Resumo: BLEU: a Method for Automatic Evaluation of Machine Translation

De Matheus Loiola Pinto Curado Silva

5 de outubro de 2023

Introdução

A avaliação humana nas traduções de máquina pesam em inúmeros aspectos, como em fidelidade, fluência e a adequação da frase com a original. Entretanto, esse processo pode demorar semanas ou meses para terminar, e isso é um problema pois os desenvolvedores das máquinas de tradução precisam monitorar seus resultados e suas mudanças de forma diária, para evitar podar as ideias ruins e melhorar o sistema.

Porém, como uma pessoa mede a qualidade de uma tradução? Quanto maior perto a tradução da máquina estiver de uma tradução profissional humana, melhor. Para isso, é necessário verificar o quão perto ela está da tradução original pela quantidade de palavras semelhantes entre eles.

The Baseline BLUE Metric

Existem várias traduções perfeitas de alguma sentença, e elas podem variar na escolha de palavras.

Assim, a principal tarefa de um implementador BLEU é comparar *n-grams* de traduções candidatas com *n-grams* da tradução de referência e contar o número de igualdades. Quanto mais igualdades, maior é seu rank entre os candidatos.

Uma sentença candidata não deve ser nem tão longa, nem tão curta, e a validação da métrica deve verificar isso. Caso a sentença seja maior que a sentença de referência, essa sentença logo é penalizada pela métrica de precisão do *n-gram*, então não é necessária penalizar de novo.

A avaliação BLEU

A métrica BLEU varia entre 0 e 1. Poucas traduções vão atingir o escore máximo a não ser que sejam idênticos a sentença de referência.

Conclusão

BLEU irá acelerar a área de pesquisa e desenvolvimento em máquinas de tradução. A força do BLEU é que ele se correlaciona altamente com os julgamentos humanos, fazendo a média dos erros de julgamento de sentenças individuais sobre um corpus de teste, em vez de tentar adivinhar o julgamento humano exato para cada sentença: a quantidade leva à qualidade.