Uma Investigação sobre Algoritmos de Diferentes Abordagens de Aprendizado Supervisionado na Classificação de Papéis Retóricos em Resumos Científicos

Vinícius M. A. de Souza[1] e Valéria D. Feltrim[2]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos – SP – Brasil

[2]Departamento de Informática (DIN) – Universidade Estadual de Maringá (UEM)
Maringá – PR – Brasil

vsouza@icmc.usp.br, valeria.feltrim@din.uem.br

Abstract. *This paper presents the results of an investigation concerning the performance of machine learning (ML) algorithms from different approaches when applied to the problem of categorize the rhetoric's role of scientific abstracts sentences. The used example sentences are from theses abstracts written in Portuguese, in the field of Computer Science. The purpose of this investigation was to identify the best ML algorithms for the implementation of systems for the automatic detection of rhetoric structures in scientific texts, aiming at the development of flexible and powerful classifiers. Results indicate that more representative and interpretable models can replace the classifier currently used without addition computational cost.*

Resumo. *Este artigo apresenta os resultados de uma investigação sobre o desempenho de algoritmos de diferentes abordagens de aprendizado de máquina (AM) quando aplicados ao problema da classificação do papel retórico de sentenças provenientes de textos científicos, mais especificamente de resumos de teses e dissertações escritos em português, na área da Ciência da Computação. Tal investigação teve como objetivo identificar os melhores algoritmos de AM a serem utilizados na implementação de sistemas para a detecção automática da estrutura retórica de textos científicos, visando à geração de classificadores robustos e flexíveis. Os resultados apontam que modelos mais representativos e interpretáveis podem substituir o algoritmo atual de classificação sem adição de custo computacional.*

1. Introdução

A natureza padronizada do discurso científico tem motivado várias propostas de modelos estruturais, em termos de componentes da estrutura retórica, para textos do gênero acadêmico [Huckin and Olsen 1991], [Weissberg and Buker 1990], [Swales 1990], [Trimble 1985]. Tais modelos podem ser aplicados em diferentes contextos, como no auxílio à escrita científica por meio de ferramentas computacionais [Aluísio et al. 2001], [Anthony and Lashkia 2003], [Feltrim et al. 2004], [Souza and Feltrim 2013].

Uma vertente de ferramentas computacionais de auxílio à escrita que tem sido explorada nos últimos anos está relacionada ao uso de classificadores automáticos da estrutura retórica [Burstein et al. 2003], [Anthony and Lashkia 2003], [Feltrim et al. 2004], [Gênoves et al. 2007], [Dayrell et al. 2012]. Tais classificadores utilizam algoritmos de aprendizado de máquina para atribuir automaticamente uma ou mais

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 30–39,
Fortaleza, CE, Brazil, October 21–23, 2013.

categorias para cada sentença, de acordo com o papel retórico desempenhado pela sentença (ou por suas partes) no texto. A classificação é feita com base em um conjunto de características extraído automaticamente a partir das sentenças e pode ser mono-rótulo, quando atribui uma única categoria a cada sentença, ou multi-rótulo, quando pode atribuir mais de uma categoria a cada sentença. De fato, classificadores dessa natureza têm sido aplicados não só em ferramentas de auxílio à escrita, mas também em tarefas como a avaliação da escrita (*scoring systems*) [Attali and Burstein 2006], [Yao 2013], sumarização automática [Teufel and Moens 2002], [Saggion and Poibeau 2013] e análise automática de coerência [Souza and Feltrim 2011].

Um classificador automático de estrutura retórica desenvolvido especificamente para resumos científicos escritos em português é o AZPort — *Argumentative Zoning for Portuguese* [Feltrim et al. 2004]. Esse classificador compõe o sistema de crítica do ambiente SciPo — *Scientific Portuguese* [Feltrim et al. 2006], [Souza and Feltrim 2013], um ambiente baseado em corpus e em sistemas de crítica voltado para o auxílio à escrita de resumos e introduções de dissertações e teses em Ciência da Computação. O AZPort utiliza um algoritmo Naïve Bayes similar ao utilizado pelo classificador AZ [Teufel and Moens 2002], porém adapta a extração dos atributos para a língua portuguesa e as categorias ao modelo estrutural aplicado a resumos de teses e dissertações pelo SciPo.

O classificador AZ implementa a técnica *Argumentative Zoning*, que segmenta automaticamente um artigo científico em "zonas argumentativas", isto é, trechos de texto que possuem determinados papéis retóricos dentro da estrutura retórica do texto como um todo. Essa técnica assume que o papel retórico de uma sentença pode ser "lido" da superfície do texto e representado por características que, posteriormente, poderão ser aprendidas por um algoritmo de aprendizado de máquina.

Embora outros autores tenham relatado resultados semelhantes aos alcançados com o algoritmo Naïve Bayes ao utilizar outros algoritmos de aprendizado para a tarefa de análise de estrutura retórica [Teufel and Moens 2002], [Anthony and Lashkia 2003], [Genovês et al. 2007], a realização de experimentos para a confirmação dessa hipótese no contexto do AZPort se fez necessária, em especial por dois motivos: (i) o modelo estrutural utilizado pelo AZPort é específico para resumos de teses e dissertações, enquanto os outros sistemas baseiam-se em modelos mais genéricos voltados para resumos de artigos científicos que, em geral, são textos mais curtos e condensados; e (ii) o AZPort trata textos escritos em português, enquanto os outros sistemas descritos na literatura tratam textos em inglês.

Desse modo, este artigo apresenta os resultados da investigação acerca do desempenho de algoritmos de diferentes abordagens de aprendizado de máquina supervisionado quando aplicados à classificação de papéis retóricos que compõem a estrutura de resumos científicos em português. Mais especificamente, utilizando-se o mesmo conjunto de atributos e as mesmas categorias empregadas pelo AZPort, foram realizados experimentos com algoritmos de classificação provenientes de cinco abordagens de aprendizado: (1) *bayesiano*; (2) baseado em funções; (3) baseado em instâncias; (4) baseado em regras; e (5) árvores de decisão. O principal objetivo desta investigação foi verificar a influência do algoritmo de aprendizado empregado nos resultados de classificação e verificar se o desempenho do algoritmo *bayesiano* atualmente empregado no AZPort e em outros sistemas similares da literatura de fato supera o desempenho obtido por algoritmos de outras abordagens de aprendizado.

2. Trabalhos Relacionados

Além do AZ [Teufel and Moens 2002] e do AZPort [Feltrim et al. 2004], outros sistemas que se propõem a realizar a classificação automática da estrutura retórica de resumos científicos são o Mover [Anthony and Lashkia 2003], o AZEA [Gênoves et al. 2007] e o MAZEA [Dayrell et al. 2012]. Embora a comparação direta entre os sistemas não seja possível devido a diferenças de língua, domínio e modelo estrutural, a análise dos resultados de cada um deles permite estabelecer uma margem de desempenho para a tarefa. Nesta seção é feita uma breve apresentação dos classificadores AZ, Mover e AZEA. O MAZEA não será abordado por se tratar de um classificador multi-rótulo, diferenciando-se dos demais na tarefa a que se propõe. O AZPort, por ser o classificador base dos experimentos descritos neste artigo, será apresentado na Seção 3.

O AZ foi desenvolvido com o objetivo de sumarizar artigos científicos e utiliza um conjunto de 16 atributos que busca caracterizar o papel retórico da sentença por meio da extração de conhecimento em nível léxico, sintático e estrutural. Como algoritmo de aprendizado é utilizado o Naïve Bayes. O sistema classifica cada sentença de um artigo científico em uma das categorias descritas na Tabela 1.

Tabela 1. Descrição das categorias (modelo estrutural) utilizadas pelo AZ

Categoria	Descrição
Propósito	Afirmações sobre o propósito do autor
Contraste	Afirmações sobre diferenças relativas a outros trabalhos
Base	Afirmações sobre a base teórica
Outro	Descrição neutra de outro trabalho
Contexto	Afirmações gerais aceitas na área
Estrutura	Afirmações sobre a estrutura do artigo
Próprio	Todas as outras afirmações relativas ao trabalho do autor

Os resultados obtidos pelo AZ foram estimados aplicando-se 10-*fold stratified cross-validation* em um corpus de 80 artigos publicados em conferências na área de Linguística Computacional e superaram as *baselines* utilizadas (a descrição das *baselines* pode ser encontrada em Teufel and Moens (2002)). Em termos de *Macro-F*, o AZ alcançou 0,50 e, em termos da medida *Kappa*, 0,45. Essas e outras medidas utilizadas na avaliação dos classificadores serão detalhadas na Seção 4.

O Mover, desenvolvido para auxiliar escritores não nativos na leitura e escrita de *abstracts* científicos, também emprega um classificador Naïve Bayes. Seu conjunto de atributos é modelado como um *bag of cluster*s, que é uma adaptação do modelo *bag of words* em que expressões contendo até cinco palavras são utilizadas em vez das palavras separadas. Em adição aos *clusters*, também é considerada a posição da sentença no texto para a realização de otimizações baseadas no modelo estrutural do *abstract*. O modelo estrutural usado nos experimentos relatados por Anthony and Lashkia (2003) é baseado no modelo *CARS* de Swales (1990), de modo que seu conjunto de categorias é um subconjunto das categorias previstas por Swales.

O Mover foi treinado com 554 sentenças e testado com 138 sentenças extraídas de um corpus com 100 *abstracts* de artigos da área de Tecnologia da Informação. Na avaliação o Mover obteve precisão média de 68% e superou as *baselines* consideradas: classificação aleatória (16,6%) e categoria mais frequente (26%).

O AZEA, desenvolvido para auxiliar na anotação de corpora e também no auxílio à escrita científica, utiliza um conjunto de 22 atributos para classificar sentenças de resumos de artigos científicos escritos em inglês em uma de seis categorias possíveis:

Background, *Gap*, *Purpose*, *Methodology*, *Results* e *Conclusion*. Essas categorias se diferenciam das categorias usadas pelo AZPort apenas pela exclusão da categoria *Outline*. O sistema utiliza um classificador de Máquinas de Vetores de Suporte.

Os resultados obtidos pelo AZEA foram estimados aplicando-se 10-*fold stratified cross-validation* em um corpus de 74 *abstracts* extraídos de artigos publicados em revistas científicas na área de Farmácia. O desempenho medido como a porcentagem de acerto foi de 80,39%, superando a *baseline* utilizada (associação da categoria mais frequente (44,86%)). Em termos das medidas *Kappa* e *Macro-F*, os resultados obtidos foram 0,73 e 0,78, respectivamente.

3. Argumentative Zoning for Portuguese – AZPort

Assim como o AZ e o Mover, o AZPort utiliza um classificador Naïve Bayes para estimar a probabilidade de que uma sentença *S* tenha a categoria *C*, dados os valores dos atributos extraídos de *S*. A categoria com a maior probabilidade é escolhida como saída para *S*. Os oito atributos utilizados pelo AZPort foram adaptados do conjunto de 16 atributos utilizados por Teufel and Moens (2002) e são descritos na Tabela 2.

Tabela 2. Descrição dos oito atributos utilizados pelo AZPort

Atributo	Descrição
Tamanho	Qual é o tamanho da sentença em comparação aos dois limiares (20 e 40 palavras)?
Localização	Qual a posição da sentença no resumo?
Citação	A sentença contém citações?
Expressão	Que tipo de expressão padrão a sentença contém?
Tempo	Qual o tempo do primeiro verbo finito da sentença?
Voz	Qual a voz do primeiro verbo finito da sentença?
Modal	O primeiro verbo finito da sentença é modal?
Histórico	Qual é a categoria da sentença anterior?

Todos os atributos são determinados automaticamente a partir do texto de entrada por meio de um *pipeline* de operações que envolve a identificação de expressões padrão (por ex., "o objetivo deste trabalho", "no entanto", etc) [Feltrim and Souza 2009], [Machado Jr and Feltrim 2009] e o uso de ferramentas de pré-processamento como *tokenizer*, delimitador de sentenças e etiquetador morfossintático.

Para o treinamento do classificador foram utilizados 52 resumos do CorpusDT, [Feltrim et al. 2003] manualmente anotados segundo um esquema pré-definido de sete categorias, a saber: (1) Contexto; (2) Lacuna; (3) Propósito; (4) Metodologia; (5) Resultado; (6) Conclusão; e (7) Estrutura. No total foram anotadas 366 sentenças. A distribuição das sentenças por categoria é apresentada na Tabela 3.

Tabela 3. Distribuição das sentenças anotadas de acordo com as categorias retóricas consideradas no modelo estrutural

Categoria	*Contexto*	*Lacuna*	*Propósito*	*Metodologia*	*Resultado*	*Conclusão*	*Estrutura*
Quantidade de sentenças (%)	77 (21,04%)	36 (9,84%)	65 (17,76%)	45 (12,29%)	117 (31,97%)	20 (5,46%)	6 (1,64%)

3.1. Avaliação Intrínseca do AZPort

Os resultados da classificação foram computados aplicando-se 13x13-*fold cross-validation* aos 52 resumos manualmente anotados, obtendo acurácia (% de acerto) igual a 72% e *Macro-F* igual a 0,71. Como comparação, foram utilizadas duas *baselines*: (1) escolha aleatória da categoria considerando-se a distribuição do corpus e (2) atribuição

da categoria mais frequente a todas as sentenças. Os resultados obtidos pelo classificador Naïve Bayes e pelas *baselines* são apresentados na Tabela 4.

Tabela 4. Resultados do AZPort em termos de acurácia e medida *Kappa*

Medida	AZPort	*Baseline 1*	*Baseline 2*
Acurácia (%)	**72**	20	32
Kappa	**0,65**	0	0,26

Conforme mostrado na Tabela 4, quando comparado a um anotador humano, a concordância entre eles medida por meio da medida *Kappa* foi de 0,65. Esse resultado é bastante encorajador se comparado com o resultado relatado por Teufel and Moens (2002) (K=0,45). Cabe destacar que existem diferenças significativas entre os dois sistemas quanto ao conjunto de categorias e ao corpus utilizados, o que pode justificar a diferença de desempenho. Uma análise detalhada dos resultados do AZPort incluindo o desempenho por categoria, por atributo e outras medidas de avaliação é apresentada em Feltrim et al. (2006).

Os resultados da avaliação intrínseca do AZPort mostraram que, embora seu desempenho esteja abaixo do desempenho humano para a mesma tarefa (que é em torno de K=0,69 para dois anotadores treinados), é bastante promissor e encontra-se no nível de desempenho alcançado por outros classificadores retóricos da literatura [Burstein et al. 2003], [Anthony and Lashkia 2003], [Genovês et al. 2007], encorajando o seu aperfeiçoamento. Investigações para tal aperfeiçoamento têm sido conduzidas de dois modos: (i) por meio da aplicação de outros algoritmos de aprendizado e (ii) por meio da investigação de novos atributos de classificação. Como a investigação (ii) ainda não foi concluída, o foco deste artigo são os resultados alcançados em (i).

4. Metodologia

A escolha dos algoritmos de aprendizado investigados neste trabalho foi norteada pela natureza do conjunto de dados utilizado pelo AZPort. Desse modo, foram selecionados algoritmos de aprendizado supervisionado que manipulam categorias e atributos nominais. Como material de treinamento e teste foram utilizadas as 366 sentenças empregadas na avaliação intrínseca e descritas na Tabela 3. Também foi utilizado o mesmo conjunto de atributos e procedimentos para a avaliação dos resultados.

Os algoritmos foram avaliados aplicando-se 13x13-*fold stratified cross-validation*, de modo que o conjunto de dados foi dividido em 13 subconjuntos mutuamente exclusivos de mesmo tamanho a partir da escolha aleatória dos dados. Em seguida, 12 subconjuntos foram utilizados para treinamento e o subconjunto remanescente foi utilizado para o teste do algoritmo. Esse procedimento foi repetido 13 vezes alternando-se os subconjuntos, de modo que todos sejam utilizados no teste. Por ser um procedimento de amostragem estratificada, as categorias são representadas com aproximadamente a mesma proporção tanto no teste como no treinamento. Para maior variação dos exemplos, o procedimento foi repetido 13 vezes alterando-se os dados aleatoriamente para a formação dos subconjuntos.

Os resultados obtidos foram expressos em termos das medidas *Macro-F* e *Kappa*. A medida *Macro-F* corresponde à média das *F-measures* de todas as categorias e é uma medida interessante para avaliar se o classificador não sacrifica o desempenho de uma ou outra categoria com poucos exemplos em troca de uma melhora na taxa de acerto. A medida *F-Measure* é a média harmônica das medidas *Precision* e *Recall*, sendo um modo conveniente de expressá-las em um único valor. A medida estatística

Kappa é empregada para determinar a concordância entre o classificador e a anotação manual. O valor de *Kappa* varia entre -1 e 1, sendo que -1 indica máxima discordância, 0 indica concordância ao acaso e 1 indica concordância perfeita. O uso da medida *Kappa* serve como um bom indicativo do desempenho da classificação quando o classificador automático é comparado com a anotação manual.

Com base nos algoritmos presentes no sistema Weka [Witten and Frank 2005], foram realizados experimentos que contemplaram cinco abordagens diferentes de aprendizado. Assim, para avaliar a abordagem *baysiana* foram analisados os algoritmos *NaiveBayes*, *NaiveBayesSimple* e *BayesNet*. Para a abordagem baseada em instâncias foi analisado o algoritmo *IBk*. Para a abordagem baseada em regras foram analisados os algoritmos *JRip*, *OneR*, *NNge* e *DecisionTable*. Para a abordagem baseada em funções foram analisados os algoritmos *Logistic*, *SimplesLogistic* e *SMO* (com 3 diferentes *kernels*). Por fim, para a abordagem baseada em árvores de decisão foram analisados os algoritmos *J48*, *RandomForest* e *Id3*. Desse modo, os experimentos abrangeram um total de 14 algoritmos diferentes e algumas de suas variações.

Vários dos algoritmos utilizados nos experimentos dependem de parâmetros, de modo que a otimização de seus valores pode exercer grande influência no desempenho. Por isso, durante os experimentos foi realizada uma etapa preliminar de variação dos principais parâmetros de cada algoritmo avaliado. Devido à escassez de dados para a formação de um conjunto de validação que pudesse ser utilizado exclusivamente na otimização dos parâmetros, optou-se pela variação e análise dos resultados em uma etapa de validação cruzada para a escolha dos melhores valores para os parâmetros.

Para confirmar a validade da hipótese de que determinado algoritmo apresenta de fato um desempenho superior ao algoritmo *bayesiano* Naïve Bayes, não basta apenas comparar as estimativas de erro obtidas por ambos os algoritmos, devido ao fato de ser possível que as diferenças nos resultados sejam causadas pelos procedimentos de validação utilizados no cálculo das estimativas. Assim, para uma melhor confiança na análise dos resultados, realizou-se também um teste de hipóteses. Em específico, para o contexto deste trabalho, foi realizado o Teste *T-Student* Pareado.

5. Resultados Experimentais

Os resultados obtidos na fase experimental após a etapa de otimização de parâmetros dos algoritmos são apresentados na Tabela 5. Nesta tabela é mostrada a abordagem de cada algoritmo avaliado, o nome atribuído pelo sistema Weka, os valores considerados para os principais parâmetros, os resultados obtidos pelas medidas *Macro-F* e *Kappa*, a quantidade de vitórias (>) e a quantidade de derrotas (<) de cada algoritmo sobre os demais considerando-se a medida *Macro-F* e 0,05% de nível de significância.

Pode ser observado na Tabela 5 que o algoritmo *J48* (árvores de decisão) apresentou o melhor desempenho nas duas medidas consideradas: *Macro-F* = 0,731 e *Kappa* = 0,673, superando o desempenho do algoritmo *NaiveBayes* utilizado pelo AZPort: *Macro-F* = 0,714 e *Kappa* = 0,656. Os algoritmos *DecisionTable* e *SimpleLogistic* apresentaram o segundo melhor desempenho, ambos com *Macro-F* = 0,727 e *Kappa* = 0,672 e 0,669, respectivamente. Os piores resultados foram obtidos pelos algoritmos baseados em regras *NNge* (*Macro-F* = 0,639 e *Kappa* = 0,562) e OneR (*Macro-F* = 0,650 e *Kappa* = 0,571).

Tabela 5. Desempenho dos algoritmos na classificação de papéis retóricos

Paradigma	Algoritmo	Valores dos parâmetros	Macro-F	Kappa	>	<
Bayesiano	BayesNet	– –	0,722	0,664	5	0
	NaiveBayes	– –	0,714	0,656	4	0
	NaiveBayesSimple	SimpleEstimator; a=0,5	0,714	0,656	4	0
Baseado em funções	SimpleLogistic	W=0,11; H=10	0,727	0,669	5	0
	Logistic	R=0,001	0,708	0,644	2	0
	SMO	Poly, C=1; e=1E-6	0,703	0,639	2	0
	SMO	RBF; C=3; P=1E-6; G=0,001	0,695	0,643	2	1
	SMO	Pearson; C=2; S=2; O=1; G=0,001	0,676	0,611	0	6
Baseado em instâncias	*IBk*	K=3, 1/dist	0,677	0,609	0	6
Baseado em regras	DecisionTable	– –	0,727	0,672	5	0
	JRip	N=1,06; F=3	0714	0,650	2	0
	OneR	B=1	0,650	0,571	0	10
	NNge	G=3; I=1	0,639	0,562	0	11
Árvores de decisão	J48	C=0,43	**0,731**	**0,673**	**7**	**0**
	RandomForest	I=30; K=4	0,689	0,620	1	1
	Id3	– –	0,675	0,598	0	4

É possível observar a partir da quantidade de vitórias (>) e de derrotas (<) listados na Tabela 5, que do total de 16 algoritmos avaliados, o algoritmo *J48* obteve resultado estatisticamente superior a sete outros: *SMO* (*kernel RBF* e *Pearson*), *IBk*, *Id3*, *RandomForest*, *NNge* e *OneR,* e resultado estatisticamente inferior a nenhum dos outros algoritmos. Os algoritmos *DecisionTable* e *SimpleLogistic* apresentaram resultados estatisticamente superiores a cinco algoritmos e resultados inferiores a nenhum dos outros algoritmos. Já o algoritmo de comparação *NaiveBayes*, obteve resultado estatisticamente superior a quatro outros algoritmos. Uma quantidade menor de vitórias se comparado aos algoritmos *J48*, *DecisionTable* e *SimpleLogistic*.

Os resultados descritos podem sugerir que os algoritmos *J48*, *DecisionTable* e *SimpleLogistic* apresentam desempenhos superiores ao *NaiveBayes*. Entretanto, ao aplicar um teste estatístico com 0,05% de nível de significância, nota-se que nenhum deles apresenta diferença no desempenho que seja estatisticamente significante.

Assim, o critério de escolha do algoritmo para um classificador como o AZPort deve considerar outros fatores além do desempenho, como a representatividade e a interpretabilidade do modelo gerado. Tais características podem, por exemplo, auxiliar anotadores humanos na extensão do conjunto de dados com maior confiança de anotação. Outro fator a ser considerado é a complexidade do algoritmo, que pode impactar no desempenho do sistema de classificação em termos do custo de execução.

Para o conhecimento dos erros cometidos pelos classificadores em termos das categorias retóricas, selecionamos o resultado de uma das execuções do algoritmo *J48* por ser um algoritmo que apresentou desempenho competitivo e ter a vantagem de apresentar um modelo facilmente interpretável. Assim, os erros cometidos por esse classificador são expressos na matriz de confusão apresentada na Tabela 6. É possível observar, por exemplo, que as maiores dificuldades do classificador se encontram na distinção das categorias Estrutura vs. Resultado e Conclusão vs. Resultado. Esse tipo de erro pode ser atribuído à presença de expressões padrão típicas da categoria Resultado nas categorias Estrutura e Conclusão, principalmente em sentenças mais longas, bem como ao alto desbalanceamento entre essas categorias (conforme pode ser observado na Tabela 3).

Tabela 6. Matriz de confusão obtida por uma execução de teste do algoritmo *J48* com C=0,43

		Predito						
		Contexto	*Lacuna*	*Propósito*	*Metodologia*	*Resultado*	*Conclusão*	*Estrutura*
Real	*Contexto*	**60**	9	1	1	6	0	0
	Lacuna	3	**31**	0	0	2	0	0
	Propósito	5	0	**47**	2	10	1	0
	Metodologia	1	0	1	**27**	16	0	0
	Resultado	3	1	4	10	**96**	3	0
	Conclusão	0	0	1	0	9	**10**	0
	Estrutura	0	0	0	0	3	0	**3**

7. Conclusões e Trabalhos Futuros

Este trabalho apresentou os resultados de uma investigação sobre o desempenho de algoritmos pertencentes a diferentes abordagens de aprendizado de máquina aplicados ao problema da classificação de papéis retóricos em sentenças de resumos científicos escritos em português. Após uma etapa de otimização de parâmetros, foram realizados experimentos para a avaliação de 14 algoritmos aplicados a um corpus de 52 resumos.

Na avaliação dos resultados experimentais, os testes de hipóteses mostraram que não há diferenças estatísticas entre o classificador *bayesiano* atualmente implementado no AZPort e os algoritmos *J48, DecisionTable* e *SimpleLogistic*. Entretanto, o uso de modelos de classificação baseados em árvores de decisão ou em regras apresentam vantagens que devem ser consideradas quando comparadas a modelos estatísticos (*bayesianos* ou baseados em função). Tais vantagens se devem principalmente ao fato de algoritmos estatísticos induzirem modelos "caixa preta", em que a representação do conhecimento está implícita no modelo aprendido. Por outro lado, algoritmos que induzem modelos baseados em árvores de decisão ou em regras são facilmente interpretáveis e fornecem uma representação expressiva dos dados que podem ser úteis para a compreensão dos modelos estruturais utilizados nos textos científicos. Modelos mais legíveis também podem auxiliar a anotação manual de novos textos.

A busca por melhores resultados pode ser conduzida de diferentes modos em trabalhos futuros. Primeiro, nota-se que o conjunto de dados é pequeno e relativamente desbalanceado, sendo uma extensão natural deste trabalho a coleta de novos dados e o uso de técnicas para o balanceamento das classes. Em segundo, atualmente o AZPort realiza as classificações sentença-a-sentença sem considerar uma visão global do resumo analisado, tendo apenas o atributo *Histórico* como indicativo do fluxo retórico do texto. Assim, o uso de modelos como *Hidden Markov Models* [Baum et al. 1970] que consideram a ordem das sentenças durante o processo de classificação pode levar a melhores resultados. A inclusão de um novo atributo que indique a categoria da próxima sentença (*Futuro*) também pode ser considerada. Para isso, pode-se utilizar um algoritmo que realize mais de uma rodada de classificações por todo o resumo, de modo que essas classificações sejam utilizadas para o cálculo dos atributos *Histórico* e *Futuro*.

É comum, principalmente em sentenças mais longas, que mais de uma categoria retórica caracterize uma única sentença. Esse fato exerce influência até mesmo nos resultados de anotações humanas. Assim, a construção de classificadores multi-rótulos é uma vertente importante para sistemas como o AZPort. De fato, essa vertente já vem sendo explorada para a língua inglesa, como é caso do MAZEA – *Multi-label Argumentative Zoning for English Abstracts* [Dayrell et al. 2012].

Referências

Aluísio, S.M.; Barcelos, I.; Sampaio, J.; Oliveira Jr., O.N. (2001). "How to Learn the Many Unwritten "Rules of the Game" of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases", In: Proceedings of the IEEE International Conference on Advanced Learning Technologies, p. 257–260.

Anthony, L.; Lashkia, G.V. (2003). "Mover: A Machine Learning Tool to Assist in the Reading and Writing of Technical Papers", In: IEEE Transactions on Professional Communication, 46(3), p. 185–193.

Attali, Y.; Burstein, J. (2006). "Automated essay scoring with e-rater v.2", In: *Journal of Technology, Learning and Assesment*, 4(1), p. 1–31.

Baum, L.E; Petrie, T.; Soules, G.; Weiss, N. (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains", In: *Annals of Mathematical Statistics*, v. 41, p. 164–171.

Burstein, J.; Marcu, D.; 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, 18(1), p. 32–39.

Dayrell, C.; Candido Jr, A.; Lima, G.; Machado Jr, D.; Copestake, A.; Feltrim, V.D.; Tagnin, S.; Aluisio, S. (2012). "Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora", In: Proceedings of the 8th International Conference on Language Resources and Evaluation, p. 1604–1609.

Feltrim, V.D.; Aluísio, S.M.; Nunes, M.G.V. (2003). "Analysis of the Rhetorical Structure of Computer Science Abstracts in Portuguese", In: Proceedings of the Corpus Linguistics 2003, Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (Eds.), *UCREL Technical Papers*, Vol 16, Part 1, Special Issue. p. 212-218.

Feltrim, V.D.; Pelizzoni, J.M.; Teufel, S.; Nunes, M.G.V.; Aluísio, S.M. (2004). "Applying Argumentative Zoning in an Automatic Critiquer of Academic Writing", In: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence, São Luis-MA, Brazil. *Lecture Notes in Artificial Intelligence*, 3171, Springer, p. 214-223.

Feltrim, V.D.; Teufel, S.; Nunes, M.G.V.; Aluísio, S.M. (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*. Dordrecht, The Netherlands: Springer, p. 233-246.

Feltrim, V.D.; Souza, V. M. A. (2009). "PyGER: Uma Ferramenta Geradora de Expressões Regulares a partir de um Conjunto de Expressões em Linguagem Natural", In: Proceedings of the 1st Student Workshop on Information and Human Language Technology (TILic – STIL), 2009, São Carlos – SP, p. 1–5.

Genovês Jr., L.; Feltrim, V.D.; Dayrell, C.; Aluísio, S. (2007) Automatically detecting schematic structure components of English abstracts. In Proceedings of the Recent Advances in Natural Language Proceesing, Workshop on Natural Language Processing for Educational Resources, Borovets, Bulgaria, pp. 23-29.

Huckin, T.N.; Olsen, L.A. (1991). "Technical Writing and Professional Communication For Nonnative Speakers of English", New York, USA: McGraw-Hill.

Machado Jr, D.; Feltrim, V.D. (2009). "Extração Automática de Expressões Indicativas para a Classificação de Texto Científicos", In: Proceedings of the 1st Student Workshop on Information and Human Language Technology (TILic – STIL), 2009, São Carlos – SP, p. 1–5.

Saggion, H.; Poibeau, T. (2013). "Automatic Text Summarization: Past, Present and Future", In: Multi-source, Multilingual Information Extraction and Summarization - Theory and Applications of Natural Language Processing, p. 3–21.

Souza, V.M.A.; Feltrim, V.D. (2011). "Automatic Analysis of Semantic Coherence in Academic Abstracts Written in Portuguese", In: Proceedings of the 5th International Joint Conference on Natural Language Processing, p. 1144–1152.

Souza, V.M.A.; Feltrim, V.D. (2013). "A coherence analysis module for SciPo: providing suggestions for scientific abstracts written in Portuguese", In: *Journal of the Brazilian Computer Society*, 13(1), p. 59–73.

Swales, J. (1990). "Genre Analysis: English in Academic and Research Settings", Cambridge, UK: Cambridge University Press.

Teufel, S.; Moens, M. (2002). "Summarising Scientific Articles — Experiments with Relevance and Rhetorical Status", In: *Computational Linguistics*, 28(4), p. 409–446.

Trimble, L. (1985). "English for Science and Technology: A Discourse Approach", Cambridge, UK: Cambridge University Press.

Witten, H. I.; Frank, E. (2005) "Data mining - practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, Elsevier.

Yao, X. M. (2013). "Automated Essay Scoring: A Comparative Study", In: *Applied Mechanics and Materials*, v. 274, p. 650–653.