

Pontificia Universidad Católica de Chile  
Escuela de Ingeniería  
Aprendizaje Profundo [IIC3697]

Informe de Avance de Proyecto:

Análisis de Texto con Aprendizaje Profundo

**Integrantes:**

Dan Ustilovsky  
Alberto Valdés  
Elena Villalón

**Profesor:**

Álvaro Soto

**Introducción**

En el presente informe se detalla el avance del proyecto de Análisis de Texto con Aprendizaje Profundo, el cual se realiza con el set de datos SQuAD. El informe contiene revisión bibliográfica con su respectivo resumen, análisis e implementación de código, exploración de la base de datos, planteamiento de objetivos de trabajo, y una metodología con los pasos a seguir para llevarlos a cabo.

La revisión bibliográfica incluye un resumen y análisis de los artículos leídos y discutidos por el grupo, que son “Know What you don’t know: Unanswerable Questions for SQuAD” de Pranav Rajpurkar y “CS 224N Final Proyect: Question Answering on SQuAD”, que es un proyecto oficial realizado en un curso de Sanford. De dicha documentación, se estableció que el objetivo del proyecto es analizar las técnicas utilizadas para el aprendizaje de texto, junto con la realización de un aporte propio de probar dichas implementaciones para nuevos pre procesamientos de los datos, hiper parámetros y salidas de las redes, para finalmente comparar y analizar los resultados y los desempeños.

**Problema**

El problema que se debe resolver es dado un texto y una pregunta específica, determinar si la pregunta se puede responder y en caso de que se pueda responder, dar la respuesta ideal.

**Revisión Bibliográfica**

**Know What you don’t know: Unanswerable Questions for SQuAD**:

En este paper se presenta el problema que se tiene que resolver en este proyecto, que es poder identificar si dado un texto una pregunta específica se puede responder o no, y que en caso de que se pueda responder, responderla.

Se comenta que actualmente la mayoría de las redes neuronales implementadas son aquellas en las cuales la base no se debe reconocer si las preguntas se pueden responder dado un párrafo, sino que estas redes solo se entrenan con preguntas que pueden ser respondidas.

El dataset que solo tiene preguntas que pueden ser respondidas es SQuAD 1.1 y el dataset que incluye preguntas que no pueden ser respondidas es SQuAD 2.0.

**Final Proyect: Question Answering on SQuAD**

Este documento describe un proyecto que se realizó en un curso de Stanford, donde se propone una red que funciona para el aprendizaje de texto y que logra distinguir con buen *accuracy* qué preguntas son contestables a partir del texto, y responder las respectivas preguntas. Además, propone el desafío de que se modifique la red propuesta para experimentar nuevos resultados. La red propuesta consiste en un modelo de Bidirectional Attention Flow (BiDAF), que contiene una capa de *Embedding, Encoding, Attention, Modeling y Output.* La capa de embedding utiliza la técnica *Highway Network, Encoding* utiliza una LSTM bidireccional de una capa, *Attention* computa los *scores* de los pares (c,q), donde c es contexto y q la pregunta de los datos, *Modeling* contiene una LSTM bidireccional de dos capas, y *Output* produce un vector de probabilidades.

**Exploración de la Base de Datos**

Notemos que el archivo lo abrimos de la siguiente manera:



file[‘version’] nos entrega ‘v2.0’ que es la versión correspondiente a la entrega.

Por otro lado, definimos el archivo data:

data = file[‘data’]

Este archivo data tiene 442 elementos.

Sea ‘i’ en {0, … , 441} entonces tenemos que data[i] tiene dos llaves una correspondiente a ‘tittle’ la cual indica sobre qué tema trata ese elemento como por ejemplo:

* Beyonce
* Ipod

Por otro lado, data[i]['paragraphs'] tiene 66 elementos los cuales todos están relacionados con el título mencionado anteriormente.

Así, sea ‘j’ en {0, … , 65} tenemos que data[i]['paragraphs'][j] tiene dos llaves una correspondiente al contexto que es por ejemplo:

*‘Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knwles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".’*

Y hay otra llave que es ‘qas’ donde data[i]['paragraphs'][j][‘qas’] es una lista de 15 elementos. Así sea t en {0, … , 14} tenemos que:

data[i]['paragraphs'][j][‘qas’][t] tiene una llave que es ‘question’ que corresponde a la pregunta específica y otra que es ‘answer’ que corresponde a una lista de 0 elementos en donde ese único elemento es un diccionario que en la llave ‘text’ esta la respuesta específica.

**Objetivos del Trabajo**

Se propone implementar las técnicas descritas en la revisión bibliográfica, para luego realizar variaciones de éstas, junto con un análisis de los resultados y compararlos. Las variaciones que se realizarán serán:

1. En el preprocesamiento de los datos: el cual incluirá nuevas técnicas de *embeddings*.
2. En los hiperparámetros: para evaluar cómo influyen en el desempeño de la red.
3. En el *output* de la red: para evaluar si nuevas ideas funcionan para realizar las respuestas del texto, como establecer umbrales en las probabilidades de las respuestas para determinar si la pregunta era respondible o no.

**Metodología**

La metodología que se seguirá será la de iterativamente realizar cambios a partir del código implementado de la bibliografía y evaluar y analizar resultados y desempeños. Para dicho propósito, se seguirá en búsqueda de bibliografía para entender y aplicar el estado del arte en este problema, se implementarán las nuevas ideas que se adquieran a partir de la revisión, y se analizarán los resultados, los cuales serán comparados con los de la red original respecto a su desempeño.

**Modelos propuestos**

Además del modelo presentando en el proyecto final del curso de Stanford, también se proponen los siguientes modelos:

1. Dos redes neuronales. Una que determine si una pregunta es respondible o no, y si es respondible, dar una respuesta.
2. Una red neuronal que cuando la pregunta no sea respondible que responda “No hay respuesta”.
3. El modelo propuesto en el proyecto de Stanford.