METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments

De Matheus Loiola Pinto Curado Silva

Introdução

O artigo descreve METEOR, uma métrica automática para avaliação de tradução de máquina que é baseada em um conceito generalizado de casamento de unigramas entre traduções de máquina e traduções de humano.

A métrica precisa ser consistente, confiável e generalizada. Dessa forma, METEOR foi desenvolvida para consertar inúmeras fraquezas na métrica BLEU. Essa métrica é baseada em um casamento de palavra por palavra e uma ou mais traduções referência. Essa métrica não só faz o casamento de palavras idênticas, mas como também faz o casamento entre palavras que possuem o mesmo sinônimo ou variantes dessa palavra.

Cada casamento é dado uma pontuação de acordo a precisão e recall de unigramas e o quanto as palavras estão fora de ordem na tradução da máquina. Foi verificado, também, que o recall é mais importante do que a precisão na validação das traduções.

A métrica METEOR

O principal problema na métrica BLEU é a medição de unigramas que se sobrepõem (palavras únicas) e palavras com n-grams de alta ordem. O principal componente do BLEU é a precisão, calculada a partir dos n-grams casados com os n-grams da tradução, e não leva o recall em conta diretamente, e compensa isso utilizando uma penalidade no tamanho da tradução (Brevity Penalty).

Os problemas citados pelo artigo, no geral, são:

- Falta do recall.
- Uso de uma ordem alta de N-grams.
- Falta de casamento de palavras explícito entre a tradução e a referência.

- Uso de média geometrica dos N-grams.

Assim, METEOR foi desenvolvido para tratar desses problemas. Ele avalia uma tradução calculando uma pontuação baseada em correspondências explícitas palavra a palavra entre a tradução e uma tradução de referência.

METEOR cria um alinhamento entre duas strings. Um alinhamento é como um mapeamento entre unigramas, de modo que cada unigrama de cada string é mapeado para zero ou um unigrama na outra string. Esse alinhamento é produzido incrementalmente através de uma série de etapas.

Avaliação da métrica METEOR

A comparação em nível de sistema é a seguinte:

System ID	Correlation
BLEU	0.817
NIST	0.892
Precision	0.752
Recall	0.941
F1	0.948
Fmean	0.952
METEOR	0.964

É observado que o recall, por si só, se correlaciona com avaliação humana muito melhor do que precisão, e que a combinação dos dois usando a fórmula Fmean resulta em melhorias adicionais.