ENEPEX
ENCONTRO DE ENSINO, PESQUISA E EXTENSÃO
13º ENEPE UFGD - 10º EPEX UEMS
PESQUISA E TECNOLOGIA: AÇÕES PARA UM FUTURO SUSTENTÁVEL

ESTUDO E PROTOTIPAGEM DE FUNCIONALIDADES UTILIZADAS PARA APLICAÇÃO DE *DEEP LEARNING* NA EXTRAÇÃO DE RELAÇÕES DE TEXTOS NÃO ESTRUTURADOS

SANTOS, Thiago Lourenço dos[1] (thiagosup1@gmail.com); **BATISTA JR, Joinvile**[2] (joinvile@ufgd.edu.br)

[1]Discente do curso de Engenharia de Computação da UFGD;
[2]Docente dos cursos de Engenharia de Computação e de Sistemas de Informação da UFGD.

Recentemente tem sido reportada na literatura, a utilização bem sucedida de *deep learning*, para extrações de relações a partir de fontes não estruturadas, evitando a necessidade de utilização de ferramentas de processamento de linguagem natural, que introduzem erros no processo de melhoria da extração de relações. O objetivo central deste trabalho é o estudo de propostas de utilização de *deep learning*, bem como, a prototipagem e a avaliação de uma das propostas estudadas. Os três artigos selecionados para estudo detalhado atendem os critérios estabelecidos previamente na metologia proposta: (a) utilizam representações de palavras de um extenso vocabulário em vetores, de forma a servir como vocabulário de entrada (geradas pela ferramenta Word2Vector); e (b) se caracterizam por um alto reuso entre si. A proposta selecionada para prototipagem disponibiliza uma implementação que foi adaptada para a linguagem Java, para se integrar com ferramentas geradas em trabalhos anteriores. A primeira proposta utiliza a representação de um extenso vocabulário em vetores e uma rede convolucional para classificar relações. A entrada consiste de sentenças associadas a um par de entidades e a relação esperada, utilizada na fase de treinamento e para avaliação na fase de teste. É selecionada a sentença com maior probabilidade de ocorrência para cada par de entidade, desprezando a informação contida nas demais sentenças associadas ao mesmo par de entidades e, portanto, resultando na rotulação incorreta das demais relações associadas ao par de entidades. Na segunda proposta é aplicado, adicionalmente, o processo de atenção seletiva a cada sentença de entrada, considerando a influência de todas as sentenças associadas ao mesmo par de entidades. Na terceira proposta, é utilizado um vetor de representação para todas as sentenças associadas a um dado par de entidades, gerado a partir da escolha do maior valor para cada índice dos vetores das sentenças consideradas. A segunda proposta foi prototipada. Sua base de instâncias de treinamento é composta por 522.611 sentenças e 53 relações. No teste realizado, com 15 execuções da base de teste (*epochs*) e 12 *threads*, as 172.448 sentenças da base de teste se distribuiram nos seguintes grupos: (a) 172066 sentenças com pelo menos uma predição correta; e 382 sentenças sem nenhuma predição correta. Na base de teste constavam sentenças cobrindo 32 das 53 relações existentes na base de treinamento. Das 32 relações cobertas, 12 relações foram instanciadas com sentenças com pelo menos uma predição correta. Foram encontradas 165209 sentenças com predições corretas, sendo que a razão do total de predições corretas, em relação ao total de instâncias de teste, é de 0,9580. Este valor caracteriza um alto percentual de classificações corretas de relações. A proposta prototipada, conforme informado pelos autores, superou o estado da arte na extração de relações a partir de fontes não estruturadas.

Palavras-chave: *deep learning*, redes neurais convolucionais, extração de relações.

Agradecimentos: Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão de bolsa de iniciação científica ao primeiro autor