ROUGE: A Package for Automatic Evaluation of Summaries

De Matheus Loiola Pinto Curado Silva

Introdução

ROUGE significa *Recall-Oriented Understudy for Gisting Evaluation*. Ele inclui medidas para determinar automaticamente a qualidade de um resumo ao comparar com outros resumos criados por humanos. A medição conta o número de unidades que se sobrepõem, como n-gramas, sequência de palavras e a correspondência entre pares de palavras geradas por computador.

O artigo introduz quatro diferentes tipos do ROUGE: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S.

ROUGE-N

ROUGE-N é um recall de n-grama entre o resumo candidato com os resumos referências. Essa métrica possui uma fórmula que tende muito ao recall.

Ao controlar as referências que adicionamos na métrica, é possível avaliar diferentes aspectos de resumos. Isso dá mais peso ao realizar a correspondência entre n-grams em referências múltiplas. Dessa forma, essa métrica favorece resumos que são mais similares no consenso geral.

ROUGE-L

Essa métrica é baseada na LCS (Longest Common Subsequence). Ela possui os fatores de precisão e recall, e estudos demonstram que a lógica dessa métrica produziu estimativas tão boas quanto o BLEU.

Uma vantagem de usar a LCS é que não é necessário correspondências consecutivas, mas sim em sequência. A outra vantagem é automaticamente incluir a maior sequência comum de n-gramas.

Por recompensar ocorrências em sequência, ROUGE-L também captura a estrutura de nível de sentença de uma maneira natural.

Uma desvantagem é que a LCS apenas conta a principal sequência de palavras. Dessa forma, sequências menores ou outras LCSs não são refletidas no score final.

ROUGE-W

ROUGE-W significa *Weighted Longest Common Subsequence*. A LCS vista anteriormente possui o problema de não diferenciar LCSes de outros tamanhos ou sequências. Dessa forma, foi utilizado o algoritmo de Programação Dinâmica para lembrar do tamanho de ocorrências consecutivas em uma matriz 2D, e utilizar ela para o cálculo das métricas.

ROUGE-S

Skip-bigram é qualquer par de palavras em ordem de sentença, permitindo espaços arbitrários. "Skip-bigram co-occurrence statistics"medem a sobreposição de skip- bigrams entre a tradução candidata e o conjunto de traduções referência.

Comparando skip-bigram com a LCS, o skip-bigram conta todas as palavras correspondentes em ordem enquanto a LCS conta apenas a subsequência comum mais longa.

ROUGE-SU

ROUGE-SU é uma extensão do ROUGE-S. A ROUGE-S tem um problema que é não dar qualquer crédito a uma sentença candidata se a sentença não tem nenhum par de palavras co-ocorrendo com suas referências. Para contornar isso, o ROUGE-S foi estendido com a adição de unigramas como uma unidade de contagem.

Conclusões

ROUGE-2, ROUGE-L, ROUGE-W E ROUGE-S funcionaram bem em tarefas de resumo de um único documento. ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4 e ROUGE-SU9 performaram bem em resumos pequenos (estilo resumos de *headlines*.

Em um estudo separado, ROUGE-L, ROUGE-W e ROUGE-S se mostraram muito efetivos na avaliação automática de tradução de máquina.