

Extração de Alvos em Comentários de Notícias em Português baseada na Teoria da Centralização

Frank Willian Cardoso de Oliveira¹, Valéria Delisandra Feltrim¹

¹Departamento de Informática – Universidade Estadual de Maringá (UEM)
CEP 87020-900 – Maringá – PR – Brazil

{frankwco, valeria.feltrim}@gmail.com

***Abstract.** This paper presents a prototype for target extraction in news comments in Portuguese based on Centering Theory. The prototype was evaluated and the results showed that Centering helps target extraction.*

***Resumo.** Este trabalho apresenta um protótipo para a extração de alvos em comentários de notícias da língua portuguesa baseado na teoria da centralização. O protótipo foi avaliado e os resultados mostraram que a teoria auxilia na extração de alvos.*

1. Introdução

Para realizar a análise de sentimentos de forma mais refinada é necessário conhecer sobre quais entidades ou aspectos o escritor expressou sua opinião. Assim, uma das etapas dessa análise com uma granularidade mais fina busca extrair qual é o alvo da opinião [Liu 2012].

Grande parte dos trabalhos que buscam identificar alvos se concentram na extração de aspectos em *reviews* de produtos ou serviços, nos quais as entidades já são conhecidas. Poucos trabalhos focam a extração de alvos em outros tipos de texto, como os comentários de notícias. Uma proposta voltada para comentários de notícias escritos em chinês é a de [Ma and Wan 2010]. Já para a língua portuguesa, não foram encontrados na literatura trabalhos relacionados à extração de alvos para esse domínio.

Dessa forma, este artigo apresenta um protótipo para a extração de alvos em comentários de notícias escritos em português. O protótipo é uma adaptação da abordagem proposta por [Ma and Wan 2010], que faz uso da teoria da centralização [Grosz et al. 1995] para identificar para cada sentença do comentário, um alvo.

2. Trabalhos Relacionados

Vários trabalhos da literatura buscaram extrair aspectos sobre entidades conhecidas a partir de *reviews* de produtos e serviços. [Hu and Liu 2004] utilizaram um algoritmo que busca por substantivos e sintagmas nominais frequentes para extrair aspectos a partir de *reviews* de produtos. Exemplos de trabalhos com abordagens similares são os de [Popescu and Etzioni 2005], [Siqueira 2013] e [Silva 2010].

Já no domínio das notícias, [Kim and Hovy 2006] propuseram um método para a extração do titular, do alvo e da polaridade da opinião para cada sentença proveniente de notícias *online*. Para isso, o método explora informações semânticas provenientes de *Semantic Role Labeling* e da *FrameNet*.

Visto que nosso objetivo é extrair alvos a partir de comentários de notícias, o trabalho que mais se relaciona ao nosso é o de [Ma and Wan 2010], que propuseram uma abordagem para a extração de alvos em comentários de notícias para a língua chinesa baseada na teoria da centralização. A partir da análise manual dos comentários, os autores concluíram que informações relativas aos centros de atenção poderiam ser úteis na extração de alvos. Uma vez que um centro representa o foco de atenção de um enunciado, isso seria um indicativo de que o centro de atenção é o alvo. A abordagem proposta pelos autores contempla tanto alvos implícitos (alvos não mencionados na sentença opinativa), quanto alvos explícitos (alvos mencionados na sentença opinativa). Para a identificação de alvos implícitos são utilizadas informações extraídas da notícia comentada e informações contextuais extraídas em sentenças adjacentes nos comentários. A avaliação da abordagem foi feita com 1.597 sentenças extraídas dos comentários de nove notícias relacionadas a economia, esportes e tecnologia. Para cada sentença foi extraído um único alvo e a taxa de acerto geral (alvos explícitos e implícitos) foi de 43,2%.

3. Teoria da Centralização

Assim como [Ma and Wan 2010], nossa proposta para a extração de alvos usa informações provenientes da teoria da centralização (*Centering*). Proposta por [Grosz et al. 1995], a teoria foi desenvolvida para avaliar a coerência do discurso por meio da análise das transições entre os centros de atenção de cada enunciado.

Na teoria da centralização, cada enunciado U_i possui um conjunto ordenado de centros associados chamado de *Forward-Looking Centers* $Cf(U_i)$. Esse conjunto contém todos os potenciais centros de atenção do enunciado atual e que também representam os potenciais centros dos próximos enunciados, assumindo um texto coerente. A ordenação do $Cf(U_i)$ é realizada de acordo com a função sintática dos elementos, sendo sujeito $>$ objeto $>$ outros a ordem de preferência mais comum. O primeiro elemento do conjunto $Cf(U_i)$ é o mais saliente e é denominado *Preferred Center*, sendo representado por $Cp(U_i)$. Outro elemento do Cf é o *Backward-Looking Center*, representado por $Cb(U_i)$. Cada enunciado possui um Cb , que se conecta com um elemento do $Cf(U_{i-1})$, desde que o enunciado não seja o primeiro do discurso. Em um discurso coerente, o $Cp(U_i)$ tem a maior probabilidade de ser o $Cb(U_{i+1})$.

4. Descrição do Protótipo para Extração de Alvos

O objetivo deste protótipo é a extração de alvos explícitos em comentários de notícias em português. Considerando a definição de alvo proposta por [Liu 2012], nosso foco são as entidades dos discurso, dado que o *corpus* de comentários utilizado no desenvolvimento e avaliação do protótipo tem como alvos entidades humanas, em particular, políticos.

O protótipo recebe como entrada uma base de comentários. Em uma primeira etapa é feito o pré-processamento, que inclui substituição de abreviações e gírias, correção ortográfica e análise sintática e morfológica. As bases de abreviações e gírias foram criadas manualmente a partir da observação do SentiCorpus-PT [Carvalho et al. 2011] e listas disponibilizadas na internet. O corretor ortográfico foi construído a partir da base léxica do LibreOffice¹. Para a análise sintática e morfológica foi utilizada a API da ferramenta Cogroo².

¹<http://pt-br.libreoffice.org/>

²http://ccsl.ime.usp.br/redmine/projects/cogroo/wiki/API_CoGrOO_4x

Tabela 1. Pseudocódigo baseado na Teoria da Centralização

Entrada: Um comentário com M sentenças $S=\{s_i\}$, sendo que cada sentença possui um conjunto de alvos candidatos $Cf(s_i)=\{c_i\}$.
Saída: Um conjunto de alvos $\{t_i\}$, no qual cada t_i é um alvo da sentença s_i .
1. Para Cada s_i em S 2. Se $i = 1$ (s_i é a primeira sentença) 3. Escolher o elemento de melhor <i>ranking</i> no conjunto $Cf(s_i)$ ($Cp(s_i)$) como t_i 4. Se Não 5. Para Cada c_i em $Cf(s_i)$ 6. Se c_i está relacionado com um elemento c'_i em $Cf(s_{i-1})$ 7. Adicionar c'_i no conjunto $Cb(s_i)$ 8. Se $Cb(s_i)$ não estiver vazio 9. Escolher o elemento de melhor <i>ranking</i> do conjunto $Cb(s_i)$ como t_i 10. Se Não 11. Escolher o elemento de melhor <i>ranking</i> do conjunto $Cf(s_i)$ como t_i

Após o pré-processamento é feita a extração dos alvos candidatos. São considerados candidatos todos os substantivos, nomes próprios e pronomes encontrados. Assim, para cada sentença é gerada uma lista ordenada com os possíveis candidatos. Tendo por base a teoria da centralização, a ordenação dos candidatos é feita de acordo com a sua função sintática. Neste trabalho usamos a seguinte ordem de preferência: sujeito > objeto direto > objeto indireto > objeto preposicional > outros.

A próxima etapa é a escolha do melhor candidato a alvo da sentença. Assim como em [Ma and Wan 2010], o algoritmo que seleciona o melhor candidato usa informações provenientes do Cf , Cp e Cb . Ao final do processamento, apenas um candidato é escolhido como alvo para cada sentença do comentário. O pseudocódigo do algoritmo de seleção do melhor candidato a alvo utilizado no protótipo é apresentado na Tabela 1.

5. Avaliação do Protótipo

A avaliação do protótipo foi feita com um subconjunto de comentários do SentiCorpus-PT [Carvalho et al. 2011]. O SentiCorpus-PT é composto por comentários relacionados a notícias políticas manualmente anotados com informações relativas à polaridade e aos alvos da opinião. A versão do SentiCorpus-PT utilizada neste trabalho é composta por 1.082 comentários, totalizando 2.726 sentenças opinativas.

Para o teste do protótipo foram extraídos aleatoriamente do SentiCorpus-PT 100 comentários, totalizando 255 sentenças. A quantidade reduzida de comentários usados na avaliação se deve ao fato da teoria da centralização pressupor a resolução de correferência, a qual foi realizada manualmente para os comentários extraídos.

Das 255 sentenças extraídas, 99 continham mais de um alvo. Assim como em [Ma and Wan 2010], neste trabalho apenas um alvo foi extraído para cada sentença. Dessa forma, para as sentenças com mais de um alvo, a extração foi considerada correta se o alvo extraído estava entre os alvos anotados para sentença.

Para avaliar o efeito da teoria da centralização na extração, duas *baselines* foram

construídas. A *Baseline 1* considera como alvo o sujeito da sentença. No caso de períodos compostos com mais de um candidato, a *baseline* escolhe o alvo de acordo com a seguinte ordem de preferência: nomes próprios > substantivos > pronomes. Caso os candidatos tenham a mesma classificação sintática e morfológica, é escolhido como alvo o candidato que aparece primeiro na sentença. A *Baseline 2* considera como alvo os nomes próprios, independente da classificação sintática. Caso exista mais de um nome próprio na sentença, é escolhido o primeiro encontrado.

Os resultados obtidos para as duas *baselines* e para o protótipo em termos da taxa de acerto são apresentados na Tabela 2.

Tabela 2. Resultados da Extração

	Precisão
<i>Baseline 1</i>	46,27%
<i>Baseline 2</i>	48,63%
Teoria da centralização e sem resolução de correferência	55,29%
Teoria da centralização e com resolução de correferência	63,92%

Comparando-se as *baselines*, a *Baseline 2* foi 2,36% melhor que a *Baseline 1*. Acreditamos que isso se deva a característica do corpus, em que os alvos são entidades humanas, favorecendo assim a ocorrência de alvos que correspondem a nomes próprios. Já o protótipo superou as duas *baselines*, apresentando um desempenho 17,65% melhor em comparação a *Baseline 1* e 15,29% melhor em comparação a *Baseline 2*. Isso mostra a contribuição da teoria da centralização e o seu potencial na identificação dos alvos.

Para avaliar o impacto da resolução de correferência, o protótipo foi avaliado com o mesmo corpus de 100 comentários, porém sem a resolução manual de correferência. Como era esperado, o protótipo apresentou uma queda de 8,63% na taxa de acerto, mas ainda assim foi melhor que as *baselines*.

6. Conclusões e Trabalhos Futuros

Este trabalho apresentou um protótipo para a extração de alvos em comentários de notícias escritos em língua portuguesa. Para isso foi utilizada uma abordagem baseada na extração de sintagmas nominais e na teoria da centralização para escolher o melhor candidato a alvo de cada sentença. Na avaliação do protótipo foram utilizados 100 comentários retirados do SentiCorpus-PT. O resultado final, com a taxa de acerto de 63,92%, foi comparado a duas *baselines*, demonstrando a contribuição da teoria da centralização para a identificação de alvos.

A teoria da centralização pressupõe que seja realizada a resolução de correferência. Neste trabalho esse processo foi feito manualmente, o que limitou o tamanho do corpus utilizado na avaliação. Assim, como um trabalho futuro pretendemos automatizar essa etapa e verificar qual o impacto de se utilizar uma ferramenta de resolução automática de correferência. Além disso, pretendemos testar o protótipo em outros tipos de textos, como comentários extraídos de redes sociais. Outros trabalhos futuros incluem a construção de extratores baseados em aprendizado de máquina e no uso de padrões sintáticos e morfológicos [Liu et al. 2013], permitindo avaliar o desempenho das diferentes abordagens no contexto da extração de alvos em comentários de notícias.

Agradecimentos

A Capes pelo apoio financeiro e ao Prof. Dr. Sérgio Roberto Pereira da Silva (*in memoriam*) pela motivação e apoio para o início deste trabalho.

Referências

- Carvalho, P., Sarmiento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 564–568, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225.
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.
- Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*, volume 5. Morgan Claypool Publishers.
- Liu, K., Xu, L., and Zhao, J. (2013). Syntactic patterns versus word alignment: Extracting opinion targets from online reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1754–1763, Sofia, Bulgaria. Association for Computational Linguistics.
- Ma, T. and Wan, X. (2010). Opinion target extraction in chinese news comments. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 782–790, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Silva, N. G. R. d. (2010). WhatMatter: Extração e visualização de características em opiniões sobre serviços. Master's thesis, Universidade Federal de Pernambuco.
- Siqueira, H. B. A. (2013). PairClassif - Um Método para Classificação de Sentimentos Baseado em Pares. Master's thesis, Universidade Federal de Pernambuco.

