Classificação automática de redações por tema

Rafael de Souza Freire (PIBIC/CNPq/FA/Uem), Valéria Delisandra Feltrim(Orientadora), e-mail: ra101971@uem.br

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

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: redações**,** análise de semântica latente, processamento de linguagem natural

Resumo

A avaliação automática de redações é um campo de estudo multidisciplinar envolvendo pesquisas em áreas como psicologia cognitiva, educação, linguística e computação. Do ponto de vista computacional, os estudos focam na construção de recursos e ferramentas que automatizem a avaliação de diferentes aspectos do texto. Dentre os aspectos textuais avaliados está a pertinência do conteúdo escrito a um tema. Assim, ferramentas que avaliem a redação quanto à sua aderência ao tema dado podem ser diretamente aplicadas nesses cenários. Dada a demanda gerada pelo grande volume de redações circulantes em escolas e processos seletivos por todo o país, o objetivo deste projeto foi explorar métodos para a modelagem de tópicos visando à classificação de redações por temas. Como resultado, foi elaborado um classificador tal que, dado um conjunto de temas, classifica redação em um dos temas dados.

Introdução

A avaliação automática de redações (AAR) é um campo de estudo multidisciplinar que envolve pesquisas em áreas como psicologia cognitiva, educação, linguística e ciência da computação (SHERMIS et al., 2013). Do ponto de vista computacional, os estudos focam na construção de recursos e ferramentas que automatizem (ou auxiliem) a avaliação de diferentes aspectos do texto, tais como ortografia e gramática, uso da língua, estilo, organização textual, conteúdo, entre outros. O contexto de aplicação dessas ferramentas vai desde o auxílio ao ensino-aprendizagem até o seu uso como parte do processo de avaliação de grandes volumes de redações.
Dentre os aspetos textuais avaliados por ferramentas de AAR está a pertinência do conteúdo textual a um tema. Cabe destacar que fugir ao tema proposto é um dos motivos da obtenção de nota zero nas redações escritas como parte de alguns exames, como é o caso do Exame Nacional do Ensino Médio (ENEM) (BRASIL, 2017). Assim, ferramentas que avaliem a redação

quanto a sua pertinência ao tema dado podem ser diretamente aplicadas nesses cenários.
Nesse contexto, este projeto teve como objetivo explorar métodos para a modelagem de tópicos visando à classificação de redações em temas. Especificamente foi investigado o uso de análise de semântica latente (LSA). Como resultado, foi elaborado um classificador que, dado um conjunto de temas, incluindo o título e texto de apoio do tema, determina o tema ao qual a redação pertence. Detalhes sobre o classificador construído são apresentados a seguir.

Materiais e métodos

Corpora

O desenvolvimento deste trabalho foi feito utilizando quatro *corpora* em português, sendo um deles composto por redações do ENEM e três deles compostos por notícias e textos da Wikipédia.
O *corpus* de redações foi construído por Guilherme Passero e está disponível a partir da plataforma GitHub. Esse *corpus* contém redações extraídas do Banco de redações da UOL e conta com 2.164 redações distribuídas em 112 temas, apresentando uma média de 19 redações por tema. Além do texto original das redações, o *corpus* também inclui, para cada redação: o título e o texto de apoio do tema; uma versão corrigida da redação por revisores da UOL; e a nota atribuída pelos revisores segundo os critérios usados pelo ENEM, com a diferença de que a nota disponibilizada corresponde à nota do ENEM dividida por 100.
Os outros três *corpora*, que a partir deste ponto será chamado de *corpus* NILC, foram construídos para o trabalho de Hartmann et al. (2017) e disponibilizados diretamente aos autores. São eles: o *corpus* G1, contendo notícias do portal G1; o *corpus Google News*, contendo notícias do agregador de notícias do *Google*; e o *corpus Wikipedia*, contendo o texto de páginas extraídas da Wikipédia. Juntos, esses *corpora* somam aproximadamente 485 milhões de palavras.
O pré-processamento dos *corpora* consistiu na remoção de palavras que ocorriam somente uma vez e de *stopwords*. *Stopwords* são palavras sem conteúdo, por exemplo, artigos e preposições. Além disso, foi feita uma filtragem do *corpus* de redações por nota, uma vez que redações com notas muito baixas podem fugir aos seus respectivos temas e distorcer os resultados de similaridade. A versão filtrada do *corpus* possui redações com notas iguais ou maiores que 6,0, totalizando 700 redações.

Abordagem

A abordagem adotada neste projeto foi baseada em análise de semântica latente (LSA) (FOLTZ et al., 1999) e pode ser descrita como segue. A partir de um *corpus,* usa-se LSA para construir um espaço semântico. Posteriormente, usando-se o espaço construído, estima-se a similaridade

semântica entre uma redação e um conjunto de temas. A similaridade semântica é dada pela similaridade do cosseno entre os vetores das redações no espaço semântico com os vetores do texto de apoio e título de cada tema do conjunto. Por fim, seleciona-se o tema com maior similaridade como resultado para a redação que está sendo classificada.
Todas as implementações foram feitas usando a linguagem de programação Python e o *framework* Gensim.

Experimentos

A abordagem proposta foi avaliada em dois experimentos. No primeiro, o *corpus* NILC foi usado para a construção do espaço semântico e o *corpus* de redações filtrado foi usado na avaliação. No segundo, o *corpus* de redações filtrado foi usado tanto na construção do espaço semântico quanto na avaliação. No caso do segundo experimento, é importante destacar que quando uma redação foi usada na avaliação, ela não estava presente no *corpus* usado para se construir o espaço semântico. Assim, o espaço semântico foi gerado 700 vezes: em cada vez, usou-se 699 redações para a geração do espaço e uma redação para teste. Nos dois experimentos, foram gerados espaços semânticos com 200 e 400 dimensões.

Resultados e Discussão

Os resultados dos experimentos estão sumarizados na Tabela 1. Os *corpora* indicados nas colunas correspondem aos usados na construção do espaço semântico de cada experimento. Os textos mencionados nas linhas são os do *corpus* de redações filtrado. Dessa forma, os valores mostrados correspondem ao acerto obtido sobre um total de 700 redações e 112 temas.

Tabela 1 – Acertos para o *corpus* de redações filtrado

Total/Percentual de acerto	*Corpus* NILC		*Corpus* de redações	
	200 dim.	400 dim.	200 dim.	400 dim.
Textos originais	95/13,57%	124/17,71%	165/23,57%	219/31,29%
Textos corrigidos	95/13,57%	122/17,43%	163/23,28%	211/30,14%

Como pode ser observado, os melhores resultados foram obtidos utilizando espaços semânticos de 400 dimensões. Com o *corpus* NILC como espaço semântico, o classificador acertou o tema de 124 redações, o que corresponde a uma acurácia de 17,71%. Usando o espaço gerado com o *corpus* de redações, o classificador melhorou o desempenho, acertando o tema de 219 redações e apresentando acurácia de 31,29%. Observa-se ainda que, em todos os casos, os resultados obtidos com os textos corrigidos foram similares aos obtidos com os textos originais.
A partir de uma análise dos erros do classificador, observou-se que, em certos casos de erro, há intercessão entre o tema abordado na redação e o tema selecionado pelo classificador para a redação. Por exemplo, a

classificação correta para a redação de título “O peso da pós-verdade na sociedade pós-moderna” seria o tema “Pós-verdade, opinião pública e democracia”, porém o classificador selecionou o tema “Há limites para a liberdade de expressão?”. Nota-se, nesse caso, que há uma conexão semântica entre os temas e partes do que foi abordado na redação de fato se relacionam com ambos os temas. Outro fator relacionado aos erros do classificador é o tamanho do texto de apoio. Textos de apoio menores fornecem menos informação semântica e dificultam a classificação.

Conclusões

Este trabalho apresentou um classificador de redações em possíveis temas com base em análise de semântica latente. O classificador foi avaliado com um *corpus* de redações do ENEM e experimentou-se a construção de espaços semânticos de diferentes dimensões a partir de *corpora* de domínios variados. Os resultados mostraram que o domínio do *corpus* usado como espaço semântico foi mais importante do que o seu tamanho. Mesmo menor do que o *corpus* NILC, o *corpus* de redações com 400 dimensões obteve o melhor resultado como espaço semântico. A partir da análise dos erros de classificação, foi possível identificar dois aspectos que contribuem para o aumento da complexidade da classificação: a sobreposição semântica que existe entre temas e a diferença no tamanho dos textos de apoio usados para representar os temas.

Agradecimentos

Agradeço a minha orientadora e a UEM pelo incentivo e oportunidade.

Referências

SHERMIS, M. D.; BURSTEIN, J.; BURSKY, S. A. Introduction to Automated Essay Evaluation. In: SHERMIS, M. D.; BURSTEIN, J. (Org.). **Handbook of Automated Essay Evaluation: Current Applications and New Directions**. New York: Taylor & Francis, 2013, p. 1-15.

BRASIL. Ministério da Educação. **Redação no ENEM 2017: Cartilha do Participante**, 2017.

FOLTZ, P. W.; LAHAM, D.; LANDAUER, T. K. The intelligent essay assessor: Applications to educational technology. **Interactive Multimedia Electronic Journal of Computer-Enhanced Learning**, 1(2), p. 939–944, 1999.

HARTMANN, N. S.; FONSECA, E. R.; SHULBY, C. D.; TREVISO, M. V.; RODRIGUES, J. S.; ALUÍSIO, S. M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2017. **Proceedings...** 2017. p.122–131.