Estadística multivariada y datos categóricos
Resumen - Similitud semántica

Héctor Efrén Juárez Guadarrama 195699
Miguel López Cruz 197967
Eduardo Moreno Ortiz 151280

24 de mayo de 2021

Profesor: Alfredo Garbuno Iñigo

La **similitud semántica** es una tarea que utiliza un conjunto de documentos u oraciones, donde la idea de semejanza entre elementos se basa en la similitud de su significado (contenido semántico) en contraste a la similitud lexicográfica. Utiliza un conjunto de herramientas matemáticas que miden la semejanza del significado entre oraciones, basándose en una descripción numérica que se obtiene con la comparación de su significado. A lo largo de su historia, esta tarea ha evolucionado y no solamente permite medir la similitud entre 2 textos, sino que se ha desarrollado un score de "similitud".

A pesar de que el humano no tiene una definición formal de la similitud entre conceptos, es capaz de ser juez y definir la relación entre conceptos. Por ejemplo, un niño (probablemente) puede decir que manzana y pera están más relacionados que manzana y calabaza.

BERT (Bidirectional Encoder Representations from Transformers) es un modelo bidireccional de representación del lenguaje natural que permite representaciones bidireccionales profundas de texto sin etiquetas tanto en el sentido izquierda-derecha como derecha-izquierda, proveyendo al modelo una visión más amplia del entorno de cada palabra enriqueciendo su significado en base al contexto en el que se encuentra. Antes de introducir el modelo BERT, es muy importante mencionar que se ha demostrado que una fase de pre-entrenamiento de modelo de lenguaje tiene un efecto positivo al momento de realizar tareas de lenguaje natural para ello existen 2 estrategias:

- Estrategia basada en características
- Estrategia de afinación

En la **estrategia basada en características** se logra extraer "insights" del documento a analizar y en la **estrategia de afinación** se afinan todos los parámetros pre-entrenados. Sin embargo, éstas 2 estrategias se basan de modelos unidireccionales de lenguaje natural teniendo fuertes limitaciones en su desempeño. Es aquí donde BERT tiene una ventaja competitiva muy fuerte, ya que, gracias al modelo de lenguaje oculto (masked language model MLM) es posible obtener una estructura bidireccional. La estrategia MLM funciona de la siguiente manera:

- Se selecciona aleatoriamente un porcentaje FIJO de tokens, los cuales serán ocultos y el objetivo del modelo será predecir tales tokens.

Dicha estrategia permite utilizar tanto la información obtenida de la dirección izquierda-derecha, así como de la dirección derecha-izquierda permitiendo pre-entrenar un transformador bidireccional profundo.
Como se mencionó previemante, el pre-entrenamiento general de representación de lenguaje ha permitido considerables mejoras al análisis de lenguaje. Dentro de las aproximaciones más utilizadas para esta etapa se encuentran:

- Aproximaciones basadas en características no supervisadas

- Aproximaciones de afinación no supervisada

En la primera de ellas, se utiliza un modelo unidireccional de izquierda a derecha, así como el uso de la información tanto de la izquierda como de la derecha para la elección de palabras correctas. La segunda de las estrategias, permite que pocos parámetros sean aprendidos desde cero, lo cual es una ventaja computacional. Para la implementación del modelo BERT es necesario de evaluar 2 pasos:

1. Pre-entrenamiento
2. Afinación

En la primera etapa se obtiene una configuración de parámetros del modelo usando datos no etiquetados y en la segunda se afinan dichos parámetros con datos etiquetados. Cabe mencionar que BERT se basa en una arquitectura de un codificador bidireccional multicapa, lo cual permite que entre ambas etapas exista una conexión y no sean 2 procesos totalmente independientes.
La estructura de los datos que alimenta BERT es específica, pues permite identificar una oración o par de oraciones en una sola secuencia de token. En específico, en este proyecto se tiene la siguiente estructura:

1. El primer token de cada secuencia siempre es un toke especial de clasificación [CLS], que es utilizada en la última capa oculta para la tarea de clasificación
2. Las oraciones son separadas con el token especial [SEP]
3. Se añade un elemento para definir si el token pertenece a la oración A o a la oración B.

Como se mencionó previamente, una de los elementos más importantes de BERT es que es un modelo bidireccional, es por ello que no puede ser pre-entrenado con modelos tradicionales unidireccionales, por lo que es necesario de 2 tareas no supervisadas:

1. MLM: gracias a esta tarea (ocultar tokens aleatoriamente), es posible que el modelo observe un panorama más amplío del target a predecir. Sin embargo, hay que ser cuidadosos en esta tarea, ya que, se puede romper la conexión entre el pre-entrenamiento y la afinación. Es por ello que se recomienda realizar las siguientes tareas para evitar dicha desconexión:
   (a) Seleccionar aleatoriamente los tokens a ocultar
   (b) Cada token oculto es remplazado con un token aleatorio (10% de las veces), reemplazarlo por el token mismo (10% de las veces), reemplazarlo con el [MASK] (80% de las veces)
   (c) Utilizar la penúltima capa par predecir el token oculto utilizando la función de pérdida entropía cruzada.

2. NSP (Next Sentence Prediction - predicción de la siguiente oración)

Finalmente, para la segunda tarea (la afinación de parámetros) BERT usa el mecanismo de atención para conectar las tareas de codificar independientemente pares de texto y aplicar la atención cruzada bidireccinal.
Para aplicaciones que involucran pares de texto una ruta común a seguir es codificar independientemente pares de texto antes de aplicar la atención cruzada bidireccional, en lugar de ello BERT usa el mecanismo de propia atención para conectar ambas tareas, esto es, codificando texto en pares concatenando con atención propia incluyendo el mecanismo de atención propia entre ambas oraciones.