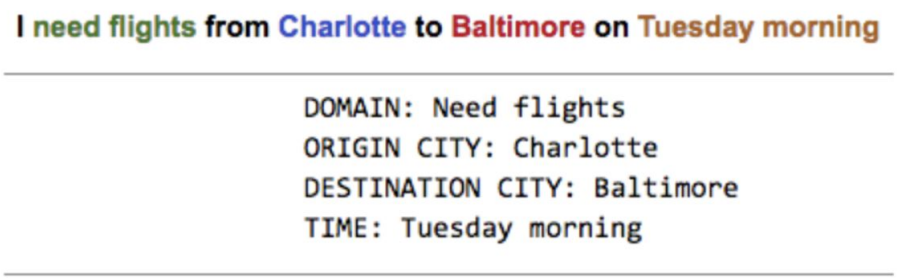
Implementation of a Spoken Language System

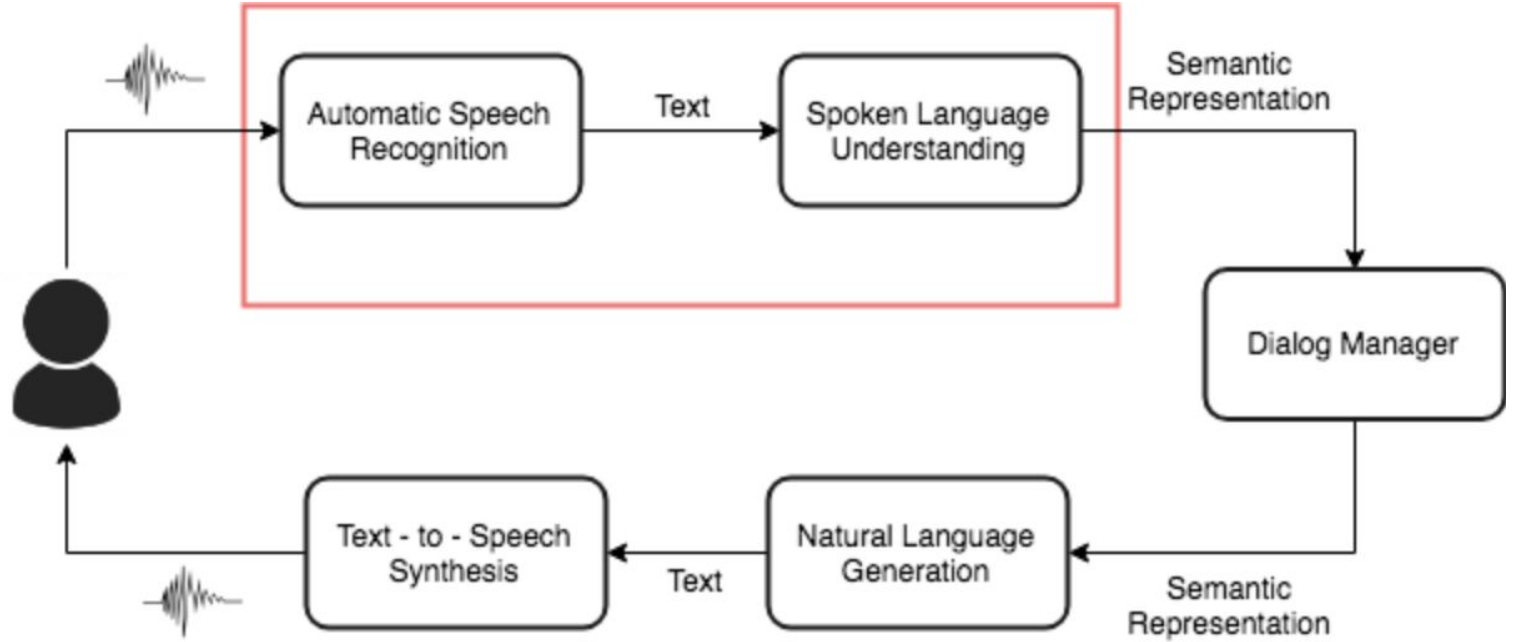
Trabajo Fin Grado
Área: Inteligencia Artificial
Estudiante: Jessica Pérez Guijarro
Tutor: David Isern Alarcón

Objetivos del proyecto

- Profundizar los conocimientos adquiridos a lo largo del grado en el área de AI, concretamente en la disciplina de Natural Language Processing.
- Desarrollar un sistema Spoken Language System que permita:
  - Reconocer las órdenes de voz de un usuario/a y transformarlas a texto → Módulo ASR
  - Generar una representación semántica del texto previamente obtenido. Dicha representación deberá identificar aquellos elementos de la frase que permitan responder a la orden enunciada por el usuario/a → Módulo SLU



Arquitectura del Spoken Language System



Stack de tecnologías y recursos empleados para el desarrollo del proyecto

ATIS

VoxForge

HTK - Toolkit

Continuous Speech Recognition Julius Engine

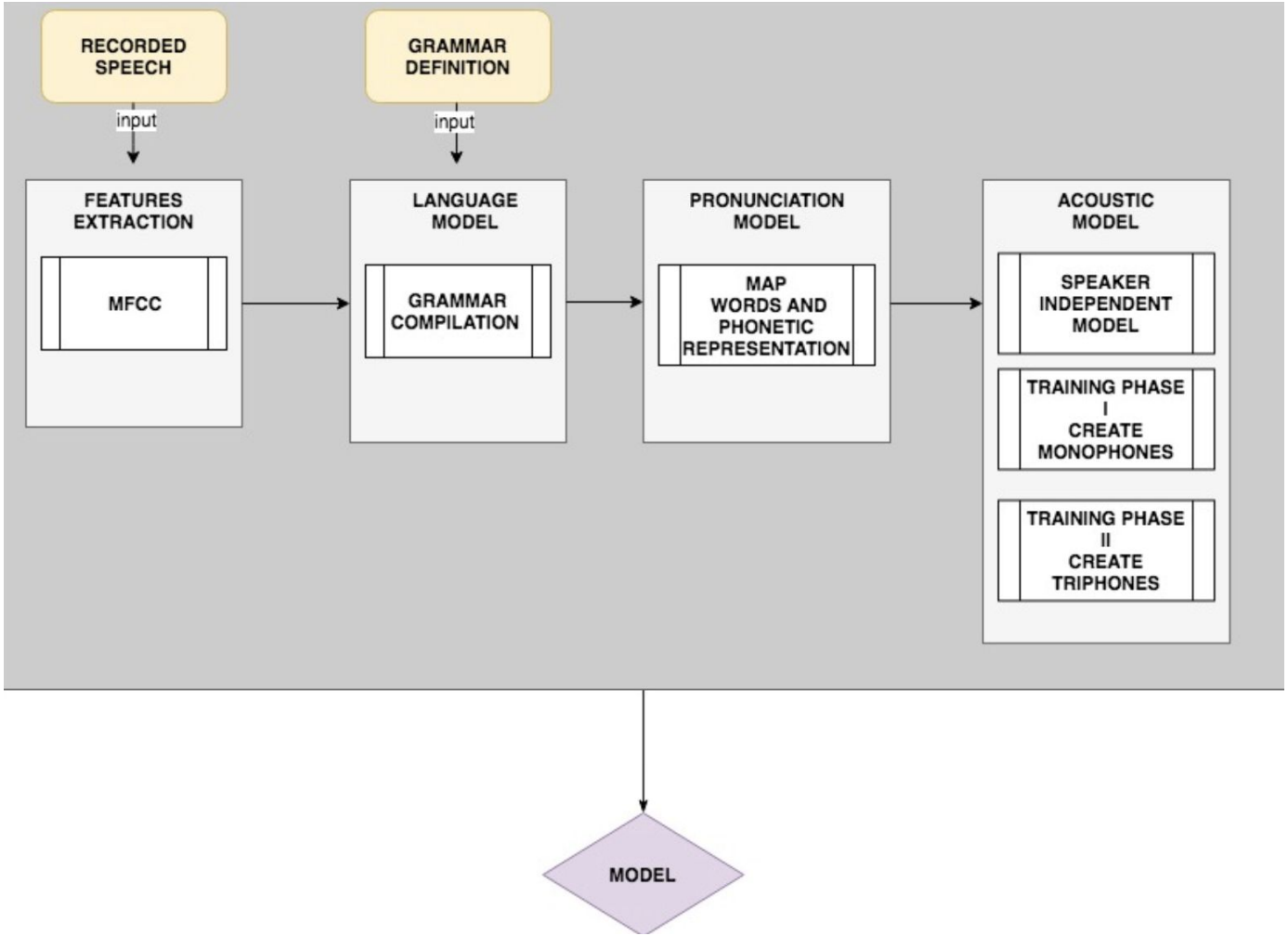
Linux Bash

Python

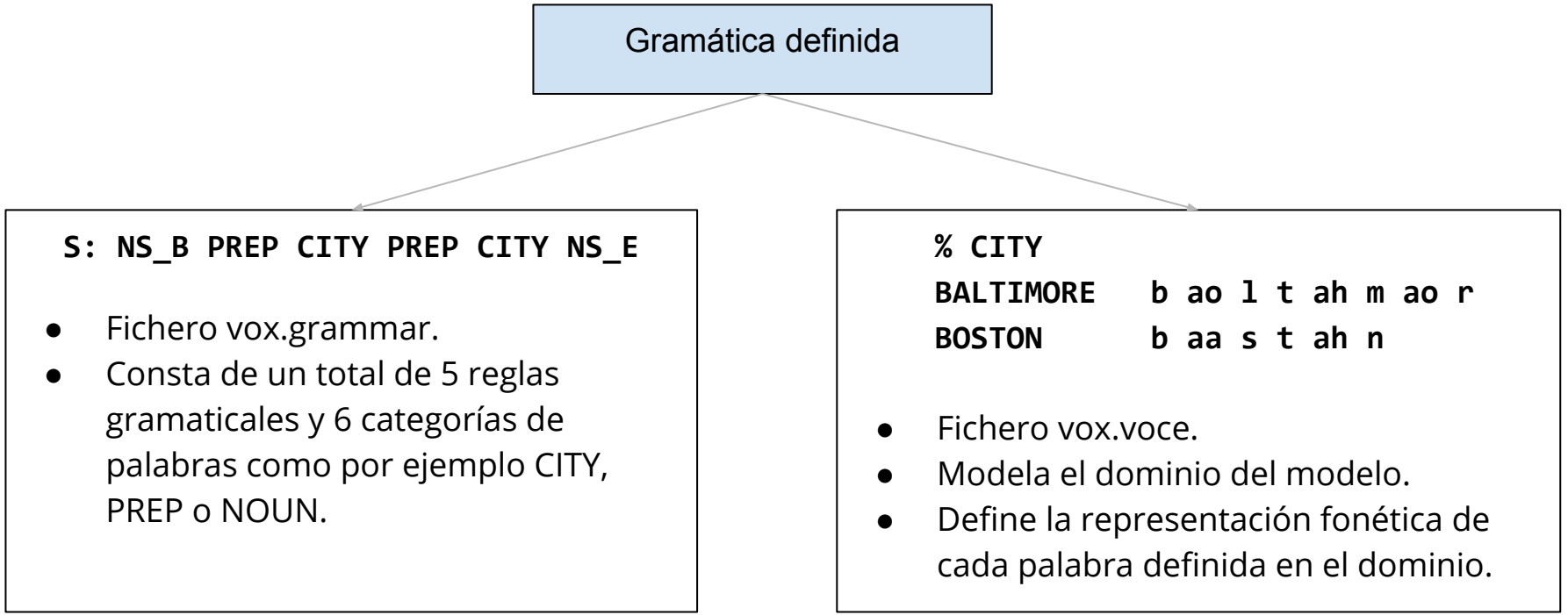
Keras

Módulo ASR

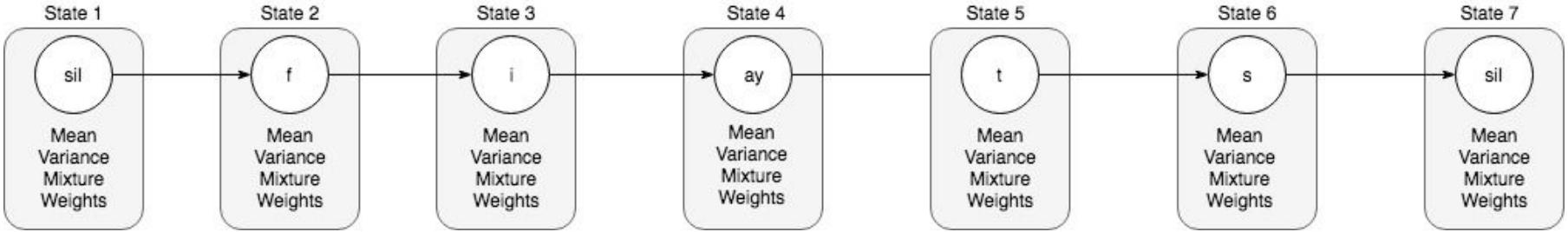
Arquitectura



Modelo de lenguaje basado en reglas gramaticales

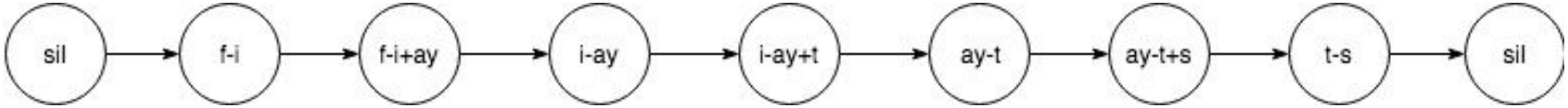


Modelo acústico **monofónico** basado en HMMs



- Cada palabra se representa fonema a fonema.
- Los fonemas son independientes del contexto.
- Se crea un conjunto de HMMs monofónico que se entrenará durante n iteraciones a lo largo de la primera fase de entrenamiento.
- Cada HMM estado está representado por la media, varianza y el peso estadístico computado de los samples de training proporcionados.

Modelo acústico **trifónico** basado en HMMs



- Cada palabra agrupa los fonemas de tres en tres.
- Los fonemas son dependientes del contexto de cada uno de ellos.
- Es un modelo más robusto que modelo anterior.
- La creación del modelo consiste en la transformación de las representaciones monofónicas a representaciones trifónicas.

Resultados parciales

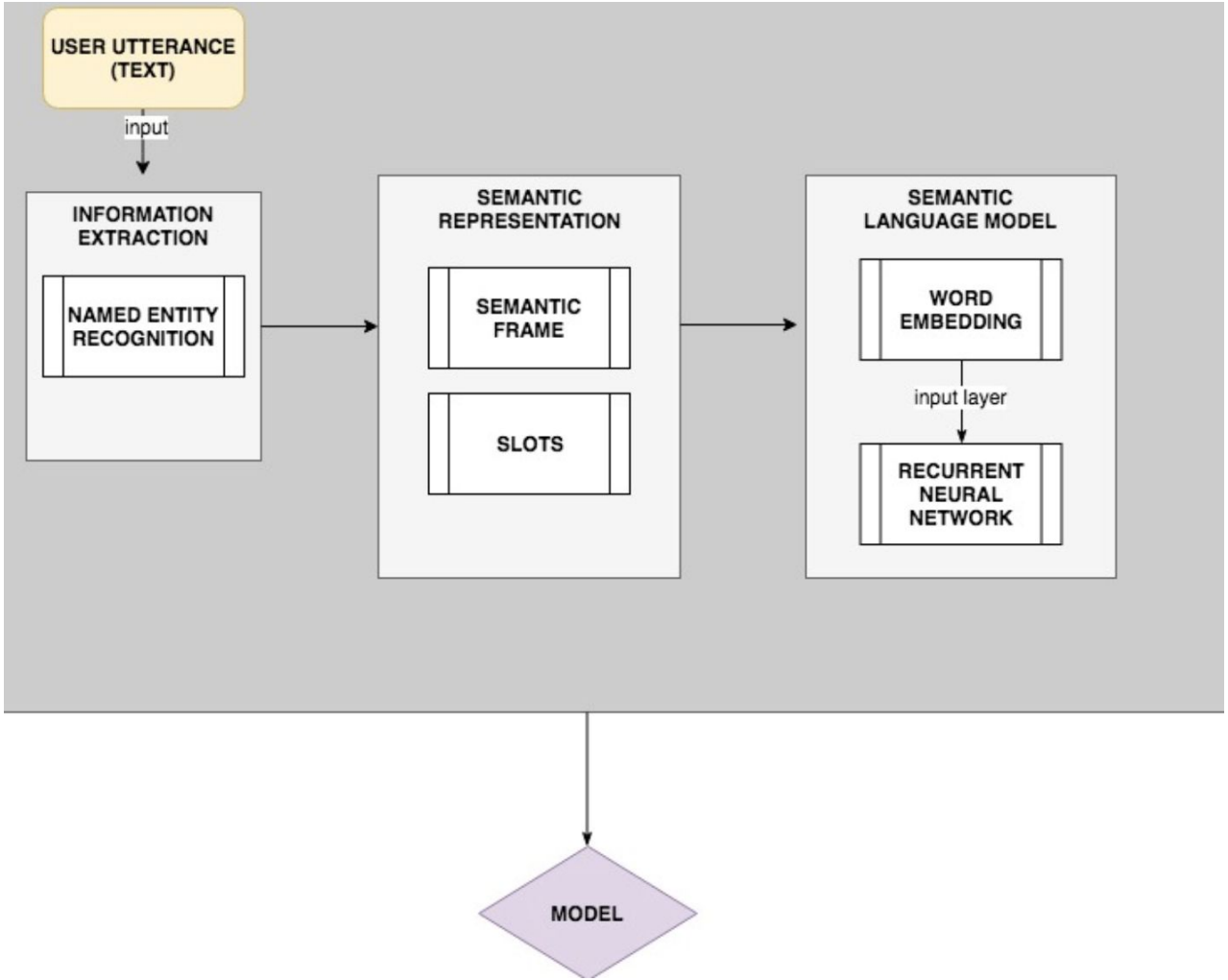
- El modelo presenta una precisión del 61.95 %.
- %correct de palabras aisladas es del 63.11%
- %correct de frases completas es del 17.89%.
- Error global del sistema es del 38.05%

Módulo SLU

Spoken Language Understanding

- La tarea principal del componente es proveer de comprensión a los enunciados del usuario, detectando y extrayendo información de dichos enunciados.
- Los tres factores que permiten analizar la comprensión de un enunciado son:
  - Clasificación del dominio. De qué está hablando el usuario (vuelos, libros ...)
  - Intención del enunciado (Show me flights from Barcelona to Berlin)
  - *Filling Slots*. Tarea que determina qué *slots* y que *fillers* han de ser extraídos de la orden que da el usuario para posteriormente entenderla.

Arquitectura



Representación semántica

- Enfoque basado en *semantic frame*
- Dominio conocimiento limitado
- Cada frame contiene diferentes elementos llamados *slots.*
- Tarea: Slot filling → tipo parsing semántico
- Objetivo: identificar los frames adecuados de un enunciado y extraer la información clave para rellenar los slots de dicho frame.

I need flights from Charlotte to Baltimore on Tuesday Morning

```
<frame name="needFlight" type="Void">
      <slot name="flight" type="Flight">
</frame>
<frame name="flight" type="Flight">
      <slot name="origCity" type="City">
      <slot name="destCity" type="City">
      <slot name="day" Type="day">
      <slot name="partDay" type="PartDay">
</frame>
```

Modelo Semántico basado en Recurrent Neural Network (RNN)

- Los modelos basados en Redes neuronales relacionan una representación distribuida de cada palabra que compone el diccionario y calculan la probabilidad conjunta de dar diferentes secuencias de palabras a dichos vectores.
- Las características principales de la RNN implementada son:
  - Arquitectura de Elman → Simple Recurrent Network (SRN)
  - El input de la RNN es una capa preentrenada con el método Word - embedding (one - hot representation).
  - Conexiones cíclicas de la red.
  - Propagación feed-forward.
  - Capa “escondida” o “hidden” simula la memoria de la RNN.

Resultados parciales

- El modelo presenta una precisión del 93.3 %.
- El % de error que se da al evaluar la diferencia del output real y la predicción del output esperado es del 33.3%.

Demostración

Conclusiones

- El sistema creado es un sistema robusto y de fácil mantenimiento.
- En cuanto al módulo ASR decir que es capaz de reconocer de manera más eficiente palabras aisladas. Por un lado, es más fácil determinar el final de dicha palabra y por otro, la pronunciación de la siguiente palabra no interfiere con la de la anterior.
- En cuanto al módulo SLU decir que pese a que tras su evaluación se ha determinado con 93% su tasa de acierto, la performance que lleva a cabo cuando ambos módulos funcionan integrados es peor que cuando trabaja en solitario. Esto es debido principalmente a que el input del SLU está condicionado por el output generado por el ASR.

Mejoras y Líneas de trabajo futuro

- Integrar ambos módulos en un único componente.
- Trabajar con técnicas de *deep learning* como Redes Neuronales para mejorar la precisión del modelo acústico del módulo ASR.
- Expandir la gramática del modelo de lenguaje que reconoce el ASR, a una gramática que permita reconocer eventos temporales y estructuras sintácticas más complejas.
- Experimentar con otro tipo de Redes Neuronales como *LSTM* para el desarrollo del módulo SLU y comparar cuál proporciona mayor precisión y performance.
- Desarrollar un sistema de diálogo que incorpore el *spoken language system* creado.

Gracias por su atención