CLASSIFICAÇÃO AUTOMÁTICA DE ESTRUTURA RETÓRICA EM REDAÇÕES DO GÊNERO ARTIGO DE OPINIÃO

Mariana Soder (PIC/UEM), Valéria Delisandra Feltrim (Orientadora),
e-mail: ra95381@uem.br

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

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: estrutura retórica, artigo de opinião, processamento de linguagem natural

Resumo:

Este trabalho teve como objetivo investigar o desempenho de classificadores automáticos de estrutura retórica para redações do gênero artigo de opinião que foram produzidas como parte do vestibular da Universidade Estadual de Maringá. O treinamento e o teste dos classificadores foram feitos usando um *corpus* de redações manualmente anotado de acordo com um modelo de estrutura retórica de sete categorias. Foram usadas características superficiais, como localização e tamanho das sentenças, e características que codificam informações morfossintáticas. Os classificadores foram induzidos por meio de dois algoritmos de aprendizagem de máquina supervisionada, o *Support Vector Machine* (SVM) e o *Conditional Random Fields* (CRF). Os resultados experimentais mostraram que o classificador CRF com todas as características obteve o melhor desempenho, atingindo a média de 88% de medida-F, o que mostra a relevância das características investigadas neste projeto para a classificação de estruturas retóricas.

Introdução

No contexto do ensino de língua portuguesa, a produção textual é uma das prioridades em todos os níveis de ensino. Isso é constatado também nas universidades, uma vez que a proficiência em diferentes gêneros textuais tem sido exigida nos processos seletivos para ingresso nas mesmas. No caso da Universidade Estadual de Maringá (UEM), a prova de redação exige que o candidato produza textos em gêneros específicos que são pré-definidos em uma lista divulgada com antecedência e periodicamente atualizada, sendo que um desses gêneros é o artigo de opinião.
Considerando a demanda por ferramentas computacionais que auxiliem tanto no processo de produção textual quanto na avaliação do grande volume de textos produzidos em processos seletivos como os vestibulares, este projeto teve como objetivo a construção e a avaliação de classificadores automáticos de estrutura retórica para redações do gênero artigo de opinião

baseados em características que codificam informações extraídas a partir da superfície das sentenças, como localização e tamanho, e informações morfossintáticas, como tempo e voz do verbo principal.

Materiais e métodos

Corpus

O *corpus* utilizado neste estudo é composto de redações do gênero artigo de opinião produzidas por candidatos do vestibular da UEM dos anos de 2014 e 2016. Os textos fornecidos pela instituição foram digitalizados, segmentados em sentenças e armazenados em formato XML. Ao todo são 271 redações que totalizam 2.562 sentenças.
O *corpus* foi manualmente anotado por três anotadoras com base em um modelo de estrutura retórica proposto especificamente para o gênero artigo de opinião em contexto de vestibular. O modelo composto de sete categorias retóricas é mostrado na Figura 1. Após um período de treinamento, as anotadoras foram instruídas a atribuir uma das sete categorias do modelo para cada sentença do *corpus*.

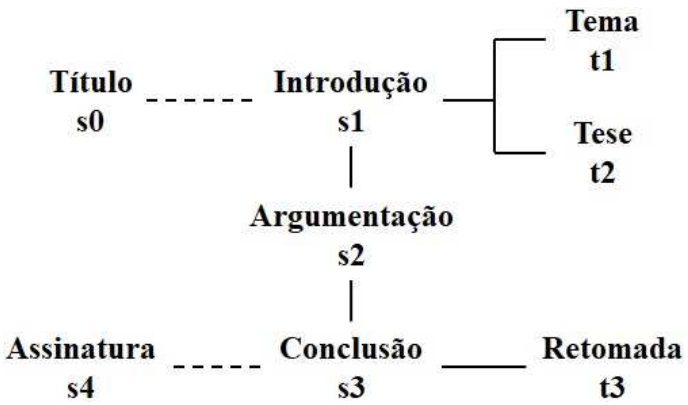

Figura 1 – Modelo de estrutura retórica para o gênero artigo de opinião em contexto de vestibular.

A concordância observada entre as anotadoras foi substancial, com estatística *Kappa* de 0,78 (N = 2.532, k = 3, n = 7). Desse modo, concluiu-se que a anotação é reprodutível o suficiente para ser usada no treinamento de um classificador automático. A partir dessas anotações, foi criada a versão anotada final do *corpus*, na qual a categoria de cada sentença foi dada pela votação da maioria.

Extração de Características

As características usadas neste projeto foram adaptadas dos trabalhos de Feltrim et al., (2006) e de Andreani e Feltrim (2015). Para cada sentença do *corpus* foram extraídas sete características, a saber: (1) posição relativa

(início, meio ou fim); (2) posição absoluta; (3) tamanho da sentença em palavras (pequeno, médio e grande); (4) voz do verbo principal; (5) tempo do verbo principal; (6) presença de auxiliar modal; e (7) posição relativa ao segmento retórico. A característica (7) refere-se à posição que a sentença ocupa dentro de um mesmo segmento retórico, i.e., uma sequência de sentenças de mesma categoria retórica. As características (1), (3), (4), (5) e (6) foram extraídas por meio do classificador AZPort (FELTRIM et al., 2006). A extração das características (2) e (7) foi implementada usando a linguagem Python.

Aprendizagem

Os classificadores foram induzidos usando dois algoritmos de aprendizagem, o *Support Vector Machine* (SVM) e o *Conditional Random Fields* (CRF). Nos dois casos, foram usadas as implementações fornecidas pela biblioteca Python Scikit-learn.

Experimentos

A etapa de experimentação consistiu na avaliação da combinação de diferentes subconjuntos de características com os diferentes algoritmos de aprendizagem. Os subconjuntos de características foram gerados usando-se uma abordagem “todas menos uma” e, para alguns casos, “todas menos duas”. Além disso, nos experimentos com o CRF, foi adicionado ao vetor da sentença s_i as características das sentenças s_{i-1} e s_{i+1}, sempre que possível. Todos os resultados de avaliação foram coletados por meio de validação cruzada com dez partições sobre o *corpus* de redações. No caso do SVM, as partições foram criadas a partir do total de sentenças. Para o classificador CRF, as partições foram criadas a partir do total de redações, uma vez que esse classificador prevê sequências de categorias.

Resultados e Discussão

Os melhores resultados obtidos em termos das médias das métricas de precisão, revocação e medida-F para os classificadores SVM e CRF são mostrados na Tabela 1. Em ambos os casos, os classificadores usaram todas as características disponíveis.

Tabela 1 – Melhores resultados para os classificadores SVM e CRF.

	Precisão	Revocação	Medida-F
CRF	0,88	0,88	0,88
SVM	0,73	0,74	0,73

Nos experimentos com o classificador SVM, observou-se uma queda significativa de desempenho quando foram retiradas as características (2) posição absoluta e (7) posição relativa ao segmento retórico, indicando que

essas características tiveram maior relevância para esse classificador. Em relação ao classificador CRF, também foi observada uma queda no desempenho com a retirada da característica (7), enfatizando a sua contribuição para a classificação independentemente do algoritmo utilizado. Por outro lado, a retirada das características (5) tempo do verbo principal e (6) presença de auxiliar modal não influenciou o desempenho do CRF, indicando menor relevância dessas informações para esse classificador.
Por fim, os resultados do classificador CRF foram superiores aos do SVM, mostrando que o uso de algoritmos que fazem a predição de sequências é vantajoso para a classificação de estruturas retóricas.

Conclusões

Este trabalho investigou o desempenho de classificadores automáticos de estrutura retórica para redações do gênero artigo de opinião. Usando um *corpus* de redações manualmente anotado, foram construídos classificadores baseados em características superficiais e morfossintáticas.
Os resultados mostraram que as características investigadas são relevantes para a tarefa da classificação de estrutura retórica, com destaque para as características que codificam a posição ocupada pela sentença tanto no texto quanto em um segmento retórico. Mostraram também que o algoritmo CRF apresenta melhor desempenho do que o SVM.
A partir desses resultados, conclui-se que informações a respeito do contexto de ocorrência das sentenças contribuem para a obtenção de resultados melhores, sejam elas codificadas em características de posicionamento ou pelo próprio algoritmo de aprendizado.

Agradecimentos

À minha orientadora e ao programa PIC-UEM pela oportunidade.

Referências

FELTRIM, V. D.; TEUFEL, S.; NUNES, M.G.V.; ALUÍSIO, S.M. Argumentative zoning applied to critiquing novices' scientific abstracts. In: SHANAHAN, J.; QU, Y.; WIEBE, J., (eds.) **Computing Attitude and Affect in Text: Theory and Applications**, Dordrecht, Netherlands: Springer, 2006. p. 233-246.

ANDREANI, A. C., FELTRIM, V. D. Campos Aleatórios Condicionais Aplicados à Detecção de Estrutura Retórica em Resumos de Textos Acadêmicos em Português. In: X BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY. **Proceedings...** Natal, 2015. p.111-120.