ANÁLISE AUTOMÁTICA DA ESTRUTURA RETÓRICA DO GÊNERO ARTIGO DE OPINIÃO

Karina Soares dos Santos (PIBIC/FA/UEM), Valéria Delisandra Feltrim (Orientadora), e-mail: karinasoares.uem@gmail.com

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

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: estrutura retórica, artigo de opinião, aprendizado de máquina.

Resumo:

Em vista da necessidade de abordagens automáticas voltadas tanto ao ensino-aprendizagem, quanto à avaliação de textos produzidos em contexto de vestibular, este projeto teve por objetivo a elaboração de um analisador estrutural automático para textos do gênero artigo de opinião que foram produzidos em vestibulares da Universidade Estadual de Maringá. A partir da anotação manual e da análise de um *corpus* de redações, foram aplicadas técnicas de processamento de linguagem natural e aprendizagem de máquina para a construção de um classificador de estrutura retórica. Os resultados obtidos com o classificador foram encorajadores, mostrando a viabilidade da abordagem proposta.

Introdução

O artigo de opinião é um gênero textual que tem se popularizado em diferentes cenários, como em vestibulares. Essa mudança tem estimulado diversos estudos, em especial linguísticos e computacionais, visando à criação de subsídios e ferramentas que auxiliem tanto na produção quanto na avaliação desses textos com um menor custo e agilidade.

Visando a criação de abordagens automáticas de auxílio e avaliação da escrita, neste projeto foi elaborado um analisador estrutural de textos do gênero artigo de opinião redigidos como parte das provas de vestibular da Universidade Estadual de Maringá (UEM). Para isso, foi proposto um modelo de estrutura de retórica para o gênero abordado como também foi utilizado para a anotação manual de um *corpus* de redações. A partir do *corpus*, foram aplicadas técnicas de processamento de linguagem natural para a extração de características textuais, que, posteriormente, foram usadas por um algoritmo de aprendizagem de máquina para a indução de um classificador de estrutura retórica específico para o gênero. As atividades realizadas e os principais resultados obtidos no projeto estão descritos a seguir.

Materiais e métodos

As atividades realizadas neste projeto seguiram a seguinte esquemática:

Revisão da Literatura

A busca na literatura por referenciais teóricos corroborou para a elaboração de um modelo de estrutura retórica do artigo de opinião, portanto, envolveu estudiosos da Linguística Textual e os estudos da Análise Sistemática dos Discursos.

Formatação do corpus

O *corpus* utilizado nesta pesquisa foi provido pela UEM em formato de imagem. Dessa forma, todos os textos foram digitados manualmente em formato de arquivo de texto. O *corpus* resultante é composto por 271 textos, sendo 200 referentes ao vestibular de Inverno de 2016 e 71 referentes ao processo seletivo do inverno de 2014.

Análise preliminar das redações

Após a formatação do corpus, foi realizada uma análise preliminar das redações para subsidiar a proposta de um modelo retórico para o gênero artigo de opinião em contexto de vestibular. Nesse processo foram investigados os recursos discursivos usados nas sentenças do *corpus*, bem como as relações discursivas em nível de investigação linguística dadas as estruturas locais e globais do discurso.

Proposta do modelo de estrutura retórica

Zanini (2017) contribuiu para os estudos sobre o gênero e a teoria sociointerativa adotada nos manuais didáticos e documentos oficiais relacionados ao ensino-aprendizado de produções textuais. No que concerne à noção de estrutura retórica voltada para aplicações computacionais, os estudos de van Dijk e Kintsch (1983) e van Dijk (1980) auxiliaram na compreensão das estruturas globais e locais das redações.
A partir desses estudos e da análise preliminar do *corpus*, foi possível formalizar um modelo de estrutura retórica representativo da estrutura do artigo de opinião em contexto de vestibular e com granularidade adequada para o treinamento de um classificador automático.

Anotação das redações

Com base no modelo de estrutura retórica proposto, a anotação manual do *corpus* foi realizada por três anotadoras instruídas a atribuir uma categoria para cada uma das sentenças das redações. Para medir a concordância

entre as anotadoras e, dessa forma, verificar a reprodutibilidade do modelo proposto, foi empregada a estatística *Kappa*.

Extração dos atributos e experimentação com os algoritmos de aprendizado

O pré-processamento, a extração de características e o treinamento e teste dos classificadores foram feitos usando a biblioteca scikit-learn.
O pré-processamento incluiu a segmentação de sentenças e a tokenização. Como características foram usados valores TF-IDF extraídos a partir do conjunto de treinamento. A distribuição χ^2 foi usada para selecionar as melhores características e reduzir a dimensão dos vetores.
Foram avaliadas diferentes configurações para os vetores de características resultantes da combinação de TF-IDF extraídos de diferentes formas (considerando unigramas, bigramas e trigramas) com diferentes valores de corte para o χ^2 (50, 100 e 1000). Também foi considerada como característica a posição absoluta da sentença na redação. Em todos os casos possíveis, o vetor para uma sentença s_i continha, além das suas próprias características, as características das sentenças s_{i-1} e s_{i+1}.
O algoritmo de aprendizado empregado foi o *support vector machines* (SVM) de *kernel* linear com valores padrão para os demais parâmetros. Em todos os experimentos os classificadores induzidos foram avaliados por meio de validação cruzada de 10 partições.

Resultados e Discussão

O modelo de estrutura retórica para o artigo de opinião em contexto de vestibular resultante deste projeto é mostrado na Figura 1.

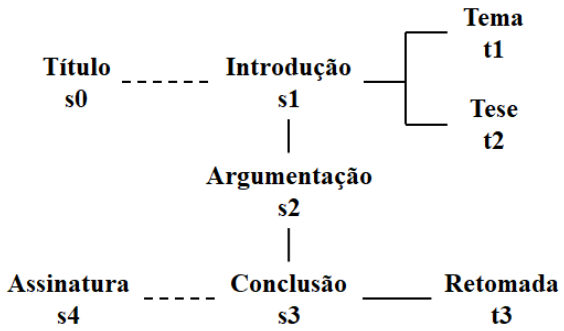

Figura 1: Modelo de estrutura retórica para o artigo de opinião

Com base nas categorias da Figura 1, foi feito o experimento de anotação e o valor *Kappa* obtido foi de 0,78, o que indica uma concordância substancial entre as anotadoras. O corpus resultante desse experimento foi então usado para os experimentos de aprendizado.
Os melhores resultados para a classificação automática foram obtidos com os vetores de características contendo: os 100 valores TF-IDF mais discriminantes (selecionados com χ^2) gerados a partir de unigramas, a

posição absoluta da sentença e as características das sentenças anterior e posterior. Os resultados por categoria em termos de precisão, revocação, medida-f, bem como a distribuição de categorias no *corpus* (suporte), são apresentados na Tabela 1.

Tabela 1. Resultados do classificador SVM

Categoria	Precisão	Revocação	Medida-f	Suporte
Título (s0)	0,90	0,97	0,93	213
Tema (t1)	0,67	0,54	0,60	544
Tese (t2)	0,63	0,44	0,52	259
Argumentação (s2)	0,69	0,90	0,78	958
Conclusão (s3)	0,66	0,58	0,62	218
Retomada (t3)	0,63	0,15	0,24	115
Assinatura (s4)	1,00	0,98	0,99	255
Média/total	0,72	0,73	0,71	2562

Conclusões

Neste projeto foi conduzido um estudo a respeito da estrutura retórica de artigos de opinião produzidos no contexto de vestibular. A partir da revisão bibliográfica e da análise manual de um *corpus*, foi proposto um modelo de estrutura retórica que visa representar a estrutura dos artigos de opinião produzidos nesse contexto específico.
Um *corpus* de redações foi manualmente anotado e os resultados experimentais mostraram que o modelo é reprodutível. Esse *corpus* também foi usado para construir um classificador que atribui uma categoria retórica a cada sentença de um artigo de opinião. O melhor classificador obteve medida-f média de 0,71, o que é um resultado promissor considerando-se que foram empregadas características superficiais apenas e que o *corpus* possui um tamanho reduzido.

Agradecimentos

À Fundação Araucária pelo apoio financeiro.

Referências

VAN DIJK, T.A. **Macrostructures:** An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. New Jersey: Lawrence Erlbaum Associates, 1980.

VAN DIJK, T. A., KINTSCH, W. **Strategies of discourse comprehension**. New York: Academic Press, 1983.

ZANINI, M. Artigo de opinião: do ponto de vista à argumentação. In: ANTONIO, J.D, NAVARRO, P. (Org.). **Gêneros textuais em contexto de vestibular**. Maringá: Eduem, 2017. p. 43–58.