

Proyecto Final

**Detección de Intenciones y
Reconocimiento de Entidades
con una RNN de arquitectura
Seq2Seq y mecanismo de
Atención**

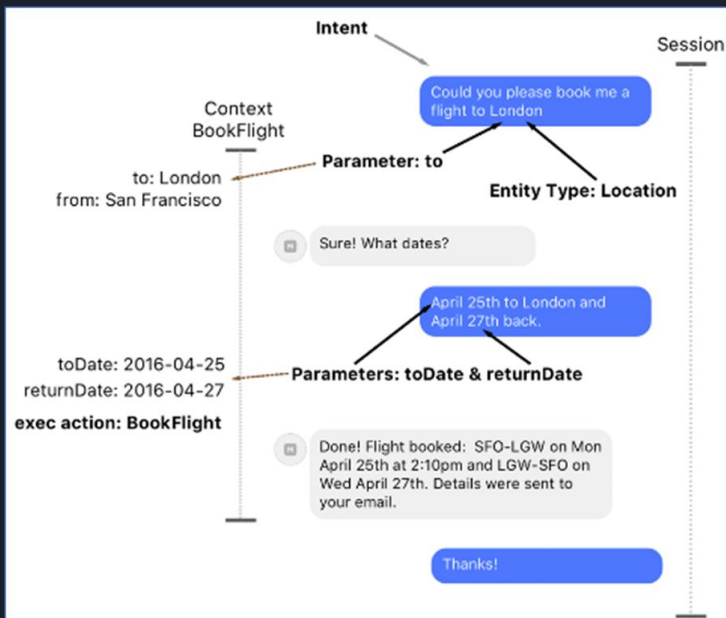
Integrantes

Daniel Arteaga Melendez
Juan Carlos Tovar Galarreta
Rubén Córdova Alvarado

INTRODUCCIÓN



Introducción: Detección de Intenciones y Reconocimiento de Entidades

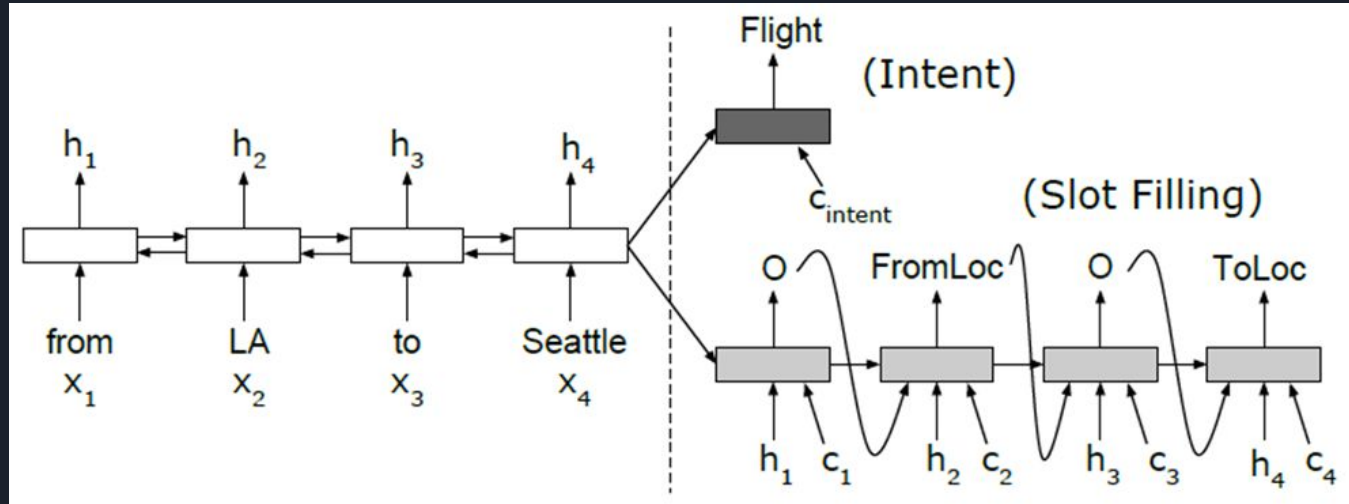


I need two economy tickets from San Francisco to London departing next Sunday back on Oct 30

i need **two** **economy** tickets from **san francisco** to **london** on **united** departing **next sunday** back **on oct 30**

place_from	: san francisco	: OAK, SFO, SJ
place_to	: london	: LHR, LTN, LCY, LGW, STN
date_depart	: next sunday	: 2017-10-15
date_return	: oct 30	: 2017-10-30
num_tickets	: two	: 2
class_type	: economy	: economy
airline	: united	: UA

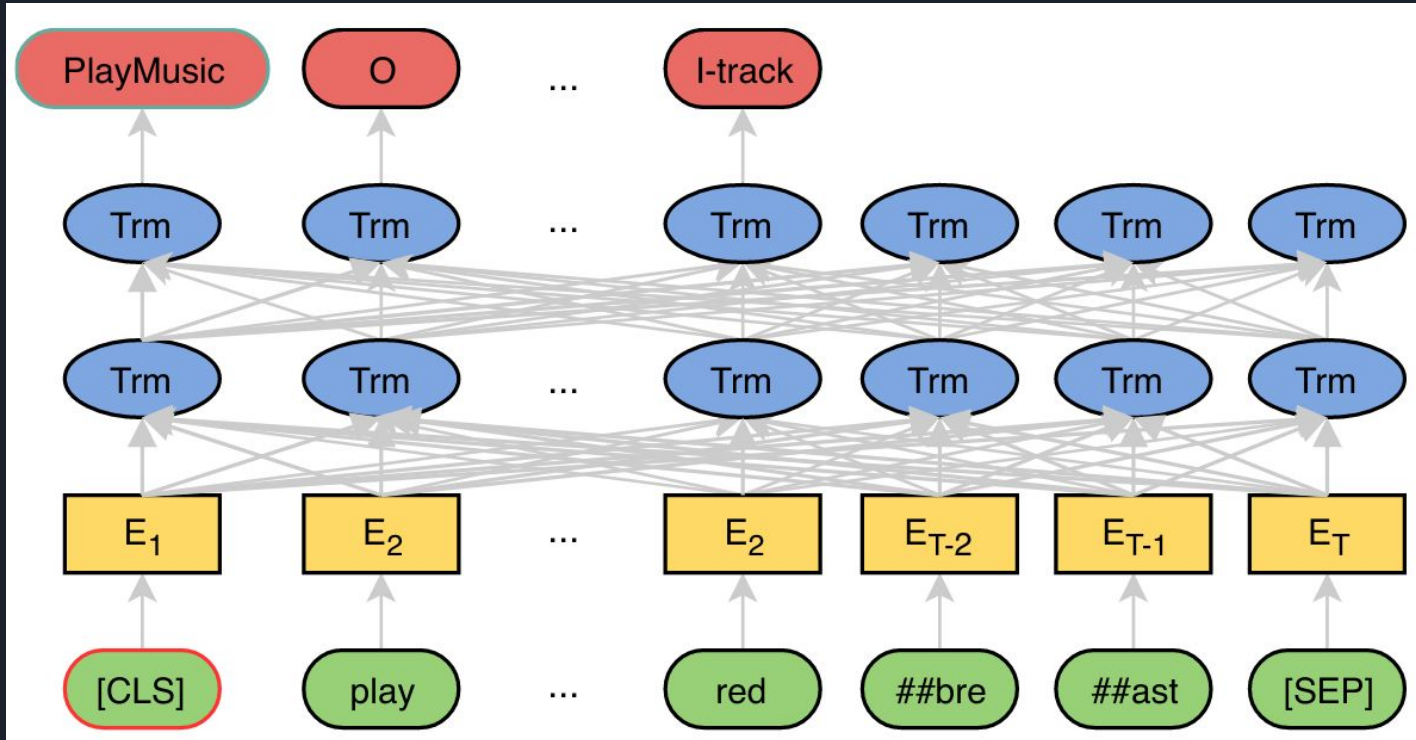
Introducción: Secuencia a Secuencia (Seq2Seq) con Atención



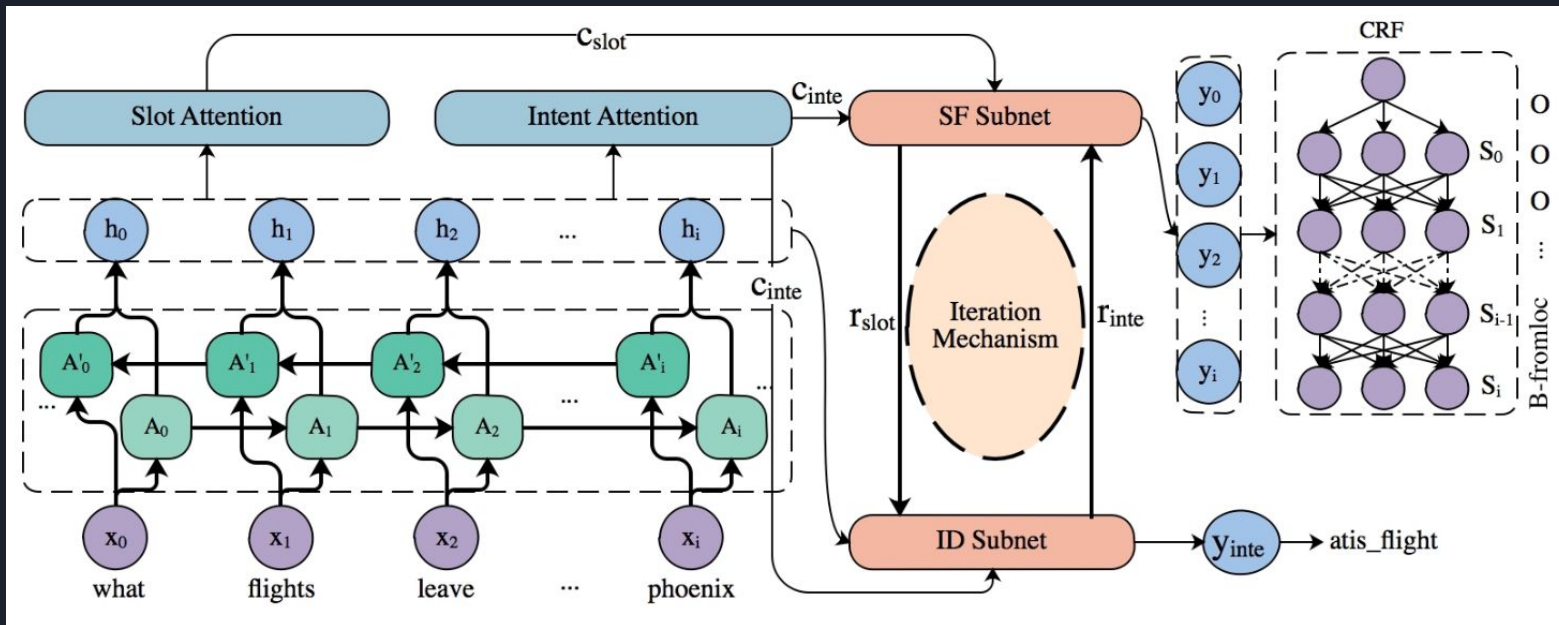
ESTADO DEL ARTE



BERT for Joint Intent Classification and Slot Filling



Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling

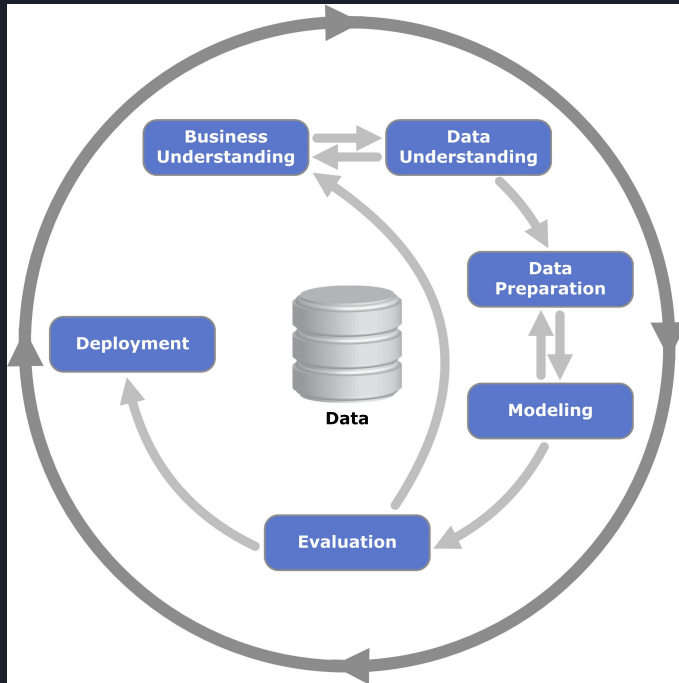


METODOLOGÍA



Metodología: Marco

CRISP-DM



01	Modelos	<ul style="list-style-type: none"> Modelo seq2seq con mecanismo de atención Modelo BERT Modelo BLSTM con CRF
02	Métricas de evaluación	<ul style="list-style-type: none"> Accuracy F1-Score
03	Configuración de los hiper parámetros	<ul style="list-style-type: none"> Tamaño de los embeddings Dimensión latente de las capas LSTM Tamaño del batch Learning rate

Metodología: Experimentación con modelos Seq2Seq

- Comparar el rendimiento del modelo propuesto con otros modelos sobre el conjunto de datos ATIS
- Comparar la influencia del padding al inicio de la secuencia (prepadding) con el padding al final de la secuencia (postpadding)
- Probar el modelo propuesto con otro conjunto de datos manualmente etiquetados en el dominio de consultas sobre cursos de capacitación

1	BOS i want to fly from baltimore to dallas round trip EOS	0 0 0 0 0 0 B-fromloc.city_name 0 B-toloc.city_name B-round_trip I-round_trip atis_flight
2	BOS round trip fares from baltimore to philadelphia less than 1000 dollars round trip fares from denver to philadelphia less than 1000 dollars round trip fares from pittsburgh	
3	BOS show me the flights arriving on baltimore on june fourteenth EOS	0 0 0 0 0 0 0 B-toloc.city_name 0 B-arrive_date.month_name B-arrive_date.day_number atis_flight
4	BOS what are the flights which depart from san francisco fly to washington via indianapolis and arrive by 9 pm EOS	0 0 0 0 0 0 0 0 B-fromloc.city_name I-fromloc.city_name
5	BOS which airlines fly from boston to washington dc via other cities EOS	0 0 0 0 0 B-fromloc.city_name 0 B-toloc.city_name B-toloc.state_code 0 0 0 atis_airline
6	BOS i'm looking for a flight from charlotte to las vegas that stops in st. louis hopefully a dinner flight how can i find that out EOS	0 0 0 0 0 0 0 B-fromloc.city_name 0 B-t
7	BOS okay and then from pittsburgh i'd like to travel to atlanta on september fourth EOS	0 0 0 0 0 B-fromloc.city_name 0 0 0 0 B-toloc.city_name 0 B-depart_date.month_name B-
8	BOS show me all the flights from philadelphia to cincinnati EOS	0 0 0 0 0 0 B-fromloc.city_name 0 B-toloc.city_name atis_flight
9	BOS okay i'd like a flight on us air from indianapolis to san diego in the afternoon what's available EOS	0 0 0 0 0 0 0 B-airline_name I-airline_name 0 B-fromloc.city_na
10	BOS on tuesday what flights leave phoenix to st. paul minnesota and leave after noon EOS	0 0 B-depart date.day name 0 0 0 B-fromloc.city name 0 B-toloc.city name I-tolo

[illegible]

Metodología: Preprocesamiento de los datos (textos)

Prepadding:

Input	PAD	PAD	PAD	...	PAD	BOS	w1	w2	w3	w4	w5	w6
Entities	PAD	PAD	PAD	...	PAD	BOS	e1	e2	e3	e4	e5	e6

Postpadding:

Input	BOS	w1	w2	w3	w4	w5	w6	EOS	PAD	PAD	PAD	...
Entities	BOS	e1	e2	e3	e4	e5	e6	PAD	PAD	PAD	PAD	...

Metodología: Preprocesamiento de los datos (textos)

```
1 from keras.preprocessing.text import Tokenizer
2 tokenizer = Tokenizer()
3 tokenizer.fit_on_texts(data[:,0])
4 tokenizer.word_index
```

```
{'i': 1,
 'it': 2,
 'so': 3,
 'no': 4,
 'like': 5,
 'much': 6,
 'good': 7,
 'hate': 8,
 "don't": 9,
 'at': 10,
 'all': 11,
 'bad': 12,
 'feel': 13,
 'nothing': 14,
 'love': 15,
 'one': 16,
 'can': 17,
 'overcome': 18}
```

Input	PAD	PAD	PAD	...	PAD	BOS	w1	w2	w3	w4	w5	w6
Tok	0	0	0	...	0	2	4	78	23	453	20	1522

Entiti es	PAD	PAD	PAD	...	PAD	BOS	e1	e2	e3	e4	e5	e6
Tok	0	0	0	...	0	0	1	5	6	1	1	12

Metodología: Preprocesamiento de los datos (textos)

Prepadding:

Inp. Entities	PAD	PAD	PAD	...	PAD	BOS	e1	e2	e3	e4	e5	e6
Tok	0	0	0	...	0	1	2	5	6	2	2	12
Out. Entities	PAD	PAD	PAD	...	PAD	e1	e2	e3	e4	e5	e6	EOS
Tok	0	0	0	...	0	2	5	6	2	2	12	1

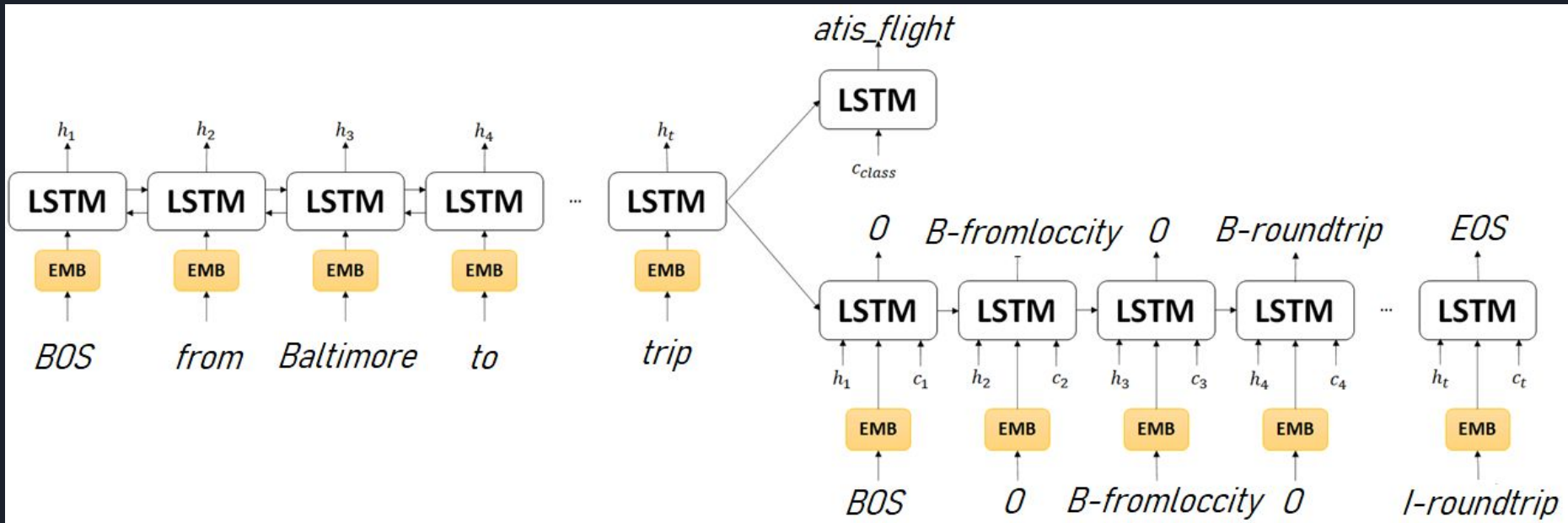
Postpadding:

Inp. Entities	BOS	e1	e2	e3	e4	e5	e6	PAD	PAD	PAD	PAD	...
Tok	1	2	5	6	2	2	12	0	0	0	0	...
Out. Entities	e1	e2	e3	e4	e5	e6	EOS	PAD	PAD	PAD	PAD	...
Tok	2	5	6	2	2	12	1	0	0	0	0	...

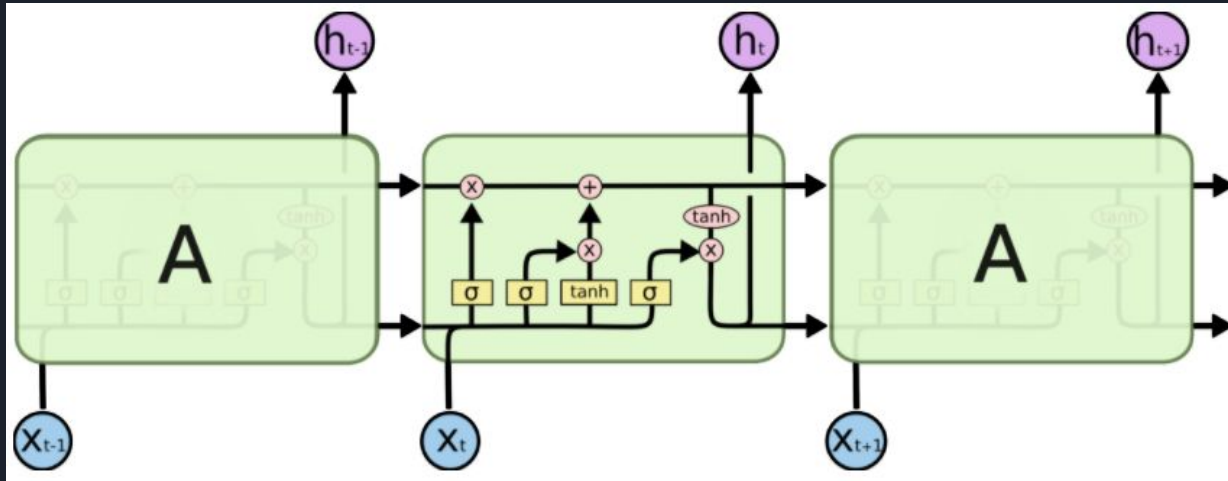
- | | queries | entities | intents |
|---|--|--|--------------------------|
| 0 | estimados sres . | 0 0 0 | saludo |
| 1 | favor de indicar si ya se puede recoger el certificado correspondiente del curso en mencion o si pueden remitirlo vía correo al menos hasta que se pueda recoger de forma personal , | 0 0 0 0 0 0 0 0 0 B-documento_certificacion 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | entrega_certificado |
| 2 | saludos cordiales , | 0 0 0 | despedida |
| 3 | heber | B-nombre_usuario | informacion_usuario |
| 4 | muy buenos días agradeceré dar un alcance del curso de cctv ya es mi voluntad de seguirlo . | 0 0 0 0 0 0 0 0 0 0 B-nombre_curso 0 0 0 0 0 0 0 0 | informacion_curso |
| 5 | gracias | 0 | gracias |
| 6 | ¿ podría pagar todo de una sola vez ? es decir s/. 660 para terminar de pagar todo el modulo ii . | 0 0 0 0 0 0 0 0 0 0 0 0 B-monto I-monto 0 0 0 0 0 0
0 0 0 | informacion_proceso_pago |
| 7 | aqui adjunto la constancia de transferencia así como la información solicitada : | 0 0 0 B-documento_pago 0 B-medio_pago 0 0 0 0 0 0 0 | pago_realizado |
| 8 | como accedo al 30 por ciento del descuento ? . | 0 0 0 B-descuento I-descuento I-descuento 0 0 0 0 | descuentos |

	intents	frequencies
0	error	5
1	consulta_aprobacion	6
2	sin_clase	8
3	si	15
4	cotizacion	15
5	certificacion	15
6	medio_pago	18
7	solicitud_devolucion	25
8	descuentos	31
9	solicitud_accesos	43
10	despedida	50
11	confirmacion_inscripcion	57
12	transparente	58
13	informacion_proceso_pago	63
14	inicio_curso	71
15	entrega_certificado	86
16	gracias	108
17	saludo	265
18	otro	298
19	pago_realizado	322
20	informacion_curso	419
21	informacion_usuario	2037

Metodología: Modelo Propuesto



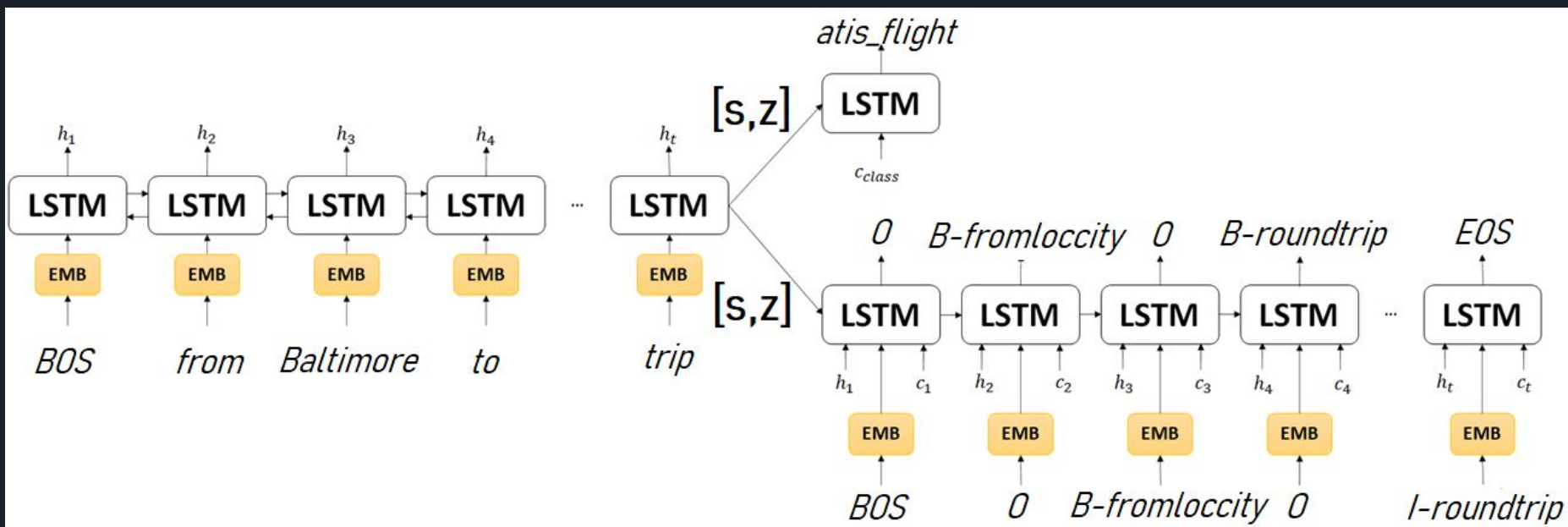
Metodología: Modelo Propuesto



$$s_0 = \tanh(W_s \overleftarrow{h_1})$$

$$z_0 = \tanh(W_z \overleftarrow{cs_1})$$

Metodología: Modelo Propuesto



Metodología: Ajuste de Hiperparámetros

- Tuning manual en base a la evolución de resultados:
 - word emb: 64, latent dim: 64, batch size: 16, lr:0.001
 - word emb: 128, latent dim: 64, batch size: 16, lr:0.001
 - word emb: 128, latent dim: 128, batch size: 16, lr:0.001
 - word emb: 128, latent dim: 128, batch size: 32, lr:0.001
 - word emb: 256, latent dim: 256, batch size: 32, lr:0.001
 - word emb: 128, latent dim: 128, batch size: 32, lr:0.001, decay steps: 5, decay rate: 0.95

EXPERIMENTACIÓN Y RESULTADOS



Experimentación: Conjunto de datos



❖ Dataset ATIS

- 4978 oraciones para entrenamiento
- 893 oraciones de prueba

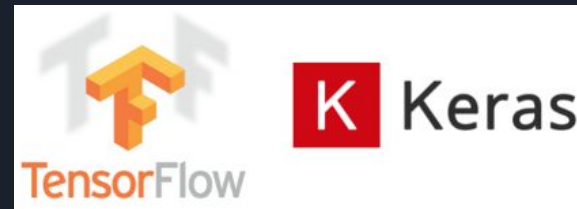
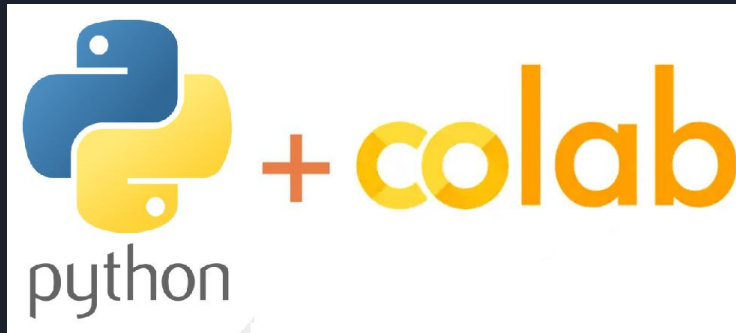
Sentence	first	class	fares	from	boston	to	denver
Slots	B-class_type	I-class_type	O	O	B-fromloc	O	B-toloc
Intent	airfare						

❖ Dataset Propio

- 2878 oraciones para entrenamiento
- 320 oraciones de prueba

Oraciones	Secuencia de Entidades	Intenciones
Informacion del curso de SISTEMAS DE video vigilancia CCTV porfa	<o> <o> <o> <o> B-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> <o>	informacion_curso

Experimentación: Entorno




Experimentación: Reproducción de resultados

Para poder comparar los resultados se empleó:

- Tamaño del batch: 32
- Tamaño máximo de la secuencia: 50
- Postpadding
- Conjunto de datos ATIS

Modelo	F1-Score (Entidades)	Exactitud (Intenciones)
BERT	96.10	97.5
SF-ID Network	95.80	97.76

In the top left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram, both tilted at an angle.

Experimentación: Entrenamiento del Modelo Propuesto

La configuración de hiperparámetros que se utilizó para los resultados finales con el conjunto de datos ATIS fue la siguiente:

- Tamaño máximo de la secuencia: 50
- Tamaño del embedding de palabras: 128
- Dimensión latente de las capas LSTM: 128
- Tamaño del batch: 32 (para entrenamiento) y 64 (para validación)
- Learning rate: Se empleó un decaimiento exponencial cada 5 épocas con un ratio de 0.95 y un valor inicial de 0.001

Experimentación: Entrenamiento del Modelo Propuesto

Para el caso del conjunto de datos propio, las diferencias en la configuración de hiperparámetros con respecto a ATIS fue:

- Tamaño máximo de la secuencia: 150.
- Tamaño del embedding de palabras: 250.

Función de pérdida:

Métrica para la detección de intenciones:

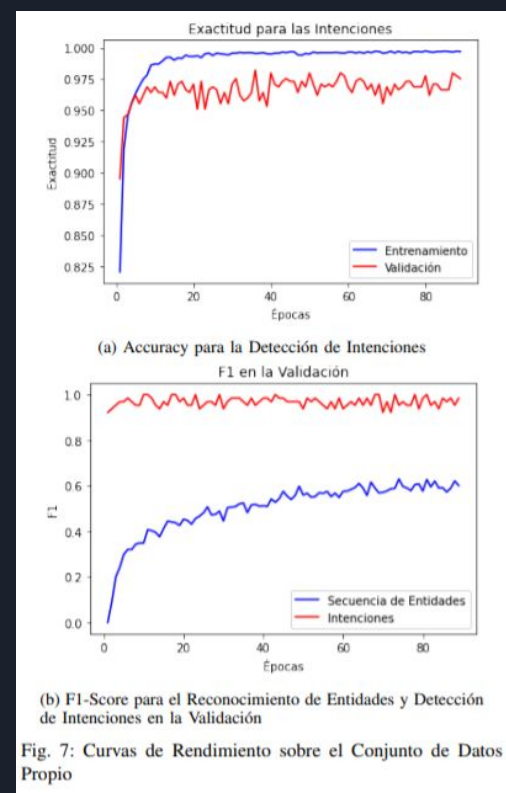
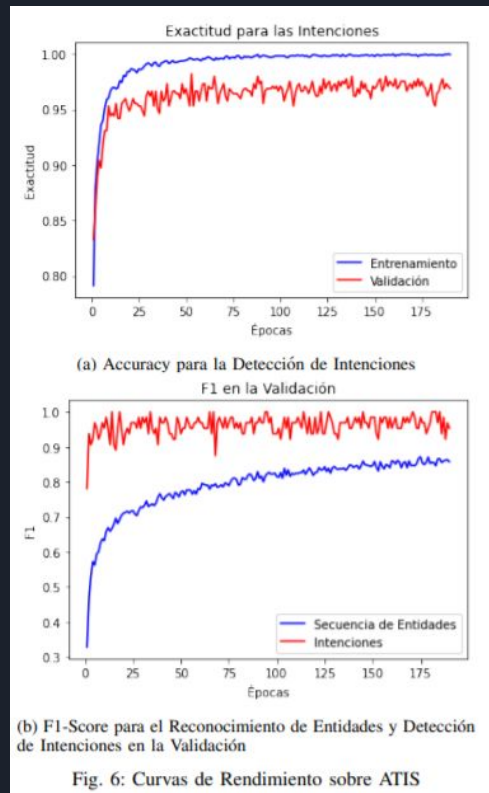
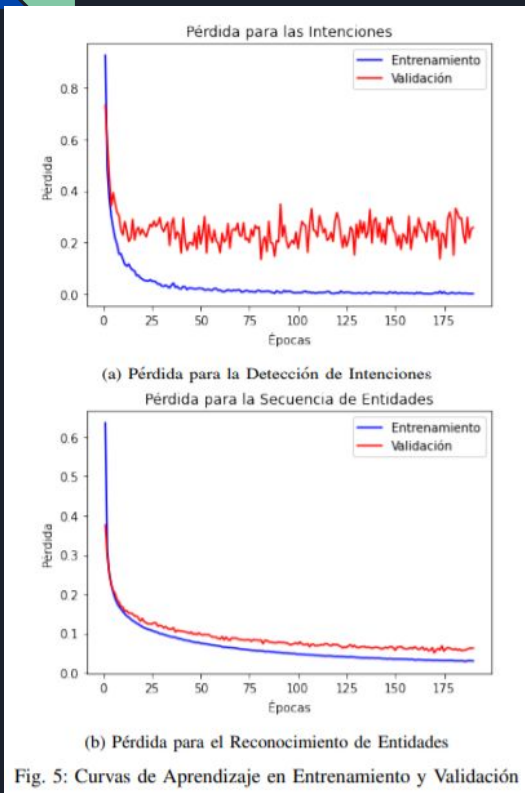
Métrica para el reconocimiento de entidades:

Entropía Cruzada Categórica Sparse

Exactitud Categórica Sparse, F1-Score

F1-Score

Experimentación: Entrenamiento del Modelo Propuesto



DISCUSIÓN



Comparación de Modelos

- Comparación realizada empleando el **conjunto de datos ATIS con post padding**.
- Batch sizes de 32 (entrenamiento) y 64 (validación), así como un tamaño máximo de secuencia igual a 50

TABLE I: Comparación de modelos

Modelo	F1-Score (Entidades)	Exactitud (Intenciones)
BERT	96.10	97.5
SF-ID Network	95.80	97.76
Modelo Propuesto	87.13	97.54

Comparación de Modelos

- F1-score (reconocimiento de entidades)
En pre-padding ligeramente menor comparado con post-padding.
- Tiempo de entrenamiento
Con pre-padding es aprox. una hora más que con post-padding (a pesar de realizar 19 iteraciones menos).

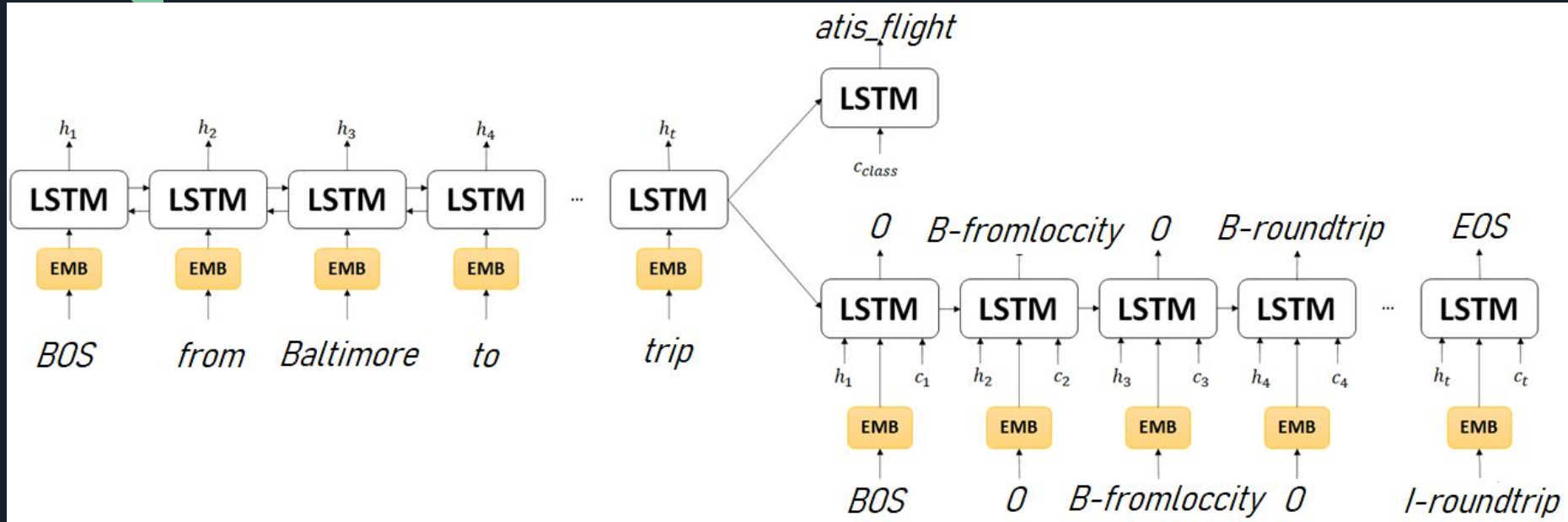
TABLE II: Comparación de prepadding y postpadding

Modelo	F1-Score	Exactitud	Épocas	Tiempo
Postpadding	87.13	97.54	190	4h 25min
Prepadding	85.95	97.54	171	5h 17min

Interpretación de los resultados

- La tarea de reconocer exactamente las entidades es una tarea más compleja que la tarea de identificar intenciones
- Los modelos más actuales presentan mejores resultados pues presentan mayor complejidad (p. ej. basados en redes atencionales y embeddings contextualizados)
 - Pueden incorporar contexto en cada posición de la secuencia de entrada
 - Poseen representaciones de los tokens de acuerdo al contexto.
- El hecho de haber empleado una mayor cantidad de tokens (triple) en las secuencias es una de las razones de la disminución del rendimiento del modelo debido a que es más difícil mantener un contexto con una secuencia de mayor longitud.

Metodología: Modelo Propuesto



- Post-padding es más eficiente que pre-padding debido a que el decodificador (unidireccional) procesa la secuencia iniciando por los tokens de las entidades, construyendo un contexto que debe recordar hasta que llega a los tokens de padding.

Limitaciones experimentadas con el Modelo Propuesto

Principal limitación: procesamiento de secuencias de texto largas

- Mayor dificultad para representar **secuencias largas** en un **vector de contexto de tamaño fijo** (128 en LSTM)
- Reflejado en el **F1-score** del modelo (reconocimiento de entidades):
 - Dataset propio con secuencias de hasta 150 palabras (0.6)
 - Dataset ATIS con un tamaño de secuencia máximo de 50 (aprox 0.9)

Mejoras futuras del sistema

- Aumentando el tamaño del embedding de palabras y la dimensión latente de las capas LSTM (modelo más complejo).
- Disponer de un **conjunto de datos de buena calidad** y con una **adecuada cantidad de ejemplos** (uno de los problemas en NLP)
 - **Calidad de los datos:** realizar un buen etiquetado de los datos (semiautomatizado, requiere de la intervención humana de expertos en el dominio).
 - **Cantidad de ejemplos:** emplear Data Augmentation con Autoencoders (Autoencoders Variacionales o VAE).

CONCLUSIONES Y TRABAJO A FUTURO





Conclusiones

- Los modelos basados en **redes atencionales y embeddings condicionales** superan el **rendimiento del modelo propuesto** (incluso aplicando **alineamiento y atención** de codificador a decodificador).
- Se ha comprobado su **potencial y facilidad de implementación** (en comparación con los modelos del estado del arte) para poder aplicarse en sistemas como chatbots.
- Trabajo a futuro: se propone explorar la aplicación, viabilidad e integración de modelos de mayor complejidad como BERT en aplicaciones como chatbots en el **idioma español**.



GRACIAS