Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación
CC6205-1 Procesamiento de Lenguaje Natural

Tarea 2 - Named Entity Recognition

Fecha de Entrega: Viernes 10/01/2020 a las 23:59.

Esta tarea, al igual que la anterior, consiste en participar en una competencia. El objetivo de esta será entrenar un modelo de redes neuronales recurrentes (**RNN**) para la detección de entidades nombradas (**NER**) en español.

Named Entity Recognition (NER)

Esta tarea consiste en localizar y clasificar los tokens de una oración que representen *entidades nombradas*. Es decir, tokens que simbolicen *(1) personas, (2) organizaciones, (3) lugares y (4) adjetivos, eventos* y *otras entidades que no entren en las categorías anteriores* deberán ser taggeados como *(1)* **PER,** *(2)* **ORG,** *(3)* **LOC y** *(4)* **MISC** respectivamente. Adicionalmente, dado que existen entidades representadas en más de un token (como *La Serena*), se utiliza la notación **BIO** como prefijo al tag: Beginning, Inside, Outside. Es decir, si encuentro una entidad, el primer token etiquetado será precedido por **B**, el segundo por **I** y los n restantes por I. Por otra parte, si el token no representa ninguna entidad nombrada, se representa por **O**. Un ejemplo de esto es:

```
La          O
quinta      O
edición             O
del         O
Festival            B-MISC
de          I-MISC
Teatro      I-MISC
en          I-MISC
la          I-MISC
Calle       I-MISC
,           O
que         O
organiza            O
anualmente          O
el          O
Ayuntamiento        B-ORG
de          I-ORG
Villanueva          I-ORG
```

Dado los recientes acontecimientos, esta tarea será de carácter **opcional.** La nota de tareas será equivalente a la nota máxima entre las tareas realizadas.
Por otra parte, los recursos que les podrían ser útiles se encuentran en el github del curso.
En particular, estos links son los más indicados para comenzar:

- Tagging: http://www.cs.columbia.edu/~mcollins/cs4705-spring2019/slides/tagging.pdf
- RNN: https://github.com/dccuchile/CC6205/blob/master/slides/NLP-RNN.pdf

Algunos detalles de la competencia:

- Para participar, deben registrarse en la competencia en Codalab en grupos de máximo 2 alumnos. Cada grupo debe tener un nombre de equipo. (¡Y deben reportarlo en su informe!)
- Es solo 1 problema de Sequence Labelling. Por ende, se evaluará solo dicho problema. Las métricas usadas serán Precisión, Recall y F1.
- **Link al notebook con el baseline**: https://drive.google.com/file/d/1IxYj4GBcPLe8RXgYRACAajpi4zB_nHXi/view?usp=sharing . Si bien, no es necesario, se recomienda ejecutar la tarea en colab. Este aporta el entorno de desarrollo como también la GPU para entrenar las redes. De lo contrario, tendrán que instalar todo a mano.
- **Link a la competencia**: https://competitions.codalab.org/competitions/21613?secret_key=762dae08-556f-4441-a94b-d5f86ab83b5f
- En total pueden hacer un **máximo de 4 envíos**.
- Hagan varios experimentos, variando los hiperparámetros de la red: cantidad de parámetros de la capa de embeddings, cantidad de capas RNN, cantidad de parámetros de las capas de RNN, embeddings preentrenados, cantidad de épocas de entrenamiento, arquitectura de la red, otros modelos de RNN como GRU o de Transformers, cambiar el optimizador, learning rate, batch size, usar una CRF-loss, etc…
- Es requisito tanto participar en la competencia como enviar el informe para ser evaluado.
- Verificar que los resultados obtenidos por el clasificador coincidan con la especificación de la competencia. De lo contrario, no se corregirán bien.

Además de participar, deben enviar un reporte en formato Jupyter-Notebook a Ucursos con la siguiente estructura:

1. **Introducción**: presentar el problema y los métodos utilizados en la tarea.
2. **Trabajo relacionado**: describir brevemente el trabajo anterior relacionado con el problema. Pueden sacar más detalles del problema en este paper.
3. **Algoritmos y representaciones**: describir los algoritmos de clasificación, los atributos y las métricas de evaluación utilizadas en sus experimentos.
4. **Experimentos**: reportar todos sus experimentos. Comparar los resultados obtenidos utilizando diferentes algoritmos y atributos. Incluyan todo el código de sus experimentos aquí. Es vital haber realizado varios experimentos para sacar una buena nota!
5. **Conclusiones**: discutir resultados.

Nuevamente, recuerden poner su nombre de usuario y de equipo en su reporte.