Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros

1. Introdução

O presente trabalho descreve a construção de um corpus de **Re**senhas de **Li**vros manualmente anotado quanto à expressão de opinião, o ReLi. Especificamente, relatamos as opções linguísticas tomadas durante o processo de anotação, apresentamos os resultados de testes de concordância, problematizando o processo de concordância no âmbito de uma tarefa complexa como a identificação de opinião, e exploramos brevemente o corpus criado.

A criação do ReLi foi duplamente motivada: do lados dos estudos da linguagem, a escolha com um corpus de textos produzidos na internet, não jornalísticos, se justifica pelo crescente interesse na descrição da linguagem utilizada na rede, além de contribuir para estudos com base em corpus que tratem da linguagem avaliativa. Do lado do processamento automático de linguagem natural (PLN), a criação de um corpus anotado quanto à expressão de opinião interessa principalmente à Mineração de Opinião (Opinion Mining), área vinculada à Análise de Sentimento (Sentiment Analysis), que lida com a identificação de opiniões, avaliações e atitudes das pessoas com relação a entidades como pessoas, produtos e organizações, expressas em texto.

Na área de mineração de opinião, a maioria dos trabalhos parte de léxicos de emoções/sentimentos, que podem ser gerais, como SentiWordNet, ou específicos para determinados domínios ou tarefas (Riloff & Wiebe 2003; Poirier et al. 2011 e, para o português, Pasqualotti & Vieira 2008 e Silva et al.2012), sendo poucos os trabalhos que consideram corpora anotados, devido principalmente ao seu alto custo de elaboração. O MPQA Opinion Corpus (Wiebe et al. 2005) é o corpus que mais se assemelha ao apresentado neste trabalho. Trata-se de um corpus de 10 mil frases de artigos de jornal, ricamente anotado no nível da palavra e do sintagma, com informação relativa a estados mentais e emocionais, visando à distinção entre informação subjetiva de informação factual. Especificamente com relação à identificação de opinião em avaliações de produtos, destacamos os trabalhos de Zagibalov et al. 2010, que criaram corpora comparáveis em inglês e russo a partir de resenhas de livros publicadas na Amazon.com, e de Poirier et al. 2011, que relatam a extração da opinião em resenhas de filmes publicadas em um site colaborativo.

Liu et al.2005 classificam as avaliações de produtos em 3 tipos: nas do tipo (1) é pedido ao avaliador que escreva os prós e contras separadamente; no tipo (2) além dos prós e contras, é pedido ao avaliador que escreva uma avaliação detalhada; o tipo (3) tem formato livre: o avaliador pode escrever livremente, e não há separação formal entre pós e contras.

As resenhas que compõem o ReLi se enquadram no tipo (3), e foram extraídas do sítio Skoob.com, uma rede social de livros e leitores, na qual os leitores/colaboradores participam de forma ativa e entusiasmada comentando os livros que leram.

O corpus

O corpus é composto por 2056 resenhas de 13 livros (7 autores), totalizando cerca de 300 mil palavras e 15mil frases. Cada livro contém cerca de 200 resenhas e, quando esse número não pôde ser atingido, completamos com outras obras do mesmo autor até chegarmos a um número próximo a 200. O corpus foi anotado com informação de PoS e de chunks pelo F-EXT (Fernandes, et al. 2009).

As opiniões anotadas referem-se exclusivamente às opiniões sobre o livro resenhado. Com relação à linguagem temos um corpus com uma linguagem bastante informal, com gírias, abreviações, neologismos e emoticons.

A opinião nas resenhas foi anotada no nível da frase e do sintagma, e a anotação ocorre em 5 etapas:

A- Identificação, na resenha, das frases que veiculam opinião sobre o objeto;
B- Anotação da polaridade da frase com (+) ou (-). Como a mesma frase pode conter uma opinião positiva e uma negativa, anotamos a polaridade da frase como um todo.
C- Identificação do alvo da opinião (o livro ou partes dele, como enredo, personagem, etc). Quando a frase não explicita o objeto da opinião (Adorei!), consideramos, por padrão, que a opinião será sempre sobre o livro.
D- Identificação do segmento que expressa a opinião.
E- Associação da polaridade ao segmento que expressa a opinião.

Com relação ao item D, notamos que nem sempre é fácil (ou possível) identificar o trecho exato da opinião. Nesses casos, a preferência é por anotar todo o trecho (ou frase):

(1) Mas em um contexto todo, o livro *conseguiu suprir minhas expectativas* [+]. (+)
(2) *impossível abandonar o livro pela metade* , [+]. (+)
(3) *a tradução deveria ser CRAPúsluco* [-].(-)

Quanto à segmentação, consideramos sempre o menor núcleo possível e uma mesma frase pode conter diversas opiniões, coordenadas, bem como opiniões contrastivas.

Talvez a principal dificuldade tenha estado/ esteja em distinguir informação subjetiva de informação factual. Embora não previssemos lidar com esse tipo de problema em um corpus de resenhas de livros, não raro era muito difícil decidir se estávamos diante de uma descrição do personagem ou de uma opinião sobre ele.

Na anotação, assumimos uma posição conservadora e, quando não conseguimos saber se o trecho relata uma característica do personagem ou uma opinião sobre ele, optamos por não anotar.

Outra dificuldade esteve no estabelecimento dos limites do que seria considerado parte do objeto e, portanto, alvo da opinião. Se, por um lado, é previsível a ocorrência de partes generalizáveis no domínio dos livros, como *capítulos, personagens, linguagem*, por outro, também foi frequente a menção a partes específicas de cada livro ("Não contente em derrapar na parte romântica, Stephenie Meyer também resolve deturpar o clássico e marcante vampiro..."). Embora nossa primeira intenção tenha sido desconsiderar tais objetos justamente por sua especificidade com relação ao livro, logo percebemos que estávamos impondo um limite talvez artificial à tarefa de identificação de opinião em livros isto é, estávamos tomando uma decisão arbitrária tendo em vista somente a facilidade na anotação. Assim, optamos por considerar toda e qualquer opinião sobre o livro ou partes dele.

O processo de anotação

O corpus foi anotado por 3 anotadores, alunos de graduação do curso de Letras. Todos passaram por um processo de treino até que estivessem familiarizados com a tarefa, com as instruções e com a ferramenta de anotação, sendo enfatizada a importância das interpretações serem sempre feitas no contexto particular da resenha. Durante o processo de treino e também durante todo o processo de anotação, os anotadores foram encorajados a perguntar e discutir suas opções e, à medida que surgiam casos não previstos, as soluções eram discutidas e incorporadas ao manual. Depois das primeiras semanas, todo o material anotado foi revisto por um dos autores deste artigo, e eventuais dúvidas foram resolvidas.

Estudo da concordância entre anotadores

Depois de cerca de 400 resenhas anotadas, realizamos um estudo da concordância entre os anotadores. A avaliação da concordância considerou a anotação dos três anotadores em um mesmo conjunto de 200 resenhas. Embora as instruções de anotação dessem alguma informação quanto à segmentação, era de se esperar alguma variação quanto à extensão das unidades selecionadas. Um dos desafios, portanto, esteve em definir a concordância nos casos em que os anotadores identificaram a mesma opinião, mas divergiram quanto aos limites da unidade.

De fato, trata-se de uma tarefa de avaliação mais complexa que a de julgamento de atribuição de polaridade a sentenças, e nos apoiamos no processo de avaliação de Wiebe et al (2005) em dois pontos fundamentais: (i) quanto à segmentação, consideramos expressões como parte final e final expressões equivalentes, e (ii) utilizamos a métrica *agr*, que tem como objetivo avaliar se os anotadores identificaram o mesmo conjunto de objetos e de opiniões.

Os seguintes pontos foram considerados na avaliação da concordância: concordância quanto às frases selecionadas; concordância quanto à polaridade das frases selecionadas; concordância quanto aos objetos selecionados; concordância quanto às opiniões selecionadas; concordância quanto à polaridade das opiniões selecionadas.

Breve exploração do corpus

A seguir relatamos alguns dados do corpus, considerando apenas um subconjunto revisto. De um total de 6.000 frases revistas (850 resenhas), 26% contêm opinião, 76% delas positivas o que não surpreende, considerando as características do corpus. Em 32,5% das frases que expressam opinião encontramos opiniões contrastivas, isto é, aspectos positivos e negativos do livro na mesma frase. Em 18% das frases a opinião não pode ser pontualmente localizada, estando na frase inteira. A tabela 1 apresenta a distribuição das expressões de opinião conforme o número de palavras:

Tamanho do n-grama	frequência
1-3	69%
4- 6	15%
7+	15%

Outro dado interessante está em palavras ou expressões cuja polaridade varia conforme o contexto, ainda que em um mesmo domínio, ou mesmo palavras que, no contexto, assumem uma polaridade diferente da usual. Assim, "diferente" poder ser positivo ou negativo. O mesmo para "linguagem fácil"", que para alguns leitores é característica positiva e, para outros, negativa. Expressões com diminutivo podem indicar tanto um julgamento positivo quanto negativo e a frase "E no final você ficará com ódio, pode ter certeza!", no contexto, está associada a um julgamento positivo sobre o livro.

Por fim, a presença de neologismos ("A Bella é muito *tonga*"), expressões típicas da internet ("rá!"; "Nah neh noh") e emoticons, além de frases mal formuladas e com erros, deixa clara a necessidade de anotadores robustos, capazes de lidar com fatos novos.

Considerações Finais

Apresentamos o ReLi, um corpus manualmente anotado quanto à expressão de opinião sobre livros.

Motivado principalmente pela tarefa de identificação de opinião em textos, o fato de o corpus ser constituído por resenhas de um site também contribui para o estudo dos gêneros na Web, somando-se aos ainda escassos corpora do português que contêm textos não jornalísticos escritos na internet.

A análise da expressão da opinião aponta para a importância do contexto quando do julgamento da polaridade, sugerindo que uma abordagem baseada exclusivamente em léxicos pode deixar de considerar aspectos relevantes da linguagem avaliativa.

Por fim, salientamos que a granularidade da anotação pode - e deve - ser abstraída para os casos em que não se precisa de tanta informação semântica, como o treino de sistemas capazes de detectar a polaridade de frases, por exemplo, dentre outras aplicações de PLN, como extração de informação, perguntas e respostas.

Referências:

Fernandes, E., Milidiú, R., Santos, C., 2009, "Portuguese Language Processing Service", In: 18th International World Wide Web Conference. Madrid, 2009

Liu, Bing, Hu, Minqing e Cheng ,Junsheng. (2005) "Opinion Observer: Analyzing and Comparing Opinions on the Web". Proceedings of the 14th international World Wide Web conference (WWW-2005). Chiba, Japan.

Riloff, Ellen e Wiebe, Janyce. (2003) "Learning Extraction Patterns for Subjective Expressions", Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03) .

Pasqualotti, Paulo; Vieira, Renata. (2008) "WordnetAffectBR: uma base lexical de emoções para a língua portuguesa". RENOTE. Revista Novas Tecnologias na Educação, v. 6, p. 1-10, 2008.

Poirier, D., C. Bothorel, E. Guimier, M. Boullé. (2011) "Automating Opinion Analysis in Film Reviews: the Case of Statistic versus Linguistic Approach". In Ahmad, Khurshid (Ed.), *Affective Computing and Sentiment Analysis: Metaphor, Ontology, Affect and Terminology*. Springer Edition, 2011.

Silva, Mário J., Carvalho, Paula e Sarmento, Luís. (2012) "Building a Sentiment Lexicon for Social Judgement Mining", In Lecture Notes in Computer Science (LNCS) / Lecture Notes in Artificial Intelligence (LNAI), International Conference on Computational Processing of Portuguese (PROPOR), 17-20 April, 2012, Coimbra.

Wiebe, Janyce, Wilson, Theresa e Cardie, Claire (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.

Zagibalov, Taras, Belyatskaya, Katerina e Carroll, John. (2010) "Comparable English-Russian book review corpora for sentiment analysis". In Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.