CRIAÇÃO DE CLASSIFICADORES RETÓRICOS PARA RESUMOS CIENTÍFICOS DA PUBMED

Alessandra Harumi Iriguti (PIC-UEM), Valéria Delisandra Feltrim (Orientadora), e-mail: vfeltrim@din.uem.br.

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

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: processamento de linguagem natural, aprendizado de máquina, classificação retórica

Resumo

Conhecer a estrutura retórica de um gênero textual auxilia a escrita e a compreensão de textos do gênero. Desse modo, sistemas que classificam o texto de acordo com um conjunto pré-definidos de categorias retóricas podem ser aplicados em vários contextos. O objetivo deste trabalho foi construir classificadores retóricos sentenciais para resumos científicos escritos em português e inglês. Os resumos em inglês foram extraídos da base de dados MEDLINE/PubMed e os resumos em português foram extraídos de uma biblioteca de teses e dissertações. Foram avaliados diferentes algoritmos de aprendizagem de máquina combinados com atributos superficiais por meio de validação cruzada. O algoritmo que apresentou melhor desempenho para ambas as línguas foi o SVM, com F1 média de 0,94 para o inglês e 0,57 para o português.

Introdução

Os elementos de um discurso e sua organização são chamados de estrutura retórica. Uma estrutura retórica é formada por um conjunto de relações retóricas e são essas relações que definem como o conteúdo está relacionado e como cada parte do texto contribui para satisfazer as intenções do autor (ROMEIRO, 2016). A estrutura retórica pode ser representada de forma mais geral (para qualquer discurso) ou mais específica (para um gênero textual em particular).

Weissberg e Buker (1990) investigaram estruturas retóricas específicas do gênero científico e propuseram um modelo para a estruturação de resumos. A motivação dos autores foi auxiliar a escrita de textos científicos, uma tarefa reconhecidamente difícil, especialmente para escritores iniciantes.

Com base em modelos de estrutura retórica pré-definidos, classificadores automáticos podem ser construídos por meio de aprendizado de máquina supervisionado. A partir de um conjunto de exemplos rotulados, algoritmos indutores são usados para treinar um classificador capaz de reconhecer as categorias (ou rótulos) presentes no conjunto.

O objetivo deste trabalho foi construir classificadores retóricos para resumos científicos com base em atributos que pudessem ser extraídos da superfície do texto, como n-gramas e posição da sentença. Os experimentos foram feitos com corpora em inglês e português. Por meio de experimentos com corpora de tamanhos variados, buscou-se também avaliar o impacto do tamanho do corpus de treinamento na classificação retórica.

Materiais e métodos

Ao todo foram usados cinco corpora de resumos. Dois deles foram extraídos da base MEDLINE/PubMed e são compostos por resumos escritos em inglês. Os outros três se encontram na língua portuguesa e são constituídos de resumos de teses e dissertações (ANDREANI; FELTRIM, 2015).
Os dois corpora em inglês estão anotados com as categorias retóricas *conclusion*, *method*, *objective* e *result*. O primeiro, chamado *dev*, contém 8.341 resumos (90.364 sentenças) e o segundo, chamado *data*, contém 93.909 resumos (922.149 sentenças). Os corpora em português estão anotados com as categorias contexto, lacuna, propósito, método, resultado, conclusão e estrutura. Os *corpora* 366 e 466 possuem 52 resumos cada um (366 e 466 sentenças, respectivamente). O *corpus 832* corresponde à junção dos dois anteriores (104 resumos e 832 sentenças). A distribuição de categorias nos cinco corpora é apresentada na Tabela 1.

Tabela 1 – Distribuição de categorias dos corpora

	N° de sentenças			N° de sentenças		
Categoria	***dev***	***data***	Categoria	***366***	***466***	***832***
Objectives	13.325	132.517	Contexto	77	179	256
Methods	26.241	255.730	Lacuna	36	36	72
Results	35.366	363.734	Propósito	65	68	133
Conclusions	15.432	170.168	Método	45	59	104
			Resultado	117	103	220
			Conclusão	20	20	40
			Estrutura	6	1	7

Para a indução dos classificadores foram usados os seguintes algoritmos de aprendizado: *Support Vector Machine* (SVM), *Nearest Neighbors* (K-NN), *Naive Bayes* (NB) e *Decision Trees* (DT). Todos os algoritmos foram importados da biblioteca Python *scikit-learn* (http://scikit-learn.org) e foram usados com parâmetros *default*. O SVM foi usado com *kernel* linear e o K-NN foi testado com o número de vizinhos igual a 15.
Como atributos foram usados valores de TF-IDF calculados para unigramas, bigramas e trigramas, bem como a posição absoluta da sentença no resumo. Também foram adicionados aos vetores os atributos da sentença anterior e da posterior. Para diminuir a dimensionalidade dos vetores, os K melhores atributos foram selecionados por meio do teste chi^2.

Para cada *corpus*, vários experimentos foram realizados variando-se tanto o algoritmo empregado quanto a configuração dos vetores de atributos. Em todos os casos, os resultados foram estimados por meio de validação cruzada de 10 partições.
Devido a limitação de memória do computador usado nos experimentos, os algoritmos K-NN e DT não foram aplicados aos corpora *dev* e *data*. Por esse mesmo motivo, não foram usados trigramas para o corpus *data*.

Resultados e Discussão

A Tabela 2 apresenta os resultados dos experimentos em termos do valor médio de medida F1. Nessa tabela, X_1 representa o algoritmo X utilizando como atributos os valores de TF-IDF da sentença, das sentenças anterior e da posterior, e a posição da sentença; X_2 representa o algoritmo X utilizando como atributos os valores de TF-IDF da sentença e a sua posição.

Tabela 2 – Valores médios de F1 obtidos pelos algoritmos em cada corpus

	dev	***data***	***366***	***466***	***832***
SVM_1	0,92	0,94	0,57	0,57	0,57
SVM_2	0,90	0,91	0,56	0,54	0,55
NB_1	0,88	0,90	0,34	0,41	0,38
NB_2	0,84	0,87	0,34	0,39	0,37
K-NN_1			0,40	0,34	0,38
K-NN_2			0,40	0,33	0,40
DT_1			0,52	0,49	0,49
DT_2			0,55	0,50	0,48

Na Tabela 2 observa-se que o SVM com maior número de atributos (SVM_1) obteve, em média, os melhores resultados, tanto para os corpora em inglês como em português. Para os corpora *dev* e *data*, esses resultados foram alcançados usando bigramas sem seleção de atributos. Já para os corpora 366, 466 e 832, esses resultados foram obtidos usando unigramas e seleção de K = 500 atributos. A diferença de tamanho dos vetores de atributos que levaram aos melhores resultados nos corpora em inglês e português se deve principalmente à grande diferença existente no número de sentenças desses corpora. Enquanto o maior corpus em inglês (*data*) possui aproximadamente 922 mil sentenças, o maior corpus em português possui apenas 832.
Como pode ser visto na Tabela 2, a diferença de tamanho entre os corpora também se refletiu no desempenho dos classificadores. Em comparação com F1 média do SVM para o *corpus data*, os resultados para o português foram 39,4% mais baixos. Cabe destacar, no entanto, que a classificação dos resumos em inglês pode ser vista como uma tarefa mais simples, uma vez que o número de categorias usadas na anotação desses resumos é menor do que a usada na anotação dos resumos em português.

Conclusões

O objetivo deste trabalho foi construir classificadores retóricos para resumos científicos que usassem atributos superficiais. Usando-se corpora em inglês e português, foram avaliados diferentes algoritmos de aprendizado, juntamente com diferentes configurações de vetores de atributos.

Para todos os corpora, os melhores resultados foram obtidos com o algoritmo SVM. A melhor média de medida F1 (0,94) foi alcançada para o *corpus data* quando se usou bigramas na geração dos valores TF-IDF e se incluiu os atributos das sentenças anterior e posterior nos vetores de atributos. Os experimentos mostraram que a alta dimensionalidade dos vetores influenciou positivamente os resultados para o corpus *data*, o que pode ser atribuído ao número elevado de sentenças desse *corpus*.

Para os corpora em português, a melhor média de medida F1 foi de 0,57. Esses resultados foram alcançados quando os vetores foram construídos a partir de unigramas e foi feita a seleção dos 500 melhores atributos por meio do teste chi^2. Uma vez que mesmo o maior *corpus* em português é pequeno, uma configuração que gerou vetores de dimensionalidade mais baixa foi mais efetiva para esses corpora.

Comparativamente, o resultado para o português foi 39,4% mais baixo do que o resultado para o inglês. Embora a tarefa de classificação em inglês tenha sido mais simples devido às categorias usadas, essa diferença mostra o quanto o tamanho da amostra de treinamento impacta nos resultados dos classificadores.

Referências

ANDREANI, A. C.; FELTRIM, V. D. Campos Aleatórios Condicionais Aplicados à Detecção de Estrutura Retórica em Resumos de Textos Acadêmicos em Português. In: **Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology**, 2015, p. 111 – 120.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Introdução às Máquinas de Vetores Suporte (Support Vector Machine). **Relatórios Técnicos do ICMC**, Nº 192, São Carlos, 2003, 58p.

REZENDE, S. O. **Sistemas Inteligentes: Fundamentos e Aplicações.** Editora Manole Ltda, 1ª Edição, 2003, 525p.

ROMEIRO, A. K. Q. Um estudo sobre o uso da teoria da estrutura retórica (RST) para sumarizar a sabedoria da coletividade. **Dissertação** (Mestrado em Computação), Universidade Federal Fluminense, Niterói – RJ, 2016, 109p.

WEISSBERG, R.; BUKER, S. **Writing up research: experimental research report writing for students of english**. Englewood Clis, NJ: Prentice Hall Regents, 1990, 202p.