

¿Modelos lingüísticos como bases de conocimiento?

Las bases de conocimiento son soluciones eficaces para acceder a datos relacionales anotados y estándar al permitir consultas como (DANTE, born-in, x). Sin embargo, en la práctica a menudo necesitamos extraer datos relacionales del texto u otras modalidades para poblar estas bases de conocimiento. Esto requiere complejos de PNL que involucren extracción de entidades, resolución de correferencias, vinculación de entidades y extracción de relaciones, componentes que a menudo necesitan datos supervisados y esquemas fijos. En su lugar, podríamos intentar consultar modelos de lenguaje neuronal en busca de datos relacionales pidiéndoles que completen tokens enmascarados en secuencias.

2.1 Modelos de lenguaje unidireccional

Dada una secuencia de entrada de tokens $w = [w_1, w_2, \dots, w_n]$ los modelos de lenguaje unidireccional comúnmente asignan una probabilidad $p(w)$ a la secuencia factorizándola de la siguiente manera:

$$P(w) = \prod_{t=1}^n p(w_t | w_{t-1}, \dots, w_1)$$

Una forma común de estimar esta probabilidad es utilizar modelos de lenguaje neuronal con $p(w_t | w_{t-1}, \dots, w_1) = \text{softmax}(w_t \cdot h_t + b_t)$

donde $h_t \in \mathbb{R}^k$ es el vector de salida de una red neuronal en la posición t y $W \in \mathbb{R}^{V \times k}$ es una matriz de parámetros aprendidos que asigna h_t a puntuaciones no normalizadas para cada dificultad en el vocabulario V .

fairseq-conv: en el lugar de las redes neuronales recurrentes comúnmente utilizadas, utilizan múltiples capas de convoluciones cerradas.

Transformer-XL: introducen un modelo de lenguaje a gran escala basado en Transformer. Puede tener en cuenta un historial más largo almacenando en caché las salidas anteriores y utilizando codificación posicional relativa en lugar de absoluta.

2.2 Bidireccional "Modelos de lenguaje"

Formalmente, dada una secuencia de entrada $w = [w_1, w_2, \dots, w_N]$ y una posición $1 \leq i \leq N$, queremos estimar $p(w_i) = p(w_i | w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_N)$ usando la curva izquierda y contexto correcto de esa palabra.

ELMO: Para estimar esta probabilidad, proponen ejecutar un LSTM hacia adelante y hacia atrás, lo que da como resultado h_i y h_i que en consecuencia se utilizan para calcular la probabilidad logarítmica de un modelo de lenguaje hacia adelante y hacia atrás.

BERT: En lugar de un objetivo de modelo de lenguaje estándar, proponen muestrear posiciones en la secuencia de entrada al azar y aprender el lenguaje a completar la palabra en la posición enmascarada.

3. Trabajos relacionados

El trabajo existente se centra en comprender las propiedades lingüísticas y semánticas de las representaciones de palabras o en qué medida las representaciones de oraciones y los modelos de lenguaje previamente entrenados transfieren el conocimiento lingüístico a tareas posteriores.