Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Estudo exploratório de categorias gramaticais com potencial de indicadores para a Análise de Sentimentos

Júlia Santos Nunes Rodrigues, Adriana S. Pagano, Emerson Cabrera Paraiso

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

Pontifícia Universidade Católica do Paraná (PUCPR) – Curitiba, PR – Brasil

juliasnrodrigues@ufmg.br, apagano@ufmg.br, paraiso@ppgia.pucpr.br

Resumo: *Este trabalho apresenta uma pesquisa em andamento sobre categorias gramaticais que podem ser exploradas como indicadores de emoção em textos escritos. Diferentemente de pesquisas sobre análise de sentimentos que se concentram em itens lexicais, este estudo baseia-se na gramática sistêmico-funcional [Halliday e Matthiessen 2014] a fim de mapear padrões de escolhas em sistemas gramaticais que podem ser associadas à construção de emoções na linguagem. A metodologia baseia-se na anotação manual de amostras de textos do tipo notícia por meio de planilhas de acordo com categorias dos principais sistemas gramaticais no nível da oração da escala de ordens. A frequência das categorias anotadas e dos agrupamentos das mesmas é investigada para verificar quais as categorias mais produtivas para a análise de sentimentos.*

Abstract: *This paper reports on work in progress on grammatical categories that may be explored as indicators of emotion in written text. Unlike state-of-the-art research on sentiment analysis focused on lexical items, this study draws on systemic-functional grammar [Halliday & Matthiessen 2014] in order to map patterns of choice in grammatical systems that can be linked to emotion construal in language. The methodology is based on manual annotation of news report text samples carried out on a spreadsheet with categories pertaining to the main grammatical systems at clause level in the rank scale. Frequency of annotated individual categories and category clusters is examined with a view to identifying the most productive categories to probe sentiment in text.*

1. Introdução

A anotação de textos em pesquisas sobre Análise de Sentimentos é feita com intervenção de seres humanos, visando-se o aprendizado de máquina para rotulação sem intervenção humana. Em geral, a anotação é ad-hoc por não estar pautada em teorias linguísticas suficientemente abrangentes para explicar os distintos recursos que na linguagem humana constroem sentimentos. O índice de concordância entre anotadores tende a ser baixo, o que retarda a criação de um sistema automático eficaz para a Análise de Sentimentos [Dosciatti et al. 2015].

Há também pesquisas que fazem uso da prosódia semântica das palavras para auxiliar a anotação automática. São desenvolvidos glossários de palavras com distribuição de valores númericos que indicam a intensidade da positividade ou da negatividade da palavra no co-texto do *corpus* [Taboada et al. 2011]. Os estudos privilegiam o polo lexical do que pode ser teorizado como um contínuo, sendo o polo gramatical, sistemas que dizem respeito ao Modo, à Transitividade e ao Tema-Rema da oração.

Este trabalho explora uma metodologia para a Análise de Sentimentos baseada numa teoria linguistica abrangente que contempla o polo lexical e o polo gramatical. Prevê uma etapa inicial de anotaçao por humanos, a qual subsidiará uma futura implementaçao automática, contribuindo para que o processo de anotação seja menos subjetivo e utilize um aporte teorico linguístico que fundamente as escolhas do anotador.

2. As Emoções Humanas e a Análise de Sentimentos

Segundo [Ekman 1970] há seis emoções básicas universais – alegria, surpresa, medo, raiva, repugnância e tristeza, as quais podem funcionar como rótulos para a análise de textos no escopo da Análise de Sentimentos.

De acordo com [Dosciatti et al. 2015] a Análise de Sentimentos é uma área de pesquisa voltada para o estudo e identificação de emoções em diferentes midias, que surgiu da necessidade de se buscar, de forma automática, opiniões manifestadas na *internet*. Tal necessidade requer a análise de textos e a identificação de itens na linguagem que apontem para avaliações por parte dos usuários.

A maioria dos métodos desenvolvidos para a Análise de Sentimentos visa a análise de dados textuais. O tipo de texto mais analisado são manifestações espontâneas em blogs, foruns e chats. Textos jornalísticos escritos geralmente não são objeto desse tipo de análise. Eles apresentam desafios adicionais, pois neles pode não haver ocorrências de itens claramente associados a uma emoção em particular, como em: " 'Brasil poderá ter uma presidente mulher', diz Dilma: Declaração foi dada após encontro com Michelle Bachelet em SP."

Uma dificuldade adicional diz respeito à identificação de emoções predominantes, quando há palavras associadas a emoções contraditórias. Em "Andorinhas mudam rotina em cidade paraense: Elas chegam a Parauapebas e dão espetáculo no céu. Entretanto, sujeira deixada pelas aves incomoda moradores.", o texto como um todo pode gerar no leitor humano a emoção *surpresa*. Todavia, para uma implementação automática, a máquina deveria ser instruída sobre quais indicadores linguísticos seriam prototipicos da emoção *surpresa*. De fato, cada oração poderia ser atribuída a uma emoção distinta: Andorinhas mudam rotina em cidade paraense (*surpresa*); Elas chegam a Parauapebas e dão espetáculo no céu (*alegria)*; Entretanto, sujeira deixada pelas aves incomoda moradores (*repugnância*).

3. A Teoria Sistêmico-Funcional como arcabouço teórico para a Análise de Sentimentos

A Teoria Sistêmico-Funcional [Halliday e Matthiessen. 2014] considera a linguagem como um conjunto de sistemas utilizados para produzir significado e possibilitar a interação e representação da experiência humana. A organização da linguagem é estratificada: significados no estrato da semântica são realizados no estrato da gramática que por sua vez são realizados no estrato da fonologia.

No estrato da gramática, os sistemas que organizam escolhas estão organizados num contínuo, pelo qual escolhas progressivas em sistemas gramaticais concluem em um item lexical. Por exemplo, dentre os tipos de Processos que realizam orações de fala estão os Processos verbais, e dentro deles, os Processos verbais que constroem significados de semiose, dentro dos quais estão os Processos verbais que relatam eventos de forma neutra, sendo um dos verbos prototípicos para essa subespecificação o verbo "dizer".

A unidade de análise é a oração, na qual confluem os sistemas de Transitividade, Modo e Tema, relativos às três Metafunções fundamentais da linguagem: Ideacional, responsável pela representação da experiência humana; Interpessoal, pela troca de relações sociais entre falante/escritor e ouvinte/leitor e Textual, pela organização da mensagem, compreendendo aspectos relativos à coesão do texto. O Quadro 1 ilustra a análise de uma oração de acordo com as funções em cada Metafunção.

QUADRO 1 – Análise Sistêmica e Estrutural de uma oração

ESTRUTURA / METAFUNÇÃO SISTEMAS	*Andorinhas*	*mudam*	*rotina*	*em cidade paraense*
TEXTUAL	TEMA	REMA		
INTERPESSOAL MODO INDICATIVO SUJEITO: PLURAL: RECUPERADO	SUJEITO	PREDICADOR	COMPLEMENTO	ADJUNTO
	MODO	RESÍDUO		
IDEACIONAL EXPERIENCIAL MATERIAL: ORAÇÃO TRANSITIVA EFETIVA	PARTICIPANTE: ATOR	PROCESSO MATERIAL	PARTICIPANTE: META	CIRCUNSTÂNCIA DE LOCALIZAÇÃO: ESPACIAL
IDEACIONAL LÓGICA ORAÇÃO SIMPLES, FINITA				

4. Anotação manual de um *Corpus* de notícias

Examinamos dados obtidos de um *corpus* composto de 2.000 linhas finas retiradas de notícias, extraídas automaticamente da Internet através da ferramenta FeedReader[1] e originalmente escritas em português brasileiro. O *corpus* foi compilado por pesquisadores da Pontifícia Universidade Católica do Paraná (PUCPR), que trabalham com aprendizagem de máquina aplicada para a rotulação de sentimentos [Dosciatti et al. 2015]. As linhas finas do *corpus* apresentam, em média, 23 *tokens* cada uma, abordam temas internacionais, políticos, policiais e economicos e foram rotuladas de acordo com as emoções de [Ekman 1970] por anotadores voluntários, todos profissionais com experiência em linguística da PUCPR e da Universidade Tecnológica Federal do Paraná (UTFPR).

Para o presente estudo, uma amostra de aproximadamente 10% dos textos desse corpus que apresentaram concordância total entre os anotadores foi segmentada em 371 orações e estas anotadas manualmente de acordo com funções gramaticais. A anotação foi feita em planilhas eletrônicas, processadas no ambiente R [R Core Team 2017] para extração de frequências e identificação de coseleções de duas ou mais categorias. Os padrões observados foram identificados como candidatos passíveis de informar algoritmos a serem usados na análise de outras amostras do corpus.

5. Resultados Preliminares

[1] http://feedreader.com/

Dentre os resultados obtidos, destacamos os Processos mais frequentes para os textos rotulados com a emoção *raiva* e *alegria*, dispostos nas Tabelas a seguir:

TABELA 1 – Frequência de ocorrência de Tipo de Processo em textos rotulados com emoção *raiva*

Emoção predominante	Tipo de Processo	Exemplo	Frequência	
Raiva	Material	"Garota de 14 anos **engravidou** do próprio pai em Itariri, no interior de SP."	Absoluta 32	Relativa 60,4%
	Verbal	Mãe **confessa** [ter matado recém-nascido à tesouradas.]".	9	17%
	Relacional Atributivo	"Filhos **são** suspeitos [de abandonar pai idoso]"	7	13,2%
	Mental	"Polícia **crê** [que além do pai e do padrasto, outros estupraram a menina]"	4	7,5%
	Relacional identificativo	"Stuart Hall, de 83 anos, **é** um "predador oportunista.	1	1,9%
Total			53	100%

TABELA 2 – Frequência de ocorrência de Tipo de Processo em textos rotulados com emoção *alegria*

Emoção predominante	Tipo de Processo	Exemplo	Frequência	
Alegria	Material	"No local, 200 barracas devem ser **montadas**/para **vender** comidas e bebidas "	Absoluta 43	Relativa 81,1%
	Relacional Atributivo	"Segundo o ministro do Planejamento, governo **está** otimista."	5	9,4%
	Verbal	"[Brasil já saiu da crise] **diz** Paulo Bernardo".	3	5,7%
	Relacional identificativo	"André Cintra, paulistano de 34 anos, **tornou-se**, nesta sexta-feira, o segundo brasileiro..."	2	3,8%
Total			53	100%

Os Processos mais frequentes para a emoção *raiva* foram: Material (60,4%), Verbal (17%), Relacional Atributivo (13,25%). Já os Processos com maior número de ocorrências para a emoção *alegria* foram: Material (43%), Relacional Atributivo (9,4%) e Verbal (5,7%). Esses resultados sugerem semelhanças em relação ao tipo de Processo mais frequente para as emoções *raiva* e *alegria.* Contudo, a emoção *alegria* tem maior probabilidade de apresentar orações com Processo Material, já que a porcentagem desse tipo de Processo nos textos de tal emoção é maior. Uma diferença entre as emoções *raiva* e *alegria* é a ocorrência do Processo Mental (7,5%), presente apenas na emoção *raiva*.

6. Conclusão

Os achados iniciais apontam para características dos textos rotulados com uma dada emoção, as quais podem ser mais investigadas para efeitos da Análise de Sentimentos.

7. Referências

Dosciatti, M. M., Ferreira, L. P. C., Paraiso, E. C. (2015) Anotando um Corpus de Notícias para a Análide de Sentimento: Um relato de experiência. In Proceedings of Symposium in Information and Human Language Technology (STIL), 121-130.

Ekman, P. (1970) Universal Facial Expressions of Emotion. California Mental Health Research Digest.

Halliday, M. A. K., Matthiessen, C. M. I. M. (2014) An Introduction to Functional Grammar. Routledge, London.

Taboada, M. et al. (2011) Lexicon-based Methods for Sentiment Analysis. In: Journal Computational Linguistics**,** 267-307.