Proyecto 1. Postagging usando word2vec y Deep Learning

Prof. Raúl Ernesto Gutierrez de Piñerez Reyes, Ph.D.
PAIS-EISC

1 Introducción

La tarea de postagging es una de las más importantes de PLN, y consiste en la clasificación de la categoría de las palabras (tags) para una sentencia dada en cualquier lenguaje. Para una secuencia de palabras $w_1, w_1, \ldots, w_n$, la tarea es encontrar un conjunto de etiquetas (categorías) $t_1, t_1, \ldots, t_n$, . Esta es considerada una tarea de PLN de etiquetado secuencial (Sequential labeling). En este proyecto se pretende implementar un clasificador de postagging usando Deep Learning y word2vec. Los métodos tradicionales para este tipo de clasificación van desde el uso de las cadenas de Markov, pasando por el uso de las CSPs, MEMMs y CRFs. En este trabajo la idea es implementar un modelo de clasificación teniendo como entrada la vectorización de palabras del corpus ANCORA usando word2vec y el uso de una red neuronal BLSTM.

2 Problema

El proyecto debe resolver el problema del etiquetado secuencial de una sentencia en tres partes; 1) El preprocesamiento de texto y vectorización de subpalabras. 2) El entrenamiento de la red neuronal y optimización. 3) El tageador se deberá poner en funcionamiento en el editor de texto que tiene como tokenizador a SentencePiece.

3 Partes del proyecto (Por ahora 50%)

Los siguientes son los temas a tener en cuenta para la evaluación:

- Procesamiento de texto y vectorización de palabras usando el corpus Spanish Billion https://github.com/dccuchile/spanish-word-embeddings
- Implementación del modelo de red neuronal BLSTM y el uso de un CRF para la capa de salida.
- Implementar el tageador en el tokenizador de SentencePiece usando solo la tokenización de palabras. Estas palabras deben ser vectorizadas con

word2vec y probadas con el modelo generado por de red entrenada + CRF.