Tarea 5. Seq2Seq a Nivel Palabra

Aprendizaje de Automático I

April 6, 2021

Descripción. Implementar un traductor Seq2Seq a nivel palabras para traducir oraciones del inglés al español.

Utilizando el modelo seq2seq que funciona a nivel carácter en `https://www.cimat.mx/~mrivera/cursos/aprendizaje_profundo/seq2seq/seq2seq.html` implementar la versión que funcione a nivel palabra.

Luego haga inferencia (prediccin de traducciones) de una serie de frase en inglés al español.

1. Busque una base de datos con frases de inglés y su traducción en español. Se recomienda usar Open Parallel corpUS (OPUS).

   - Artículo explicando el proyecto opus y su uso `http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf`.
   - La página del proyecto esta en `https://opus.nlpl.eu`. Para ver los *parallel corpora* de *English* y *Spanish* seleccione :
     Search download resources: $<< en(English) >><< es(Spanish) >><< all >>$.
   - Descripción del formato de los datos la puede encontra en `https://opus.nlpl.eu/trac/wiki/DataFormats.html`.
   - Librería OpusFilter (Python) para filtrar y combinar corpus paralelos. Utiliza la biblioteca OpusTool para descargar datos de la colección de corpus OPUS `https://github.com/Helsinki-NLP/OpusFilter`.

2. Use secuencias alienadas crudas, sin etiquetas por tipo de palabra (sustantivo, adjetivo, etc.).

3. Encuentre una codificación adecuada de cada token (palabra) en un vector. Sugerencia usar Word2Vec de inglés y de Español. Por ejemplo, corpus y vectores de codificación disponibles en

   - `http://vectors.nlpl.eu/repository/`

- `https://radimrehurek.com/gensim/models/word2vec.html` para inglés y `https://crscardellino.github.io/SBWCE/` para español.
- `https://github.com/dccuchile/spanish-word-embeddings` para español.

NOTA: Los corpus alineados y los embeddings descritos son solo ejemplos, usted puede usar otras bases de datos, siempre que sean de secuencias alienadas en inglés - español.
Entrega de la tarea La tarea se entrega como el fuente del notebook de jupyterlab (.pynb) con la última ejecución. Indicando al inicio que copus y embedding es utilizó para cada lenguaje.
Enviar la tarea a aprendizaje.maquina@cimat.mx. Con asunto: "Tarea *número_de_tarea*. grupo *nombre_del_curso_inscrito*". Ejemplo: Tarea 5. grupo Aprendizaje Automático I
Fecha de entrega: 15 de abril 2021 a las 12pm (límite).
Penalización por retraso: la calificación de la tarea se multiplicará por 0.9^n donde $n \geq 0$ son los días de retraso.

Material de apoyo

Ligas arriba inlcuidas: Notas del curso en internet y ligas a los corpus-embeddings.