Extração de Atributos para Classificação de Papéis Retóricos em Textos Científicos

Alison Rafael Polpeta Freitas[1], **Valéria Delisandra Feltrim**[1]

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
Av. Colombo, 5.790 – 87020-900 – Maringá – PR – Brazil

alisonrafael@ymail.com, valeria.feltrim@din.uem.br

Abstract. *This work presents the analysis and implementation of attributes for AZPort, an statistical classifier that automatically detects the rhetorical structure of academic abstracts written in Portuguese.*

Resumo. *Este trabalho apresenta a análise e implementação de atributos de classificação para o AZPort, um classificador estatístico que faz a detecção automática da estrutura retórica de resumos acadêmicos escritos em português.*

1. Introdução

O AZPort (*Argumentative Zoning for Portuguese*) [Feltrim et al. 2006] é um classificador estatístico que tem por objetivo detectar de forma automática a estrutura retórica de resumos acadêmicos escritos em português. Tal classificador é parte do SciPo (*Scientific Portuguese*) [Feltrim et al. 2006], um ambiente de auxílio à escrita voltado especialmente para escritores iniciantes e que abrange várias etapas do processo de escrita, entre elas, a estruturação de resumos e introduções.

Baseado no AZ (*Argumentative Zoning*) [Teufel and Moens 2002], o AZPort implementa um subconjunto de oito atributos dos dezesseis atributos utilizados pelo AZ original para classificar cada sentença de um texto em uma das possíveis categorias retóricas, que variam de acordo com o tipo do texto classificado (resumo ou introdução). Atualmente, o AZPort tem precisão de 72% e Kappa igual a 0,65, valores esses computados aplicando-se 13-*fold cross-validation* aos 52 resumos do CorpusDT [Feltrim et al. 2006]. Em avaliações realizadas com as introduções do CorpusDT, o classificador manteve o desempenho semelhante.

Embora o desempenho do AZPort esteja abaixo do desempenho humano para essa tarefa (para resumos, o Kappa entre três anotadores é igual a 0,69), é bastante promissor e encontra-se no nível de desempenho alcançado pelos outros classificadores retóricos da literatura [Burstein et al. 2003], [Anthony and Lashkia 2003], [Teufel and Moens 2002], encorajando, assim, o seu aperfeiçoamento.

Nesse contexto, este trabalho faz a análise dos atributos utilizados pelos sistemas AZ [Teufel and Moens 2002] e Critique [Burstein et al. 2003], visando a seleção de novos atributos e/ou a modificação dos atributos já utilizados pelo AZPort, buscando o aumento na precisão e na robustez do classificador.

2. Um novo conjunto de atributos: seleção, implementação e modificação

Do conjunto de 16 atributos utilizados pelo AZ original, decidimos implementar mais 2 atributos além dos 8 já implementados. São eles: (1) *Section structure*: Posição

da sentença dentro de uma seção (que pode ser útil na classificação de sentenças de introduções) e (2) *TF-IDF* (*Term Frequency - Inverse Document Frequency*): contagem de frequência que atribui altos valores para palavras que ocorrem frequentemente em um documento, mas raramente em toda a coleção de documentos.

A partir da análise dos atribuitos utilizados pelo sistema Critique, decidimos usar informações extraídas a partir da análise RST *Rhetorical Structure Theory* [Mann and Thompson 1988] das sentenças. Para realizar a análise RST automática do córpus, utilizamos o analisador discursivo DiZer [Pardo et al. 2004]. Por meio do processamento automático da saída do DiZer, extraimos os valores de 2 atributos: (1) '*Status*' ('n' ou 's'), que sinaliza se a sentença é núcleo ou satélite de uma relação e (2) 'Relação' que determina qual a última relação RST em que a sentença está envolvida.

Além desses 4 atributos novos, decidimos pela derivação de um quinto atributo a partir do atributo Histórico já existente no AZPort, que captura a categoria retórica atribuída à sentença anterior. Esse novo atributo, chamado Histórico-2, captura o valor atribuído ao atributo Histórico da sentença anterior.

3. Avaliação do desempenho dos novos atributos

No total foram implementados 5 atributos novos (RST Status, RST Relation, TF-IDF, Loc. na Seção, Histórico-2), além dos 8 atributos já existentes no AZPort (Histórico, Modal, Voz, Tempo, Expressão, Citação, Loc. no Parágrafo, Tamanho). Para a avaliação foi utilizado o pacote Weka (*Waikato Environment for Knowledge Analysis*) [Witten and Frank 2005]. Uma vez que o AZPort implementa o algoritmo Naive Bayes, esse algoritmo foi utilizado para testar o sistema com os novos atributos. Os testes foram realizados aplicando-se 10-*fold cross-validation* a um corpus de 52 resumos e 50 introduções de textos provenientes do CorpusDT.

Utilizando-se todos os atributos na classificação dos resumos (5 novos + 8 antigos) com o algoritmo Naive Bayes, a precisão ficou em 71.7%, com valor Kappa de 0,61, resultados similares aos obtidos quando se utilizou apenas os 8 atributos do AZPort original (precisão de 72% e Kappa de 0,62). É possível que exista dependência entre os atributos novos e antigos e, uma vez o que um algoritmo bayesiano foi utilizado, isso pode ter influenciado os resultados. Para as introduções, os resultados obtidos com o mesmo algoritmo foram semelhantes aos obtidos com os resumos.

Visando encontrar um subconjunto mínimo que consiga maximizar o poder de classificação, eliminando assim possíveis dependências e redundâncias, foi utilizado o *Wrapper Subset Evaluator* [Kohavi and John 1997]. Em testes para os resumos, realizados com o *Wrapper*,com o algoritmo Naive Bayes e método de busca *Greedy Stepwise*, obtivemos o subconjunto Citação, Expressão, Histórico, Loc. na Seção, TF-IDF e RST Relation como melhor conjunto mínimo (precisão de 75% e Kappa igual a 0,66). Para introduções, o melhor subconjunto foi Loc. no Parágrafo, Expressão e Histórico (precisão de 77% e Kappa igual a 0,71).

Em trabalho anterior [Fuverki and Feltrim 2008] foram testados outros algoritmos de aprendizado juntamente com os atributos do AZPort original na classificação de resumos. O algoritmo LMT (*Logistic Model Tree*) [Landwehr et al. 2005] foi o que obteve melhor resultado. Utilizando-se o LMT juntamente com os novos atributos, o

poder de classificação para resumos aumentou (precisão de 81,32% e Kappa igual a 0,75). Para introduções, o uso do LMT não melhorou os resultados (precisão 76% e Kappa igual a 0,70).

4. Conclusões

Este trabalho partiu da necessidade de aumentar o poder de classificação do AZPort, classificador que faz a detecção automática da estrutura retórica de resumos e introduções acadêmicas. Conseguimos aumentar o poder de classificação para 75% de precisão e Kappa igual a 0,66 para resumos com apenas 6 atributos dos 13 implementados e precisão de 77% e Kappa igual a 0,71 para introduções com apenas 3 atributos dos 13 implementados, utilizando o algoritmo *Naive Bayes* em ambos os casos. Destacamos que entre os atributos selecionados para os melhores conjuntos mínimos estão inclusos alguns dos atributos novos propostos neste trabalho, como o uso de relações RST.

Referências

Anthony, L. and Lashkia, G. V. (2003). Mover: A machine learning tool to assist in thereading and writing of technical papers. In *IEEE Transactions on Professional Communication*, pages 46(3):185–193.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the write stuff: Automatic Identification of discourse structure in student essays. In *IEEE Intelligent Systems: Special Issue on Natural Language Processing*, pages 18(1):32–39.

Feltrim, V., Teufel, S., Nunes, M., and Aluísio, S. (2006). Argumentative Zoning applied to critiquing novices' scientific abstracts. In *James G. Shanahan, Yan Qu and Janyce Wiebe (Eds.) Computing Attitude and Affect in Text.*, pages 233–246, São Luis-MA, Brazil. Dordrecht, The Netherlands: Springer.

Fuverki, D. and Feltrim, V. D. (2008). Uma Investigação sobre a Aplicação de Algoritmos de Aprendizado à Classificação de Papéis Retóricos. In *VIII Fórum de Informática e Tecnologia de Maringá e XI Mostra de Trabalhos de Informática*, pages 94–104, Maringá-PR, Brazil.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273 – 324. Relevance.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.

Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. . In *Text.*, pages 8(3):243–281.

Pardo, T., Nunes, M., and Rino, L. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence, (SBIA 2004)*, pages 224–234, São Luis-MA, Brazil. Lecture Notes in Artificial Intelligence, 3171, Springer.

Teufel, S. and Moens, M. (2002). Summarising scientific articles - experiments with relevance and rhetorical status. . In *Computational Linguistics.*, pages 28(4):409–446.

Witten, H. I. and Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques*.