IDENTIFICANDO ALVOS DE OPINIÃO EM COMENTÁRIOS DE NOTÍCIAS USANDO INFORMAÇÃO SINTÁTICA

Leonardo Gabiato Catharin (PIBIC/CNPq/FA/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, análise de sentimentos, extração de alvos.

Resumo

É fato que hoje as pessoas têm demonstrado cada vez mais suas opiniões na Web. Isso a torna um grande repositório de comentários sobre diversos assuntos e traz consigo a necessidade de se extrair informações relevantes para o usuário em meio ao grande volume de informações. A Análise de Sentimentos (AS) surge então para suprir essa necessidade. Neste trabalho abordamos o problema da extração de alvos de opinião, uma das tarefas de AS. Foi adotada uma abordagem baseada em aprendizado de máquina supervisionado empregando atributos léxicos, morfológicos e sintáticos. Em especial, se buscou avaliar o impacto do uso de informação sintática na classificação. Os resultados obtidos mostraram uma melhora discreta de desempenho em relação a protótipos implementados em trabalhos anteriores.

Introdução

Nos dias de hoje é fato que as pessoas têm cada vez mais demonstrado suas opiniões em redes sociais, blogs, reviews de produtos etc. Isso tem feito com que a Web se torne um grande repositório de informações sobre diversificados assuntos, produtos e pessoas. Com esse crescimento do volume de opiniões e informações, surge também a necessidade de extrair tais opiniões e informações que sejam relevantes aos usuários. A Análise de Sentimentos (AS) surge como um campo de estudo que busca por soluções para a automatização dessa tarefa (LIU, 2012).

Este trabalho tratou de uma das tarefas de AS, a extração de alvos de opinião. Essa tarefa consiste em detectar a entidade a respeito da qual a opinião dada se refere. Foram abordados dois domínios diferentes de opinião, ambos em língua portuguesa: comentários a respeito de notícias políticas e resenhas de livros. De forma similar a Silva (2016), a extração de alvos de opinião nesses domínios foi implementada como uma tarefa de rotulação sequencial, em que as sentenças compõem as sequências e as palavras as unidades de classificação. As palavras foram caracterizadas por

meio de atributos léxicos e morfológicos, conforme proposto por Silva (2016), e também por atributos que capturam informação sintática. Com o desenvolvimento deste trabalho, se buscou avaliar o impacto do uso da informação sintática na classificação dos alvos, bem como o desempenho da abordagem proposta quando aplicada ao domínio das resenhas.

Materiais e métodos

Os experimentos foram conduzidos utilizando dois *corpora* em português anotados a respeito da polaridade e alvo das opiniões: o SentiCorpus-PT (CARVALHO, 2011) e *corpus* ReLi (FREITAS, 2012). O SentiCorpus-PT é composto por comentários de notícias sobre as eleições portuguesas, totalizando 2.515 sentenças. O ReLi é composto por 1.600 resenhas de 14 livros diferentes e totaliza 12.514 sentenças.
O protótipo construído por Silva (2016) suportava apenas o processamento do SentiCorpus-PT. Assim, foi necessária a implementação de um novo protótipo para o processamento do *corpus* ReLi, além de adequações à extração de atributos do SentiCorpus-PT.
A delimitação dos alvos nas sentenças foi feita usando duas notações diferentes: a BIO (B de *Begin,* I de *inside*, O de *Outside*) e a BILOU (BIO, L de *Last*, U de *Unit*).
Em termos dos vetores de atributos, foi adicionada ao protótipo de Silva (2016) a extração de dois atributos sintáticos: um indicando o tipo de sintagma em que a palavra participa e outro indicando a função sintática exercida pelo sintagma. Para isso foi usada a API CoGrOO. Também foi adicionado um atributo representando a própria palavra. Assim, foi criado um novo extrator que calcula sete atributos que codificam informações em nível léxico, morfológico (POS *tagging*) e sintático. Além disso, foram adicionados aos vetores os atributos das *n* sentenças anteriores e posteriores, o que foi chamado de janela de tamanho *n*.
O treinamento e teste dos classificadores foram feitos usando a implementação de Silva (2016). Esta, por sua vez, usa a ferramenta CRF Suite para treinar um classificador *Conditional Random Fields* (CRF). Os testes dos dois corpora foram feitos com validação cruzada: de 12 partições para o SentiCorpus-PT e de 10 partições para o ReLi.

Resultados e Discussão

Inicialmente foram realizados experimentos com o SentiCorpus-PT utilizando os sete atributos, uma janela de tamanho três e a notação BIO. Como os resultados obtidos não foram satisfatórios, novos experimentos foram realizados visando selecionar o melhor conjunto de atributos.
Para o SentiCorpus-PT e a notação BIO, os melhores resultados foram obtidos com um subconjunto de seis atributos e uma janela de tamanho um. São eles: *chunk* (tipo do sintagma em que a palavra participa); *shallow* (função sintática do sintagma); a palavra completa; a raiz da palavra; a etiqueta POS (*Part of Speech*) da palavra e informação se a palavra inicia

com letra maiúscula. Os resultados desse experimento são mostrados na Tabela 1, correspondendo a uma melhora de cerca de 4% em relação a Silva (2016).

Cabe destacar que, devido ao uso da notação BIO para a marcação dos alvos, a avaliação dos resultados por meio das medidas tradicionais de precisão e revocação dificultariam a interpretação, uma vez que reportariam os acertos individuais por etiqueta (I, O e B). Assim, as porcentagens de acerto foram computadas em termos do alvo em vez de por etiqueta.

Tabela 1 – Resultados para o SentiCorpus-PT com a notação BIO

Métodos	% Acerto
Acerto por alvo (qualquer parte do alvo)	41,89%
Acerto por alvo (maior que 50% do alvo)	41,48%
Acerto por alvo (100% do alvo)	41,41%

Segundo Amaral e Vieira (2014), a notação BILOU produziu resultados melhores para a tarefa de reconhecimento de entidades nomeadas. Por ser uma tarefa semelhante à deste trabalho, a notação BILOU também foi avaliada. Os resultados para o SentiCorpus-PT com a notação BILOU, usando o mesmo conjunto de seis atributos e janela de tamanho 1, são mostrados na Tabela 2.

Tabela 2 – Resultados para o SentiCorpus-PT com a notação BILOU

Métodos	% Acerto
Acerto por alvo (qualquer parte do alvo)	42,41
Acerto por alvo (maior que 50% do alvo)	42,11
Acerto por alvo (100% do alvo)	42,07

Dados os resultados do SentiCorpus-PT, o *corpus* ReLi foi avaliado apenas com a notação BILOU. Os resultados para o ReLi, usando todos os atributos e uma janela de tamanho três, são mostrados na Tabela 3.

Tabela 3 – Resultados para o *corpus* ReLi com a notação BILOU

Métodos	% Acerto
Acerto por alvo (qualquer parte do alvo)	35,61
Acerto por alvo (maior que 50% do alvo)	35,50
Acerto por alvo (100% do alvo)	35,43

Como pode ser observado, os resultados para o ReLi ficaram abaixo dos obtidos para o SentiCorpus-PT. Entre os fatores que podem ter contribuído para isso, cabe destacar a diferença na distribuição de alvos dos dois corpora. Enquanto a proporção alvos/sentenças para o SentiCorpus-PT é de 1,8, para o ReLi é de 0,22 (aproximadamente 82% das sentenças do ReLi não possuem alvos explícitos).

Conclusões

O objetivo principal deste estudo foi propor e avaliar atributos que capturassem informação sintática das palavras em sentenças de comentários em língua portuguesa. Por meio deste estudo, foi possível avaliar o impacto de tais atributos na classificação automática de alvos de opinião baseada em um classificador CRF. Também foi possível avaliar o impacto do uso das notações BIO e BILOU para a delimitação dos alvos.
A partir dos experimentos realizados, pudemos concluir que o uso da notação BILOU, juntamente com os novos atributos propostos e uma janela de tamanho 1 apresentaram uma melhora nos resultados, embora discreta (4%). A adição dos atributos sintáticos sozinha não produziu resultados melhores.
Com relação ao desempenho nos corpora utilizados, os resultados obtidos para o SentiCorpus-PT foram melhores do que os obtidos para o *corpus* ReLi, o que pode ser atribuído à diferença na distribuição dos alvos dos dois corpora. A proporção alvos/sentenças para o ReLi é mais baixa, o que torna a classificação no *corpus* ReLi uma tarefa mais difícil em comparação ao SentiCorpus-PT.

Agradecimentos

Ao programa PIBIC/CNPq/FA/UEM pelo apoio financeiro.

Referências

AMARAL, D. O. F. do; VIEIRA, R. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de Conditional Random Fields. **Linguamática** — Issn: 1647–0818, v. 6, n. 1, p.41-49, 1 jul. 2014.

CARVALHO, P.; TEIXEIRA, J.; SARMENTO, L.; SILVA, M. J. SentiCorpus-PT - Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. In: ACL (SHORT PAPERS). [S.l.: s.n.], 2011. p. 564 – 568.

FREITAS, C.; MOTTA, E.; MILIDIÚ, R. L.; CÉSAR, J. Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. **Encontro de Linguística de Corpus**, v. 11, p. 1-13, 2012.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan Claypool Publishers, 2012.

SILVA, R. H. **Extração de alvo de opinião usando aprendizado de máquina supervisionado.** Monografia de Conclusão de Curso. Departamento de Informática, Universidade Estadual de Maringá, 2016, 36p.