

# ¿Modelos de lenguaje como bases de conocimiento?

## 1. Introducción.

En el campo del Lenguaje Natural (NLP), los modelos de lenguaje preentrenados como ELMo y BERT tienen un papel más importante. Estos modelos, optimizados para predecir palabras en secuencia, almacenan vastas cantidades de conocimiento lingüístico y han demostrado ser fundamentales para el rendimiento sobresaliente en diversas tareas.

## 2. Background.

### 2.1 Modelos de Lenguaje Unidireccionales

Dada una secuencia de tokens de entrada  $w = [w_1, w_2, \dots, w_N]$ , los modelos de lenguajes unidireccionales comúnmente asignan una probabilidad  $p(w)$  a la secuencia mediante la factorización:  $p(w) = \prod_i p(w_i | w_{1:i-1}, \dots, w_i)$

Modelos Considerados:

- Fairseq-fconv: Utiliza convoluciones en capas en lugar de redes neuronales recurrentes (RNN) tradicionales. Se emplea el modelo preentrenado en la biblioteca fairseq, entrenado en el corpus WikiText-103.
- Transformer-XL: basado en el modelo Transformer, puede tener en cuenta una historia más larga mediante el almacenamiento en caché de salidas anteriores y el uso de codificación posicional relativa. Alcanza una perplexidad de prueba de 18.3 en el corpus WikiText-103.

### 2.2 Modelos de lenguaje Bidireccionales

Dada una secuencia de entrada  $w = [w_1, w_2, \dots, w_N]$  y una posición  $1 \leq i \leq N$ , queremos estimar  $p(w_i) = p(w_i | w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_N)$  utilizando el contexto izquierdo y derecho de esa palabra.

Modelos Considerados:

- ELMo: utiliza una arquitectura de LSTM con múltiples capas.
- BERT: propone muestrear posiciones en la secuencia de entrada al azar y aprende a llenar la palabra en la posición enmascarada.

## 3. Trabajo Relacionado.

Es necesario investigar en qué medida los modelos de lenguaje preentrenados retienen conocimientos factuales y de sentido común. A diferencia de investigaciones anteriores que se centran en propiedades lingüísticas, este estudio compara los modelos con bases de conocimiento simbólicas.

Se mencionan trabajos previos que evalúan representaciones de palabras y oraciones, así como la capacidad de transferencia a tareas específicas. Se destacan resultados contradictorios sobre la comprensión sintáctica de los modelos. Se excluyen modelos específicos, como GPTv2, debido a limitaciones o falta de disponibilidad de datos.

#### 4. Sonda LAMA (Análisis del Modelo de Lenguaje)

Se usará de ejemplo la Sonda LAMA para evaluar el conocimiento factual y de sentido común en modelos de lenguaje. Utiliza fuentes de conocimiento como triples sujeto-relación-objeto o pares pregunta-respuesta, convirtiéndose en declaraciones de eliminación para buscar tokens faltantes en el modelo de lenguaje.

- Fuentes de conocimiento: se descubren diversas fuentes utilizando modelos de lenguaje preentrenados como fairseq-conv y BERT-large. Se genera un vocabulario común para una comparación justa.
- Baselines: se comparan los modelos con enfoques convencionales como Freq, RE y DrQA.
- Métricas: se emplean métricas basadas en el rango, calculado la precisión media en K ( $P@K$ ) para evaluar la clasificación de objetos correctos por parte del modelo.
- Consideraciones: se tomaron decisiones de diseño, como el uso de plantillas definidas manualmente, la limitación a objetos de un solo token, y la consulta solo de espacios de objeto en triples. Se utiliza un vocabulario común para una comparación justa entre modelos.

#### 5. Discusión y Conclusiones.

En este estudio se analizó el conocimiento factual y de sentido común en modelos de lenguaje preentrenados, resaltando BERT-large al recordar este conocimiento de manera competitiva con enfoque no neurales y supervisados. Los modelos entrenados en grandes corpus podrían ser alternativas efectivas a las base de conocimiento generales. Además, es importante señalar la necesidad de investigar la variabilidad en la recuperación del conocimiento con respecto a diversas plantillas de lenguaje natural.

Rios Sicaivos Andrea