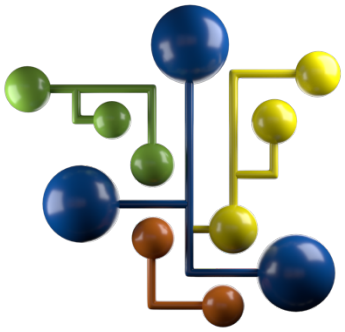

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

BLEU
Bilingual Evaluation Understudy

O Score BLEU é um modelo de avaliação dos sistemas de tradução automática. BLEU significa Bilingual Evaluation Understudy e é uma forma de avaliar automaticamente os sistemas de tradução.

Essa métrica foi introduzida pela primeira vez no documento, BLEU: A Method for Automatic Evaluation of Machine Translation, Papineni and others, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002: 311-318.

Nós vamos implementar o algoritmo de cálculo de Score BLEU em nosso modelo. Mas antes vamos entender como ele é calculado.

Vamos considerar um exemplo. Digamos que temos duas sentenças candidatas, ou seja, uma sentença prevista pelo nosso sistema MT (Machine Translation) e uma sentença de referência (ou seja, a tradução real correspondente) para alguma sentença fonte dada:

Referência 1: O gato sentou no tatame verde
Candidato 1: O gato está no tatame verde

Para ver quão boa é a tradução, podemos usar uma medida de precisão. A precisão é uma medida de quantas palavras no candidato estão realmente presentes na referência. Em geral, se você considerar um problema de classificação com duas classes (indicado por negativo e positivo), a precisão será dada pela seguinte fórmula:

$$Precision = \frac{number\ of\ samples\ correctly\ classified\ as\ positive}{all\ the\ samples\ classified\ as\ positive}$$

Vamos agora calcular a precisão para o candidato 1:

Precisão = # de vezes que cada palavra do candidato apareceu em referência / # de palavras no candidato

Matematicamente, isso pode ser dado pela seguinte fórmula:

$$Precision = \frac{\sum_{unigram \in Candidate} IsFoundInRef\left(unigram\right)}{\left|Candidate\right|}$$

Precisão para o candidato 1 = 5/6 (ou seja, 5 palavras certas em 6). Isso também é conhecido como a precisão de 1 grama, pois consideramos uma única palavra por vez. Agora vamos apresentar um novo candidato:

Candidato 2: O o o gato gato gato

não é difícil para um humano ver que o candidato 1 é muito melhor que o candidato 2. Vamos calcular a precisão: Precisão para o candidato 2 = 6/6 = 1 (embora as palavras estejam repetidas, elas estão corretas). Como podemos ver, a pontuação de precisão discorda do julgamento que fizemos. Portanto, não se pode confiar na precisão como uma boa medida da qualidade de uma tradução.

Para resolver a limitação de precisão, podemos usar uma precisão modificada de 1 grama. A precisão modificada considera o número de ocorrências de cada palavra única no candidato pelo número de vezes que essa palavra apareceu na referência:

$$p_1 = \frac{\sum_{unigram \in \{Candidate\}} Min\left(Occurences\left(unigram\right), unigram_{max}\right)}{\left|Candidate\right|}$$

Portanto, para os candidatos 1 e 2, a precisão modificada seria a seguinte:

Mod-1-gram-Precision Candidate 1 = (1 + 1 + 1 + 1 + 1) / 6 = 5/6
Mod-1-gram-Precision Candidato 2 = (2 + 1) / 6 = 3/6

Já podemos ver que esta é uma boa modificação, pois a precisão do candidato 2 é reduzida. Isso pode ser estendido para qualquer n-grama considerando n palavras de cada vez, em vez de uma única palavra.

Brevity penalty

A precisão naturalmente prefere sentenças pequenas. Isso levanta uma questão na avaliação, pois o sistema MT pode gerar pequenas sentenças para referências mais longas e ainda ter uma precisão maior. Portanto, a penalidade de brevidade é introduzida para evitar isso. A penalidade de brevidade é calculada pelo seguinte:

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}.$$

Aqui, c é o comprimento da sentença candidata e r é o comprimento da sentença de referência. No nosso exemplo, calculamos como mostrado aqui:

$$\text{BP for candidate 1} = e^{(1-(6/6))} = e^0 = 1$$

$$\text{BP for candidate 2} = e^{(1-(6/6))} = e^0 = 1$$

Finalmente, o Score BLEU

Em seguida, para calcular o Score BLEU, primeiro calculamos várias precisões de n-grama modificadas para um grupo de valores diferentes. Em seguida, calcularemos a média geométrica ponderada das precisões de n-grama:

$$BLEU = BP \times exp\left(\sum_{i=1}^{N} w_n p_n\right)$$

Aqui, wn é o peso para a precisão de n-grama modificada pn. Por padrão, pesos iguais são usados para todos os valores de n-grama. Em conclusão, o BLEU calcula uma precisão de n-grama modificada e penaliza a precisão de n-grama modificada com uma penalidade de brevidade. A precisão n-gram modificada evita valores potenciais de alta precisão dados a sentenças sem sentido (por exemplo, candidato 2).

Logo, essa é uma boa métrica para nosso NMT. Agora vejamos a implementação com o TensorFlow.