UNIVERSIDADE DA CORUÑA

Grado en Inteligencia Artificial

Fundamentos de Procesamiento de Lenguaje Natural Curso 2024-2025

Práctica 3 – Modelos neuronales de etiquetado de secuencias

En esta práctica, implementaremos una red neuronal para resolver tareas de etiquetado de palabras en oraciones escritas en lenguaje natural, y las aplicaremos a dos problemas clásicos de PLN: etiquetación morfológica (*PoS tagging*) y reconocimiento de entidades nombradas (*named-entity recognition*). La Tabla 1 ilustra gráficamente en qué consisten estas dos tareas con un ejemplo.

Inputs	En	Can	Roca	el	menú	cuesta	280	€
POS	PREP	PROPN	PROPN	DET	NOUN	VERB	NUM	SYMBOL
NER	O	B-ORG	I-ORG	O	O	O	B-PRICE	I-PRICE

Tabla 1. Ejemplo de una oración de entrada y una posible salida para las dos tareas propuestas: etiquetación morfológica y reconocimiento de entidades.

Los datasets que usaremos para ambas tareas se incluyen con los materiales de esta práctica: (i) el dataset PartUT para etiquetación morfológica y (ii) el MITRestaurant para el reconocimiento de entidades nombradas para el dominio de hostelería. Deberá entrenarse una red distinta para cada dataset.

En concreto, **se pide implementar una arquitectura neuronal basada en LSTMs** para tareas de etiquetado de secuencias. El modelo deberá ser de tipo Sequential o Functional API, y se propone una arquitectura por defecto compuesta por una capa Embedding que transforme las palabras en vectores de tamaño 100, una capa LSTM con dimensión de salida 64 que procese la secuencia de vectores y una capa de salida Dense, que calcule la distribución de probabilidad para predecir la etiqueta de cada palabra mediante una función softmax. Se deberán implementar y entrenar dos versiones del modelo: una con embeddings inicializadas aleatoriamente y otra con embeddings pre-cargadas del Word2Vec de Google.

Las funcionalidades que deberán soportar las redes serán:

Entrenar los modelos utilizando un conjunto de entrenamiento (train.txt) y un conjunto de validación (dev.txt) para la evaluación interna.

Testar los modelos entrenados. Utilizar el modelo entrenado para calcular el rendimiento de la red sobre un conjunto de oraciones test.txt:

- En el caso de etiquetación morfológica, se deberá reportar la accuracy.
- En el caso del reconocimiento de entidades, además de la accuracy, se hará uso del repositorio nervaluate para reportar la F1-score en sus variantes: ent_type, partial, exact y strict para cada tipo de entidad.

Una posible opción de presentación de la práctica es mediante un programa Python que se ejecute con el siguiente comando:

```
python main.py --embedding [random, word2vec] --task [ner, pos] --train PATH_TRAINING_SET --dev PATH_DEVELOPMENT_SET --test PATH_TEST_SET
```

donde:

`--embedding` indica el tipo de inicialización de las embeddings:

- `random`: inicializa las embeddings aleatoriamente.
- `word2vec`: utiliza embeddings preentrenadas del modelo Word2Vec de Google.

`--task` especifica la tarea para la que se entrenará el modelo:

- `ner`: reconocimiento de entidades nombradas (Named Entity Recognition).
- `pos`: etiquetado morfológico (PoS tagging).

`--train`, `--dev` y `--test` definen las rutas a los conjuntos de datos utilizados para el entrenamiento, validación y prueba del modelo, respectivamente. Los valores deben ser las rutas a los ficheros train.txt, dev.txt y test.txt del dataset correspondiente.

Entrega

La práctica debe incluir un breve manual de usuario, que indique cómo ejecutar la práctica para entrenar y evaluar los modelos en los distintos datasets. También debería incluirse un análisis de los resultados y una breve discusión de las diferencias que aprecia entre ambas implementaciones, así como sus ventajas e inconvenientes. La memoria no debe exceder las 3 páginas y deberá utilizar un tipo de letra Arial, Calibri o Times New Roman con un tamaño mínimo de 11 pt.

Las prácticas se realizarán respetando los grupos formados en la asignatura.

Defensa: La práctica deberá subirse al campus virtual, en el apartado habilitado para ello, como muy tarde el 9 de mayo a las 23:59 horas. Únicamente un miembro del grupo debe realizar la entrega. La defensa, de carácter obligatorio, tendrá lugar durante la última semana de prácticas (del 12 al 16 de mayo). Los estudiantes que no asistan a la defensa suspenderán la práctica con una calificación de 0 puntos.