USO DE EMBEDDINGS PARA A CLASSIFICAÇÃO AUTOMÁTICA DE ESTRUTURA RETÓRICA DE REDAÇÕES

Vinícius da Costa Regatieri (PIC/UEM), Valéria Delisandra Feltrim (Orientadora), e-mail: ra104016@uem.br.

Universidade Estadual de Maringá / Centro de Tecnologia / Departamento de Informática. Maringá, PR.

Área: Ciências Exatas e da Terra.
Subárea: Ciência da Computação

Palavras-chave: resposta argumentativa, *word embedding*, aprendizado de máquina, processamento de linguagem natural.

Resumo:

Este trabalho teve como objetivo explorar o uso de modelos distribuídos de língua, chamados de *embeddings*, no desempenho de um sistema de classificação automática utilizando uma base de quatrocentas e cinquenta e cinco redações do gênero resposta argumentativa dos vestibulares da Universidade Estadual de Maringá. As redações foram digitadas e anotadas manualmente por três anotadores de acordo com um modelo de estrutura retórica determinado para o gênero. Com as anotações, foi possível treinar classificadores usando três técnicas de aprendizado de máquina: *Support Vector Machine* (SVM), Regressão Logística e *Conditional Random Fields* (CRF) em trinta e seis espaços vetoriais diferentes. Com a análise dos resultados foi possível concluir que, para a base de redações deste trabalho, o desempenho do algoritmo CRF foi superior. Os modelos de *embeddings* que mais se mostraram eficientes no aprendizado foram os modelos com bases de textos maiores e com diversos gêneros, com vetores de *embeddings* medianos.

Introdução

A produção textual é um dos principais assuntos nas salas de aula no ensino fundamental e médio. Isso ocorre principalmente pelo fato de diversas universidades públicas e privadas terem a escrita de redações dos mais diversos gêneros em seus processos seletivos. No caso da Universidade Estadual de Maringá (UEM), um dos possíveis gêneros requisitados na prova de redação é a resposta argumentativa. Alguns estudos buscam caracterizar gêneros textuais, criando o que foi chamado de estrutura retórica (Swales, 1990; Teufel e Moens, 2002; Burstein *et al*., 2003; Andreani e Feltrim, 2015), para então ser possível automatizar o processo de identificação de tais estruturas presentes em um texto, auxiliando o processo

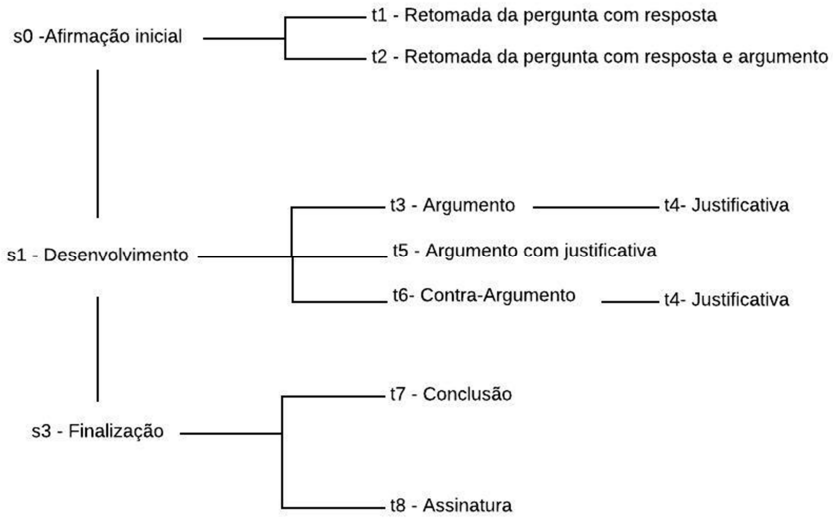
de correção e avaliação dos processos seletivos em universidades. Nesse cenário de demanda por ferramentas computacionais que contribuam na avaliação dos textos produzidos em vestibulares, este projeto propôs a construção de classificadores automáticos de estrutura retórica para redações do gênero resposta argumentativa, produzidas em provas de vestibulares da UEM, objetivando a exploração do uso de modelos distribuídos de língua, chamados de *embeddings*, no treinamento e teste dos classificadores.

Materiais e métodos

Corpus

Um total de quatrocentas e cinquenta e cinco redações do gênero resposta argumentativa, produzidas nos Vestibulares e Processos Seletivos de Avaliação Seriada da UEM, foram utilizadas para compor o *corpus* deste trabalho. Em uma primeira etapa, todas as redações foram digitadas manualmente e salvas como documento de texto. Após isso, uma estrutura retórica que serviu de base para a anotação de uma parcela do *corpus* foi desenvolvida e três anotadores anotaram as redações sem nenhum tipo de discussão. Entendeu-se que uma adequação à estrutura era necessária para a anotação do restante do *corpus* e, por isso, definiu-se uma nova estrutura retórica, ilustrada na Figura 1.

Figura 1 – Estrutura retórica final desenvolvida para anotação das redações.



Fonte: Autores (2020).

Com a nova estrutura retórica definida, três rodadas de anotações foram realizadas com cem, cinquenta e cinquenta redações, respectivamente, sendo que, em cada rodada a concordância entre os anotadores foi medida

pela estatística Kappa, com resultados 0,534, 0,652 e 0,784 para as três rodadas, respectivamente, com o último resultado atestando um alinhamento aceitável entre os anotadores, que prosseguiram com a anotação do restante do *corpus*.

Aprendizado

Primeiramente, para o aprendizado é necessário que sejam extraídos os atributos que representam as sentenças a serem classificadas. Para este estudo, as sentenças foram representadas por meio de vetores, chamados de *embeddings*, extraídos a partir de modelos de espaços vetoriais densos, sendo um deles o modelo Doc2Vec. Além da geração de espaços vetoriais a partir das redações, também foram utilizados modelos pré-treinados disponibilizados pelo Núcleo Internacional de Linguística Computacional (NILC), que utilizam o formato de *word embeddings* e necessitaram de estratégias de transformação para o formato utilizado neste estudo, sendo elas soma e média. Após a extração de atributos, o treinamento e testes dos classificadores se deu por validação cruzada de dez partições.

Resultados e Discussão

Devido ao grande número de espaços vetoriais utilizados na fase experimental do projeto, os valores exibidos na Tabela 1 correspondem às médias ponderadas obtidas para cada métrica de avaliação para os melhores modelos. Como pode ser observado, o melhor resultado foi obtido com o classificador CRF utilizando um dos modelos de *word embeddings* pré-treinado do NILC, o Wang2Vec-CBOW com vetores de 100 dimensões e o método conversor de soma.

Tabela 1 – Melhores resultados para os algoritmos SVM, RL e CRF

Algoritmo	*Precision*	*Recall*	*F1-Score*
SVM	0,51	0,52	0,51
RL	0,53	0,53	0,53
CRF	0,66	0,66	0,66

Fonte: Autores (2020).

Os resultados experimentais também mostraram que os espaços vetoriais treinados com mais textos e com gêneros diversos, como os disponibilizados pelo NILC, foram melhores para a tarefa de classificação deste trabalho, possivelmente devido ao tamanho limitado do *corpus* de redações para o treinamento de um modelo Doc2Vec específico.

Conclusões

Este projeto teve como objetivo explorar o uso de modelos de *embeddings* para a geração de atributos empregados na indução de classificadores de estrutura retórica para redações do gênero resposta argumentativa. Por meio de experimentos de validação cruzada, constatou-se que o classificador CRF obteve melhor resultado médio quando comparado aos outros dois algoritmos utilizados no trabalho.
Com relação aos algoritmos classificadores, os resultados mostraram que um classificador de sequências, como o CRF, é mais adequado à tarefa de classificação retórica do que classificadores como o SVM e o RL.
Com relação aos modelos de espaço vetoriais, foi evidente a superioridade dos espaços pré-treinados com mais textos de diversos gêneros, representados por modelos do NILC, quando comparados aos modelos menores criados para este projeto. Também foi possível concluir que os *embeddings* com dimensões maiores não melhoraram os resultados, sendo que, o tamanho dos *embeddings* que mais se mostrou proveitoso neste estudo foi 100.

Agradecimentos

Agradeço à minha orientadora pelo tempo dedicado ao trabalho e aos amigos e familiares que apoiaram o desenvolvimento do projeto.

Referências

ANDREANI, A. C.; FELTRIM, V. D. Campos Aleatórios Condicionais Aplicados à Detecção de Estrutura Retórica em Resumos de Textos Acadêmicos em Português. In: X SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA. **Anais...** SBC, 2015. p. 111-120.

SWALES, J. **Genre analysis: English in academic and research settings**. Cambridge University Press, 1990.

TEUFEL, S.; MOENS, M. Summarizing scientific articles: experiments with relevance and rhetorical status. **Computational linguistics**, v. 28, n. 4, p. 409-445, 2002.

BURSTEIN, J.; MARCU, D.; KNIGHT, K. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. **IEEE Intelligent Systems**, v. 18, n. 1, p. 32-39, 2003.