XXXIII SIC SALÃO INICIAÇÃO CIENTÍFICA

Evento	Salão UFRGS 2021: SIC - XXXIII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2021
Local	Virtual
Título	Identificação de relações semânticas em Word Embeddings
Autor	GABRIEL COUTO DOMINGUES
Orientador	JOEL LUIS CARBONERA

Identificação de relações semânticas em Word Embeddings

Aluno: Gabriel Couto Domingues
Orientador: Prof. Dr. Joel Luis Carbonera

Este projeto tem como objetivo desenvolver abordagens automáticas para identificar relações semânticas entre palavras representadas por word embeddings [1] pré-treinados.

Criamos datasets utilizando a base de dados léxicos WordNet [2], que representa palavras e relações entre elas de forma curada por especialistas, e dois modelos de word embeddings, que são vetores numéricos que representam palavras criados usando algoritmos de aprendizado de máquina.

Consideramos um conjunto de 8 relações disponíveis na WordNet para a classificação de pares de palavras: Hypernyms, Instance Hypernyms, Part Holonyms, Member Holonyms, Substance Holonyms, Similar, Antonyms e Synonyms. Para essas relações, buscamos todos pares de palavras relacionadas na WordNet, e substituímos as palavras pelos vetores correspondentes, em um conjunto de word embeddings pré-treinados. Em alguns datasets realizamos um pré-processamento para retirar pares de palavras que se relacionavam por mais de uma relação. Foram criados 8 datasets no total, sendo quatro de classificação binária, cujo foco era identificar relações taxonômicas (de hiperonímia); e quatro de classificação multiclasse, considerando as 8 relações disponíveis.

A partir destes datasets foram treinados diferentes modelos de redes neurais. O primeiro modelo foi baseado no modelo proposto em [3]. Os demais modelos testados foram variações do modelo original.

Nos experimentos realizados utilizando o modelo proposto em [3] com os datasets criados utilizando o mesmo conjunto[1] de word embeddings utilizado pelos autores, constatou-se que com pré-processamento foi possível atingir uma acurácia de aproximadamente 91% para classificação binária e de aproximadamente 87% para classificação multiclasse. Esses resultados apresentam uma melhora de aproximadamente 2% em relação à performance alcançada pelo modelo nos datasets sem pré-processamento. Esses resultados indicam que as redes neurais possuem uma performance aceitável para classificar relações usando word embeddings e que o pré-processamento proposto tem um impacto positivo sensível na performance.

Referências

1. Wang, S., Zhou, W., & Jiang, C. (2020). A survey of word embeddings based on deep learning. *Computing*, *102*(3), 717-740.
2. Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.
3. Khadir, A. C., Guessoum, A., & Aliane, H. (2020, October). Ontological Relation Classification Using WordNet, Word Embeddings and Deep Neural Networks. In *International Symposium on Modelling and Implementation of Complex Systems* (pp. 136-148). Springer, Cham.

[1] https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.en.vec