Instituto Politécnico Nacional

Centro de Investigación en Computación

Titulo de la Tesis

T E S I S

que para obtener el título de:

XXXX XXXX XXXXX

presenta:

Xxxxxxx Xxxxxxx Xxxxxxxx Xxxxx

Tutores:

DR. Xxxxxxx Sxxxxx
Dr. XXXXXxxx XXxxxx

Estados Unidos Mexicanos
Ciudad de México
2021

Índice general

Índice de figuras

Notación

Introducción

Capítulo 1

Marco teórico

1.1. Variación Dialectal del Lenguaje

Citar a de Saussure (1945), Coseriu y Montes:

La Lengua cambia para adaptarse a las necesidades y costumbres de sus hablantes.

- Cambia al igual que la cultura en varios ejes: temporal, geográfico, y social.

- El cambio puede darse en cualquier nivel de la lengua: fonético, fonológico, morfológico, sintáctico, semántico y pragmático.

- Un cambio en un nivel puede desencadenar cambios en los otros.

[Párrafo: Lenguage, lenguage estándar, y norma] Copiar y modificar del marco teórico de la tesis de maestría.

[Párrafo: Variación dialectal. Cuando hay diferencias entre la manera de usar el lenguaje en dos regiones/ poblaciones. Fonético, semántico]

[Párrafo: Teoría del language continum?, fronteras dialectales]

[Párrafo: Conexión entre variación del lenguaje y geografía]

1.2. Representaciones vectoriales de las palabras

[Párrafo de teorías distribucionales, la carreta de Bloomfield vs. Chomsky...]

[Párrafo de antecedentes: VSM, LSI, LSA, Random Indexing]

[Párrafo de word2vec, GloVe, FastText y (nuestro alcance va hasta aquí)]

[Párrafo de representaciones contextuales, Elmo, Bert, ... (aquí ya no vamos)]

1.3. Traducción Léxica Automática

[Párrafo: embeddings multilingUes] [Objetivo: mostrar como las estructuras semanticas de los embeddings son tienen estructuras similares en diferentes lenguas]

1.4. Embeddings Condicionales

[]

1.5. Análisis de la variación del lenguaje en el espacio geográfico

[HSIC y tu tesis]

Capítulo 2

Estado del arte

Capítulo 3

Avances

3.1. Redes sociales, especialmente Twitter como fuente de datos geolingísticos

3.1.1. Metadatos generales de Twitter

Metadatos geográficos de Twitter

3.2. Descripción de los datos recogidos

Los nombres propuestos para los corpus no son los oficiales, es importante consultar como llamar y sobretodo como citar el segundo corpus

3.2.1. Corpus de Tweets panispánicos

3.2.2. Corpus de Tweets geolocalizados de Claremont-Riverside

Capítulo 4

Problema de Investigación

Durante la maestría se trabajó el cambio lingüístico a nivel léxico sin profundizar en otros niveles de la lengua que tambien se pueden explorar con datos de texto (Niveles morfo-fonológico, morfo-sintáctico,). Se inició una exploración inicial del nivel semántico mediante la creación de WordEmbeddings y se realizaron pequeños experimentos de analogías geográficas usando topónimos como miembos de la analogía, pero no se exploraron las posibilidades de la información geográfica.

Nota: Revisar esta sugerencia de Segun: FastText Word Embeddings for Spanish Language Variations

[Objetivo:]

Referencias

de Saussure, F. (1945). *Curso de Lingüística General*. Buenos Aires: Losada. Descargado de https://archive.org/details/saussure-ferdinand-curso-de-linguistica-general/page/n1/mode/2up