

AUTOMAÇÃO DA REDUÇÃO DE SENTENÇAS DE FONTES NÃO ESTRUTURADAS CANDIDATAS À CONSTRUÇÃO DE ONTOLOGIAS

Willian Santos Moura (willian_98@outlook.com)
Joinvile Batista Junior (joinvile@ufgd.edu.br)

Foi realizado o estudo e análise de aplicação de deep learning para sumarização extrativa com objetivo de selecionar as sentenças mais relevantes de um texto para extração de relações. Dos cinco artigos estudados, o quinto é o mais recente e o que reportou melhores resultados, sendo escolhido para estudo detalhado, treinamento, teste e avaliação dos resultados. Arquitetura da rede neural utilizada é composta de três niveis de processamento: as sentenças dos documentos de entrada são processadas em nivel de palavras, que no próximo nível são processadas em nível de sentenças que finalmente são priorizadas para escolha das sentenças que irão compor o resumo. Os dois primeiros níveis são baseados em Redes Neurais Recorrentes que tem a habilidade de levar em conta iterações de tempo em seu processamento, de tal forma que a saída de uma iteração realimenta o processamento da próxima iteração. A rede neural utilizada é submetida inicialmente a uma fase de treinamento, na qual os pesos da rede são refinados para aprender a classificar a relevância das sentenças para compor o sumário de cada documento de entrada. Após treinada, a rede é utilizada com outros documentos na fase de teste, para extrair as sentenças que irão compor os sumários. As sentenças selecionadas para cada documento da base de teste, são comparadas com os sumários de referência da base de dados, sendo aplicadas as métricas escolhidas de sobreposição do ROUGE. Neste trabalho, são utilizadas as seguintes métricas: ROUGE 1 (sobreposição de uma única palavra ou unigram), ROUGE 2 (sobreposição de duas palavras consecutivas ou bigram), e ROUGE L (sobreposição de uma sequência de palavras não necessariamente consecutivas). Foram disponibilizados uma implementação da proposta e os pesos da rede neural pré-treinada. Os resultados obtidos com a configuração completa (treinamento e teste) diferem dos resultados obtidos com a configuração de teste a partir da base pré-treinada somente na segunda casa após a vírgula. Os resultados reportados pelos autores do artigo 5, e pelo autor da implementação utilizada são próximos. No entanto, os resultados deste experimento são inferiores às duas sequências (três métricas do ROUGE) geradas, embora sejam ligeiramente superiores em relação ao artigo 4 em relação às métricas ROUGE 1 e 2, e ligeiramente inferior em relação à métrica ROUGE L. Foram escolhidos três documentos para análise comparativa. A sumarização extrativa tem as seguintes desvantagens em relação às sentenças do sumário de referência: (a) não suporta o agrupamento de informações de mais de uma sentença, para enriquecer uma sentença original; e (b) não suporta a supressão de trechos pouco relevantes de uma sentença original. Mesmo com essas desvantagens os resultados gerados pela implementação testada da sumarização extrativa se aproximam da caracterização das informações essenciais, representadas pelas sentenças do resumo de referência.