Dr. Luis Guillermo Ruíz Velázquez

Doctor en Ciencias en Ingeniería Eléctrica Opción en Sistemas Computacionales

Ninguna

Datos de contacto:

luis.ruiz@infotec.mx

Modelo de lenguaje para el español sensible a la región usando aprendizaje profundo.

Datos Generales

Resumen ejecutivo

Se creará un modelo de lenguaje usando aprendizaje profundo que identifique la forma de hablar de los diferentes países que hablan el idioma español.

Línea General y Aplicación del Conocimiento (LGAC)

10. Inteligencia computacional en la Ciencia de Datos

Palabras clave

Modelo de Lenguaje en español, Aprendizaje profundo, Inteligencia Artifical

Objetivo General

Crear un Modelo de Lenguaje para el español que tome en cuenta las diferentes formas de hablar en los países que tienen al español como idioma oficial. El modelo debe ser lo suficientemente grande para que capture las diferencias entre las regiones pero de un tamaño aceptable para que pueda ser usadon en computadoras no especializadas. La arquitectura del modelo deberá ser flexible para que los usuarios la puedan modificar y lo puedan adaptar a una tarea en específico mediante la técnica de entrenamiento fine tunning. Se pretende poner un modelo al alcance de todos. Este modelo para un usuario común sería muy tardado o costoso de entrenar. Nuestro modelo será entrenado con unas 100 gigas de tweets seleccionados en español. Se espera que su entrenamiento dure 4 meses en una computadora especializada con 2 GPU's de 24 GB de memoria cada una. Al final, el modelo quedará disponible en un repositorio público de fácil acceso e instalación. El modelo tendrá la capacidad de responder la tarea de enmascarado de palabra, que dada una frase, se oculta o enmascara una palabra y el modelo debe predecirla. También se tendrá la capacidad de obtener los vectores de encaje de una frase y así poder usarlos para responder otras tareas como clasificación de texto, análisis de sentimiento, búsquedas por similaridad, recomendaciones, entre otras.

Objetivos especificos

Poner a disposición del público en general, modelos de Procesamiento de Lenguaje Natural (PLN) entrenados para el español que puedan ser usados en equipos de cómputo de uso común. Que los modelos puedan ser usados para las tareas de predicción de texto, análisis de sentimiento y como un generador de vectores de encaje. Los modelos podrán responder en las variaciones del español de otros países con gran número de personas hispanohablantes.

Datos del proyecto

Descripción

Una de las cualidades más apreciadas de los seres humanos es la comunicación; realizamos intercambio de ideas por medio de textos, señales, imágenes, expresiones faciales, entre otros. La principal forma en la cual nos comunicamos es el lenguaje. Ejemplos de lenguajes que surgieron y evolucionaron de manera natural podrían ser el inglés, el náhuatl o el castellano. Se cree que el lenguaje humano, tal y como lo conocemos, apareció hace unos cien mil años y gracias a él, las ideas pueden perdurar, transmitirse, evolucionar y moldear sociedades complejas. El lenguaje es una pieza clave como sistema de comunicación; sin embargo, entenderlo es una tarea sumamente compleja, incluso para los seres humanos, pues el significado de una frase puede depender de la región geográfica, el tono de la voz, el contexto, entre muchos otros factores. Esto se vuelve evidente cuando comparamos la forma de hablar entre diferentes países que hablan el mismo idioma, como es el caso del Español en Latinoamérica. El estudio de los lenguajes es el objetivo de la rama de la computación llamada Procesamiento del Lenguaje Natural. Las redes neuronales artificiales han sido empleadas en PLN con gran éxito creando modelos para encajes de palabras (word embeddings), que pueden ser usados en tareas como traducción, creación de resúmenes, respuesta a preguntas, entre muchas otras. Estas técnicas tienen dos inconvenientes principales: requieren una gran cantidad de datos y de poder de cómputo. Otro inconveniente es la falta de recursos en lenguajes diferentes del inglés. En particular, existen recursos para el español, sin embargo, no existen para las diferentes variantes que se hablan en el mundo, y en particular, los regionalismos no se intentan capturar de manera explícita por los creadores de modelos. Con este proyecto, se pretende crear y poner a la disposición de la comunidad modelos entrenados para diferentes variantes del español habladas en el mundo. Adicionalmente, se incluirán librerías para un fácil uso de los modelos, aún para usuarios poco experimentados. Para esto se requiere completar vartios pasos. Primero, obtener el conjunto de datos de entrenamiento, limpiarlo y acondicionarlo para que pueda ser usado por los algoritmos de aprendizaje profundo y puedan aprender de ellos. Después, viene el diseño e implementación de la estructura de la red neuronal. El tercer paso es el entrenamiento donde se deben usar recursos sufucientes de cómputo para el análisis de los datos y así obtener el modelo final. El último paso es la evaulación donde se comprueba la utilidad del modelo.

Antecedentes del problema a resolver

En la actualidad existen modelos pre-entrenados en grandes volúmenes de texto que pueden ser descargados y usados en alguna tarea específica. Uno de los más populares es BERT (Devlin et al.), que es un modelo pre-entrenado para predecir una palabra dentro de una oración. Desafortunadamente, BERT sólo está disponible en el idioma Inglés por lo que no puede ser usado para texto en español. BETO (Cañete et al.) es un modelo basado en BERT pero entrenado en una colección grande de texto en Español. La principal diferencia con BETO es que nuestro modelo será entrenado con mensajes marcados con una etiqueta especial para distinguir su país de origen, por lo que el modelo final incluirá información regional. Como parte de mi experiencia en el desarrollo de estos modelos de lenguaje, actualmente cuento con una librería para implementar la aqrquitectura de los Transformers, incluída la de los modelos tipo BERT. Esta librería la escribí usando las funciones de TensorFlow y la he probado en tareas como Image Captioning donde la entrada es una imagen que se quiere describir usando lenguaje natural. Como antecedente, ya tenemos una serie de modelos similares llamados BILMA (Bert In Latin America) (Tellez et al.) que creamos con un grupo de investigadores de INFOTEC y CentroGeo. Los modelos están disponibles de manera libre mediante la plataforma PIP de instalación de paquetes y se cuenta con un artículo de revista publicado. La diferencia del modelo aquí propuesto con los de BILMA es que se pretende tener un modelo único que contenga el conocimiento de todas las variantes del español. Se podrá elegir la variante mediante palabras clave adicionales que se incluirán en las frases de entrada. Con esto, en lugar de descargar un modelo para cada región, los usuarios sólo necesitarán de uno solo para todas sus necesidades.

Justificación y pertinencia

En los últimos años han surgido modelos de lenguaje que han ganado popularidad gracias a su gran desempeño. Modelos como (Petters et al.),), (Devlin et al.) o (Brown et al.) han sido ampliamente usados en gran variedad de áreas y tareas. Estos modelos son de gran tamaño y fueron entrenados en clusters con GPU’s que solo unos pocos tiene acceso. La principal limitante para Latinoamérica es que estos modelos están entrenados para el idioma Inglés, es por eso que hay un mercado enorme para modelos de este tipo en Español. El Español es el segundo de los idiomas más populares del mundo con unos 489 millones de hablantes. Es el idioma oficial en 21 países, es por eso que contar con herramientas y modelos que puedan analizar de manera automática mensajes puede ayudar en una gran variedad de tareas. En cuanto a las redes sociales, Twitter era la que más es usada en el mundo del PNL debido a su API, su importancia y la gran cantidad de texto que se genera diariamente. Por estas razones, un modelo en Español entrenado en la red social Twitter es de gran trascendencia para la comunidad científica y el público en general. El análisis de tweets pude ser usado por empresa o instituciones para dar seguimiento a quejas o problemas que tienen sus usuarios o el público en general. Pueden ser usados para análisis de opinión o detectar mensajes violentos o discriminatorios. Actualmente existe un gran interés entre la comunidad científica mexicana por tener un modelo de lenguaje que conozca las particularidades del español. Tenemos confirmado el interés de colaboración de investigadoras e investigadores del CentroGeo, CIMAT, INAOE y la UNAM.

Metas

Abonar a la oferta de modelos pre-entrenados y ser un referente en cuanto al idioma español con un modelo novedoso, de fácil acceso y con herramientas que permitan su manipulación sencilla. Crear una red de colaboración con colegas de otras instituciones que comparten el mismo interés por el desarrollo de la investigación de PLN en español. Que el modelo creado se use como una herramienta para futuras investigaciones en una amplia variedad de campos.

Metodologías

La arquitectura Representación de codificador bidireccional de transformadores (BERT, por sus siglas en inglés) (Devlin et al.), ha demostrado su efectividad en la clasificación y generación de texto. A pesar de que es muy popular, esta arquitectura se proporciona principalmente pre-entrenada en el idioma Inglés, y aunque existe su variante en Español llamada BETO (Cañete et al.), no contempla la gran variedad geográfica del Español. En este proyecto, se usará la arquitectura BERT para diseñar una red neuronal usando la librería TensorFlow que es de libre acceso. El código se escribirá usando el lenguaje Python. Para el entrenamiento se usará la gran colección de texto con la que actualmente cuenta el INFOTEC que proviene de sitios públicos como Twitter (ahora X) y Wikipedia. Para esto se hará uso de la técnica de enmascarado de palabras que es el estándar para el entrenamiento de este tipo de modelos. Debido al gran número de parámetros de la red neuronal (unos 40 millones) y a la enorme cantidad de texto de entrenamiento, el uso de tarjetas gráficas especializadas se vuelve indispensable. Para la evaluación se le pedirá al modelo que prediga la zona de origen del texto, por lo que el modelo deberá aprender las diferentes formas de hablar el español en el continente americano. Otra tarea a evaluar será la de clasificación o análisis de sentimiento. En ésta se usarán los vectores de encaje del modelo para la clasificación del texto en sentimientos como alegría, enojo, tristeza, entre otras.

Resultados esperados

Tener un modelo de lenguaje natural para el español que esté disponible para cualquier persona interesada, sea experta o no. Incluirá herramientas para un fácil uso y manipulación. Estará disponible en un repositorio público junto con un manual de uso. Se escribirá y someterá un artículo de revista para su publicación con los datos del modelo, su evaluación y el conjunto de datos usado para su entrenamiento. Consolidar la colaboración con investigadoras e investigadores de otros centros de investigación como lo son CentroGeo, INAOE o IIMAS.

Cronograma de trabajo

#	Entregable(s) comprometido(s)	Fecha inicio	Fecha fin
1	**Artículo de investigación en revista especializada**	**01/10/2024**	**15/12/2024**
2	**Impartición de docencia.**	**01/08/2024**	**31/12/2024**
3	**Participación en comité tutoral, dirección o codirección de trabajo de titulación.**	**01/04/2024**	**31/10/2024**
4	**Difusión**	**01/05/2024**	**31/08/2024**

Bibliografía relevante

Brown, Tom, et al. “Language Models are Few-Shot Learners.” Advances in Neural Information Processing Systems, vol. 33, 2020, 1877–1901. Cañete, José, et al. “Spanish Pre-Trained BERT Model and Evaluation Data.” PML4DC at ICLR 2020, 2020. Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1, vol. 1, 2019, 4171–4186. Petters, Matthew E., et al. “Deep Contextualized Word Representations.” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)", 2018, 2227–2237. Tellez, E.S., Moctezuma, D., Miranda, S. et al. “Regionalized models for Spanish language variations based on Twitter.” Lang Resources & Evaluation 57, 1697–1727 (2023).