BLEU

La métrica BLEU (Bilingual Evaluation Understudy) es ampliamente utilizada en la evaluación de sistemas de traducción automática, pero también puede ser adaptada para evaluar la calidad de resúmenes automáticos, tanto en enfoques extractivos como abstractivos.

¿Qué es la métrica BLEU?

BLEU es una métrica que compara n-gramas (secuencias de palabras) del texto generado por una máquina con los n-gramas de uno o más textos de referencia humanos. La puntuación BLEU varía de 0 a 1, donde 1 indica una correspondencia perfecta con el texto de referencia.

BLEU en Resúmenes Automáticos

Enfoques Extractivos

En resúmenes extractivos, el sistema selecciona frases o sentencias directamente del texto original para formar el resumen. La aplicación de BLEU aquí implica comparar los n-gramas del resumen extraído con los n-gramas de un resumen de referencia creado por humanos. Debido a que los resúmenes extractivos contienen frases exactamente como aparecen en el texto fuente, los n-gramas suelen coincidir más a menudo con los de los resúmenes de referencia, potencialmente resultando en puntuaciones BLEU más altas.

Enfoques Abstractivos

En resúmenes abstractivos, el sistema genera nuevas frases que pueden no estar presentes en el texto original, condensando y parafraseando la información. Aplicar BLEU a resúmenes abstractivos puede ser más desafiante, ya que la generación de nuevas frases puede no coincidir exactamente con las de los resúmenes de referencia, lo que podría resultar en puntuaciones BLEU más bajas, incluso si el contenido y la calidad del resumen abstractivo son buenos.

Aplicación de BLEU al Resumen Automático obtenido con TFIDF

TFIDF (Term Frequency-Inverse Document Frequency) es una técnica comúnmente utilizada para enfoques extractivos. Selecciona frases o sentencias del texto original que tienen términos significativos en función de su frecuencia.

Para aplicar BLEU a un resumen generado por TFIDF:

1. **Generación del Resumen**: Utilizar TFIDF para extraer las frases más relevantes del texto original.
2. **Referencia Humana**: Obtener uno o más resúmenes de referencia creados por humanos.
3. **Cálculo de BLEU**: Comparar los n-gramas del resumen generado por TFIDF con los n-gramas de los resúmenes de referencia.

Resultados Esperados

El resumen generado con TFIDF probablemente tendrá una alta puntuación BLEU en comparación con enfoques abstractivos debido a la mayor coincidencia directa de n-gramas con el texto original y, por lo tanto, con los resúmenes de referencia, especialmente si estos también son extractivos. Sin embargo, una alta puntuación BLEU no siempre indica un resumen de alta calidad en términos de comprensión y cohesión, ya que BLEU no evalúa la calidad semántica ni la estructura coherente del texto.

Limitaciones de BLEU

- **Dependencia de las Referencias**: BLEU depende en gran medida de la calidad y la cantidad de los resúmenes de referencia. Pocas referencias pueden no capturar toda la variabilidad posible de buenos resúmenes.
- **N-gramas Exactos**: BLEU mide la coincidencia exacta de n-gramas, lo que puede no reflejar la calidad semántica o la fluidez del resumen.
- **Evaluación de Resúmenes Abstractivos**: La generación de texto nuevo en resúmenes abstractivos puede recibir puntuaciones BLEU bajas injustamente debido a la falta de coincidencia exacta de n-gramas.

Conclusión

BLEU puede aplicarse a la evaluación de resúmenes automáticos tanto extractivos como abstractivos, pero es más adecuado para enfoques extractivos debido a la mayor coincidencia de n-gramas. Al aplicar BLEU a un resumen automático obtenido con TFIDF, es probable obtener puntuaciones relativamente altas, reflejando la similitud en la selección de frases con los resúmenes de referencia, aunque esto no necesariamente refleje la calidad integral del resumen. Para una evaluación más completa, se recomienda utilizar BLEU junto con otras métricas que evalúen la coherencia, la cohesión y la calidad semántica del resumen.