PLAN DE TESIS

Índice

1. Objeto y área de la tesis

Esta tesis se encuentra enmarcada en el área del procesamiento del lenguaje natural, que estudia las técnicas computacionales para realizar tareas que contengan texto escrito en alguna lengua humana, de manera automática. El objetivo de la misma será estudiar e implementar algoritmos para resolver tareas que involucran procesar lenguaje en idioma español y que además requieren una comprensión profunda del idioma para ser resueltas. Además, se pretende desarrollar métodos de evaluación del desempeño de estos algoritmos y la capacidad de ellos de ser utilizados en diferentes tareas para las que no fueron entrenadas originalmente.

2. Introducción

El procesamiento del lenguaje natural estudia los métodos computacionales con los cuales una computadora es capaz de realizar correctamente una tarea que implica trabajar con lenguaje humano en algún nivel. Existen actualmente muchas tareas que tienen gran interés comercial, algunas de las cuales pueden ser:

- La identificación del género literario al que pertenece una síntesis de un libro.
- La predicción de una enferemedad dada una descripción de síntomas de un paciente.
- La detección de opiniones positivas o negativas en una calificación de un producto en una página web como Mercado Libre.
- La asistencia virtual al usuario de un sistema operativo como Apple o Android.

Si bien existe mucho esfuerzo en la actualidad por resolver este tipo de tareas de una manera satisfactoria, la mayor parte del trabajo se realiza en texto en idioma inglés, dando lugar a una falta de bases de datos en español (sobre todo, español latinoamericano) y de algoritmos con un buen desempeño en este idioma. Más aún, por más que la tarea se realice en forma satisfactoria, muchas veces se desea evaluar el nivel de comprensión de lenguaje que es capaz de desarrollar el algoritmo al momento de resolver el la tarea. Por ello, un problema adicional a considerar es el diseño del proceso de evaluación de estos algoritmos, lo cual generamente implica (entre otras cosas) crear bases de datos en español que permitan identificar de manera no sesgada el nivel de comprensión del algoritmo sobre el problema.

Motivados por esta situación, el objetivo de la tesis será el estudio de los algoritmos de procesamiento del lenguaje en español que requieren una comprensión profunda del mismo para resolver una determinada tarea. Para ellos será necesario crear bases de datos propias en este idioma y desarrollar un procedimiento de evaluación de dichos algoritmos para mostrar su nivel de comprensión en la tarea.

3. Desarrollo previsto en la tesis

3.1. Teoría, enfoque y métodos a utilizar

3.2. Estudios conexos

El alumno ya tiene aprobadas las siguientes materias relacionadas con los temas involucrados en la tesis:

- Probabilidad y Estadística (81.04)
- Señales y Sistemas (86.05)
- Procesos Estocásticos (86.09)
- Teoría de Detección y Estimación (86.55)
- Procesamiento del Habla (86.53)
- Procesamiento de Señales I y II (86.51 y 86.52)
- Redes Neuronales (86.54)
- Teoría de la Información y Decodificación (86.11)
- Procesamiento de Imágenes (86.56)

Además, el alumno aprobó el curso "N1: Procesamiento del lenguaje natural con redes neuronales"dictado por la Escuela de Ciencias Informáticas (ECI) en 2019, y entre los años 2019 y 2020 se desarrolló como becario del Laboratorio de Procesamiento de Señales en Comunicaciones en el proyecto titulado .ªplicaciones de procesamiento de lenguaje natural (NLP) en el contexto de Internet de las Cosas (IoT)", como parte del Programa de Becas de la Sectretaría de Ciencia y Técnica de la UBA.

3.3. Alcance de la tesis