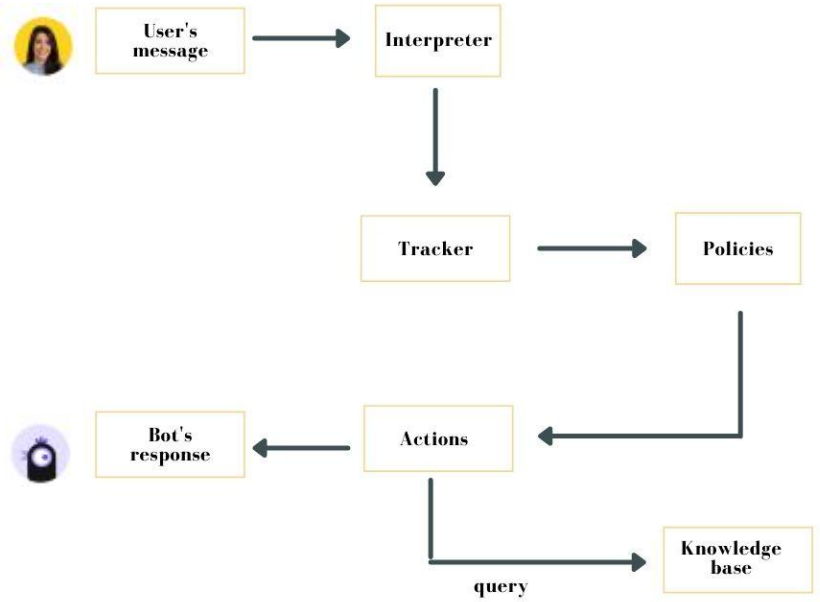
Table 2. Attributes of each university in the database

Attribute	Description
id	Id of university
name	Name of university
city	City where is located the university
careers	Careers of university
link	Link of university
ranking	Ranking of university according to https://www.topuniversities.com/
scholarship	If there is scholarships or finantial aids in the university
cost	Cost of university



Rasa is an open source machine learning framework for automated text and voice-based conversations. The two primary components are **Natural Language Understanding (NLU)** and **dialogue management.**

https://rasa.com/docs/rasa/arch-overview



Workflow of generated response in RASA Framework. It represent the complete process of end to end conversation between user and the chatbot.

A knowledge base is used to store complex data structures. The data stored in your knowledge base represent our domain knowledge.

Entity of interest (e.g. Sorbonne University)

Entity type of the entity interest (e.g. university)

Attribute of interest (e.g. fees)

Key attribute of the entity type (e.g. name)

2.4 Domain

The domain defines the universe in which the assistant operates. It specifies the intents, entities, slots, responses, forms, and actions the bot should know about.

```
intents:
  - greet
  - goodbye
  - affirm
  - deny
  - mood_great
  - mood_unhappy
  - bot_challenge
  - query_knowledge_base
  - recommendation
  - nicetomeetyou
  - thanks
  - telljoke
  - ask_howbuilt
  - ask_howdoing
  - ask_howold
  - ask_time
  - ask_weather
  - ask_whatspossible
  - rec_sports
  - rec_videos
  - rec_music
  - rec_cooking
  - fix_mispelling
```

```
entities:
  - wrong_words
  - correct_words
  - last_intent
  - object_type
  - city
  - cities
  - attribute
  - mention
  - ranking
  - name
  - USA
  - France
  - UK
  - Germany
  - Spain
  - Switzerland
```

```
responses:
  utter_greet:
  - text: "Clementine -> Hey! How are you? My name is Clementine."
  - text: "Clementine -> Hey there! My name is Clementine."
  - text: "Clementine -> Hi! I'm Clementine."
  - text: "Clementine -> Hello! I'm Clementine."
  - text: "Clementine -> Hey! My name is Clementine."

  utter_cheer_up:
  - text: "Clementine -> Here is something to cheer you up:"
    image: "https://i.imgur.com/nGF1K8f.jpg"
  - text: "Clementine -> Here is something to cheer you up:"
  - text: "Clementine -> Check this out, it could help:"
  - text: "Clementine -> By the moment, check this out:"
  - text: "Clementine -> While it gets better, check this out:"
  - text: "Clementine -> Let me try yo cheer you up with this:"
```

2.5 Configuration

In this part are defined the components used by the model to make NLU predictions, and the policies used by the model to predict the next action.

```yaml
language: en
pipeline:
  - name: WhitespaceTokenizer
  - name: RegexFeaturizer
  - name: LexicalSyntacticFeaturizer
  - name: CountVectorsFeaturizer
  - name: CountVectorsFeaturizer
    analyzer: char_wb
    min_ngram: 1
    max_ngram: 4
  - name: DIETClassifier
    epochs: 100
    constrain_similarities: true
    model_confidence: linear_norm
  - name: EntitySynonymMapper
  - name: ResponseSelector
    epochs: 100
    constrain_similarities: true
  - name: FallbackClassifier
    threshold: 0.3
    ambiguity_threshold: 0.1
    # model_confidence: linear_norm
  - name: "components.spelling_check.SpellingAnalyzer"
  # - name: EmbeddingIntentClassifier
```

```yaml
policies:
    - name: MemoizationPolicy
    - name: TEDPolicy
      max_history: 5
      epochs: 100
      constrain_similarities: true
    - name: RulePolicy
```

2.6 NLU Training data

NLU training data stores structured information about user messages. This information includes the user's intent and any entities their message contains.

```
- intent: query_knowledge_base
  examples: |
    - give me the list of [cities](object_type:cities)
    - i would like the list of universities in [USA](object_type) please
    - i dont know which universities are in [France](object_type)
    - which universities can i find in [New York](city)?
    - give me the list of universities in [Barcelona](city) [Spain](object_type)
    - list of universities in [Germany](object_type)
    - in which [city](attribute) is [EPFL](Switzerland)?
    - in which [city](attribute) is [this](mention)?
    - [IMT Atlantique](France) is in which [city](attribute)?
    - what [careers](attribute) are there in [University of Lausanne](Switzerland)?
    - what [careers](attribute) can i find in [this](mention) university?
    - list of [careers](attribute) in [five one](mention:5)
    - tell me all [careers](attribute) in [Universidad Autonoma de Madrid](Spain)
    - show [careers](attribute) in [King's College London](UK), please
    - what is the [link](attribute) of the [Technical University of Munich](Germany)?
    - give me the [link](attribute) of [UCLA](USA)
    - in the [second one](mention:2), which is the [link](attribute) ?
    - where is [its](mention) [address](attribute:link)?
    - where can I get [information](attribute:link) of [University of Manchester](UK)?
    - i would like to know [more information](attribute:link) about [Complutense University of Madrid](Spain)
    - please, [tell me more](attribute:link) about [ETH Zurich - Swiss Federal Institute of Technology](object_type:Switzerland)
    - is there any [information](attribute:link) of [University of Lausanne](object_type:Switzerland)?
    - do you know the [ranking](attribute) of [University of Edinburgh](object_type:UK)?
```

The stories are the training data used to train the assistant's dialogue management model

```
- story: query knowledgebase
  steps:
  - intent: query_knowledge_base
  - action: action_query_knowledge_base
```

```
- story: what's possible
  steps:
  - checkpoint: check_greet
  - checkpoint: check_howdoing
  - checkpoint: check nicetomeetyou
  - intent: ask_whatspossible
  - action: utter_ask_whatspossible
```

```
- story: Try to correct user's misspelling successfull
  steps:
  - intent: fix_mispelling
  - action: action_spelling_check
```

```
- story: Try to correct user's misspelling failed
  steps:
  - intent: fix_mispelling
  - action: action_spelling_check
  - intent: deny
  - action: utter_ask_rephrase
```

3. Demonstration

4. Conclusions

- The chatbot developed in this project could serve as a basis for the creation of assistants for companies in charge of providing information about studies abroad, allowing them to improve their performance by providing faster and easier information to the user

- Rasa is a very powerful tool for the creation of assistants, the framework includes pre-trained models which allow to perform a language processing which gives the bot the ability to analyze the context of a conversation and identify the user's requirements. Since rasa is open source, it is flexible to programming requirements, even allowing you to change default aspects such as answer phrases or intent classification.

- The definition and understanding of key concepts such as intentions, actions, entities, forms and slots are a crucial part of developing applications with Rasa. For a more agile development, each concept must be previously defined according to the application schema to be used.

5. For future work

- Build or find a larger database, which would allow more filtering of information such as by careers or types of degrees (masters, doctorate, bachelor)
- Train the chatbot to identify and chat in several languages.
- Modify the pipeline to allow the chatbot to identify more specific misspellings, and exclude non-English speaking names in the database, in order to improve conversational fluency.
- Connect the service to remote databases, as well as create interfaces on platforms such as Facebook or Slack.

Reference

- **Repo :** TatianaMoreno47/Chatbot (github.com)
- **Rasa framework :** Open source conversational AI | Rasa
- **Rasa tool used :** Knowledge Base Actions (rasa.com)
- **Reference Chatbot :** valeporti/imt_chatbot (github.com)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

NLP and Text Mining
Chatbot project

Students:
Johan MEJIA
Diego CARREÑO
Tatiana MORENO