USO DE WORD EMBEDDINGS PARA A CLASSIFICAÇÃO DE ESTRUTURA RETÓRICA EM REDAÇÕES DO GÊNERO ARTIGO DE OPINIÃO

Bruna Stefany Batista Marques (PIC/CNPq/Uem), Valéria Delisandra Feltrim (Orientadora), e-mail: ra103404@uem.br.

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

Ciência da Computação/Metodologia e Técnicas da Computação

Palavras-chave: artigo de opinião, *word embedding*, processamento de linguagem natural.

Resumo:

Este trabalho teve como objetivo avaliar o desempenho de um classificador automático de estrutura retórica para o gênero textual artigo de opinião baseado em *word embeddings*. Para o treinamento e avaliação do classificador foi utilizada uma base de dados de duzentas e setenta e uma redações de vestibulares da Universidade Estadual de Maringá fornecidas pelo Comissão Central do Vestibular Unificado (CVU). Todas redações foram digitadas e anotadas manualmente por três anotadores de acordo com um modelo de estrutura retórica proposto para o gênero abordado. Para a extração dos vetores de atributos foram utilizados modelos de *word embeddings* pré-treinados para a língua portuguesa. Esses vetores foram então utilizados para treinar um classificador *Support Vector Machines* (SVM) e os resultados obtidos foram comparados aos de um classificador treinado com atributos TF-IDF. A partir da análise dos resultados foi possível constatar que, para esse conjunto de redações, o desempenho do SVM com atributos TF-IDF foi melhor.

Introdução

A produção textual é de extrema importância no aprendizado de línguas tanto no âmbito escolar quanto na sociedade em geral. No ensino médio o enfoque das produções textuais, geralmente, é nos gêneros exigidos pelos vestibulares das universidades. A Universidade Estadual de Maringá (UEM) é uma das universidades que tem exigido na prova de redação os seus concursos vestibulares que o candidato produza textos em determinados gêneros, de acordo com uma lista que é divulgada antecipadamente e periodicamente atualizada pela Comissão Central do Vestibular Unificado (CVU-UEM).

O conceito de estrutura esquemática ou retórica (Swales 1990) tem sido empregado como base teórica de estudos que buscam caracterizar gêneros textuais específicos em termos de padrões esquemáticos que podem ser aprendidos e reproduzidos. Essa padronização, que se dá em maior ou menor grau dependendo do gênero textual abordado, abre caminho à construção de sistemas automáticos que possam identificar as categorias retóricas (ou esquemáticas) presentes em textos de um

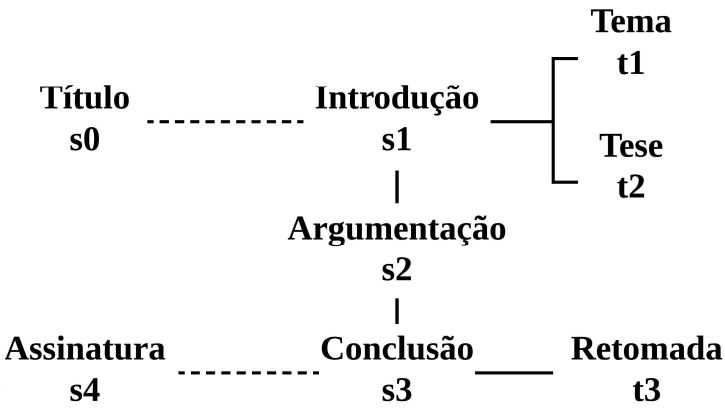
determinado gênero. Por sua vez, a informação extraída por tais sistemas pode ser empregada tanto no auxílio ao ensino e aprendizagem dos gêneros quanto na avaliação de proficiência.

Nesse contexto, este trabalho como objetivo a implementação e avaliação de um classificador automático de estrutura retórica para o gênero textual artigo de opinião empregando atributos extraídos por meio de diferentes modelos de *word embeddings*. Especificamente, procurou-se avaliar se o uso desse tipo de atributo contribuiria para a melhoria de desempenho em comparação a um classificador baseado em atributos TF-IDF.

Materiais e métodos

Foi utilizado um córpus composto por 271 redações do gênero artigo de opinião produzidas nos vestibulares da UEM de 2014 e 2016 e que receberam boas notas. Em um primeiro momento, as redações foram digitadas exatamente da forma como foram escritas (incluindo possíveis erros) e, após isso, salvas como um documento de texto. Após a digitalização das redações, elas foram manualmente anotadas por três anotadoras segundo o modelo de estrutura retórica de Soares et al. (2018) mostrado na Figura 1.

Figura 1 – Estrutura retórica utilizada na anotação das redações



A concordância entre as anotadoras foi medida por meio da estatística *Kappa*. A partir dos valores *Kappa* obtidos por categoria foi possível observar que os resultados, no geral, foram bons, visto que para a maior parte das categorias o valor *Kappa* foi acima de 0,70. Esses resultados permitiram assumir que o modelo retórico proposto é reprodutível e, assim, a anotação resultante poderia ser usada para treinar o classificador. A partir das três anotações, gerou-se um córpus único resultante da concordância da maioria.

Para a realização dos experimentos com algoritmos de aprendizado de máquina foi necessário, primeiramente, realizar a extração de atributos. Em particular, neste projeto, utilizou-se para esse fim diferentes modelos de *word embeddings*, que são representações distribuídas de palavras, ou espaços semânticos, no qual palavras com significados próximos são representadas de forma similar. Cada uma dessas

palavras é armazenada em um vetor de valores reais com uma determinada dimensão, sendo a dimensão um parâmetro do modelo.
O córpus anotado de redações, composto de 2.562 sentenças, foi armazenado em um arquivo em formato JSON. A representação vetorial das sentenças se deu pela combinação dos *embeddings* recuperados para cada palavra de uma sentença. A combinação foi feita utilizando-se três funções: soma, multiplicação e média dos vetores referentes aos *embeddings*. Os vetores resultantes de cada sentença foram utilizados para o treinamento, por meio de validação cruzada de 10 partições, de classificadores *Support Vector Machines* (SVM). Foram avaliados os seguintes modelos de *word embeddings*: FastText, Glove, Wang2Vec e Word2Vec nas dimensões 50, 100, 300, 600 e 1000, todos eles previamente treinados e disponibilizados pelo NILC-USP.
Como base de comparação para o desempenho dos classificadores baseados em *word embeddings*, foi utilizado um classificador treinado a partir de vetores gerados por meio do cálculo de valores TF-IDF. Os vetores usados por esse classificador foram calculados com base nos 100 melhores unigramas do conjunto de teste selecionados por meio do teste Chi-quadrado. Assim como nos outros experimentos, esse classificador foi avaliado por meio de validação cruzada de 10 partições.
O desempenho dos classificadores foi medido por meio das métricas de precisão, revocação e medida-f. Devido a limitação de espaço, apenas a medida-f é reportada.

Resultados e Discussão

Com relação aos diferentes modelos de *word embeddings*, não foi possível estabelecer um melhor modelo, uma vez que os resultados obtidos foram similares. Também não houve diferença significativa de desempenho com relação às dimensões do modelo. Já com relação a forma como os *word embeddings* foram combinados para gerar os *embeddings* das sentenças, no geral, as funções multiplicação e média trouxeram resultados melhores em comparação com a função soma. Comparando-se apenas os valores médios de medida-f obtidos nos diferentes experimentos, o modelo Wang2Vec obteve os melhores resultados (0,65 de medida-f). Esse resultado foi obtido tanto pelo modelo Wang2Vec-CBOW de 50 dimensões usando a função soma, quanto pelo modelo Wang2Vec-SKIP-GRAM de 100 dimensões usando a função média. Embora os dois modelos mencionados tenham obtido o mesmo valor médio de medida-f, as métricas por categoria foram diferentes. Apesar disso, foi possível observar um padrão de comportamento no desempenho por categorias. Por exemplo, os dois modelos obtiveram resultados mais altos para as categorias s0, t1 e s2 e mais baixos para a categoria s3 e s4.
O classificador baseado em TF-IDF foi melhor do que os classificadores baseados em *word embeddings*, obtendo 0,71 de medida-f média. De forma especial, foi possível observar uma grande diferença de desempenho para a categoria s4. Acredita-se que isso deva ao fato dos atributos TF-IDF serem calculados diretamente a partir do córpus de redações, sendo assim capazes de capturar particularidades desse córpus. Os modelos de *embeddings*, por outro lado, são gerais, pois foram treinados a partir de um córpus muito maior e diverso. Um resumo dos resultados obtidos, bem como a distribuição de categorias no córpus, é apresentado na Tabela 1.

Tabela 1 – Resultados obtidos utilizando os modelos de *word embeddings* e TF-IDF

Categoria	Medida-f		# Sentenças no córpus
	Embeddings	TF-IDF	
s0	0,86	0,93	213
t1	0,76	0,60	958
t2	0,57	0,52	218
s2	0,91	0,78	255
t3	0,52	0,62	544
s3	0,30	0,24	259
s4	0,31	0,99	115
Média/Total	0,65	0,71	2.562

Conclusões

Neste trabalho foi apresentado um classificador automático de estrutura retórica em redações do gênero artigo de opinião baseado em *word embeddings*. Foram avaliados 35 modelos diferentes de w*ord embeddings* em combinação com três funções de combinação de *word embeddings*: soma, multiplicação e média. Os resultados obtidos por esses modelos foram comparados com os de um classificador baseado em atributos TF-IDF. Foi possível constatar que, para o conjunto de redações utilizado, o classificador baseado em TF-IDF obteve melhores resultados. Isso se deu devido ao fato desses vetores terem capturado padrões específicos do córpus de redações, uma vez que, diferentemente dos modelos de *word embeddings* utilizados, foram calculados diretamente a partir do córpus de treinamento. Dessa forma, um possível trabalho futuro é o treinamento de modelos de *embeddings* utilizando um *córpus* de redações.

Agradecimentos

A Deus por ter me dado saúde e força para finalizar este trabalho. À minha orientadora, pelo empenho dedicado à elaboração deste trabalho. À família e amigos pelo apoio e incentivo.

Referências

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

SANTOS, K. S.; SODER, M.; MARQUES, B. S. B.; FELTRIM, V. D. Analyzing the rhetorical structure of opinion articles in the context of a Brazilian college entrance examination. In: INTERNATIONAL CONFERENCE ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE, 2018, Canela, RS, Brazil. **Proceedings...** Springer International Publishing, 2018. p 3-12.