Reconhecimento de entidades nomeadas em textos em português do Brasil no domínio do *e-commerce*

Lucas Hochleitner da Silva[1,2], **Helena de Medeiros Caseli**[1]

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
13.565-905 – São Carlos – SP – Brasil

[2]Boolabs – http://boolabs.com.br/

hochlucassilva@gmail.com, helenacaseli@dc.ufscar.br

1. Introdução

O mercado varejista *online* no Brasil está em franca expansão tendo alcançado um faturamento de R$35,8 bilhões em 2014, um crescimento nominal de 24%, já que em 2013 o resultado foi de R$ 28,8 bilhões.[1] A maior parte da informação disponibilizada em páginas de *e-commerce* é apresentada na forma de texto em língua natural. Por serem apresentadas de modo não estruturado, faz-se necessário coletar, organizar e comparar essas informações por meio da aplicação de técnicas computacionais. O processamento mais inteligente dessas informações, possível com a aplicação de técnicas de aprendizado de máquina, pode garantir a precisão no reconhecimento e classificação de textos, gerando conhecimento útil para diversas aplicações.

Neste contexto, este artigo busca, com o Reconhecimento de Entidades Nomeadas, suprir a necessidade de um processamento mais inteligente e garantir o melhor aproveitamento do conteúdo de páginas de *e-commerce*. O Reconhecimento de Entidades Nomeadas (REN) é um tipo de extração de informação que visa identificar regiões do texto (menções) correspondentes a entidades e categorizá-las numa lista pré-determinada de tipos (entidades de interesse) [Ling e Weld 2012]. REN é considerada uma tarefa fundamental para a mineração de dados [Jing 2012].

Entidades Nomeadas são, segundo [CoNLL 2002], expressões que contêm nomes de pessoas, organizações, locais, tempos e quantidades. Neste trabalho, a definição de Entidade Nomeada foi conceitualmente expandida, abrangendo entidades além das mencionadas, que são de maior relevância quando se trata de um corpus de *e-commerce*. As entidades mais utilizadas (Pessoa, Organização e Local) não se apresentam como interessantes, visto que o corpus consiste de descrições de produtos, textos técnicos e análises, onde tais entidades não são encontradas. Assim, considerando o conteúdo do corpus, as nove entidades seguintes foram selecionadas:

1. Modelo – forma de identificação do produto;
2. Marca – fabricante do produto;
3. Dimensão – dimensão(ões) física(s) do produto;
4. Grandeza – valores como peso, capacidade etc;
5. Cor – cor aplicada ao produto;
6. Utilidade – função que o produto desempenha;

[1]Dados do relatório web shoppers do EBIT. Disponível em: http://www.ebitempresa.com.br/web-shoppers.asp. Acesso em: 28 jun. 2015.

7. Material – material constituinte do produto;
8. Parte – produto ou objeto constituinte de outro produto ou objeto;
9. Objeto – mercadoria, bem de consumo.
10. Produto – bem de consumo comercializável.

Para o reconhecimento automático dessas entidades, foi aplicada a técnica de *Conditional Randon Fields* (CRF), descrita na seção 2. O treinamento do reconhecedor automático de entidades nomeadas no domínio do *e-commerce* (seção 3) e os resultados obtidos até o momento (seção 4) são apresentados em seguida. Por fim, a seção 5 traz algumas propostas de trabalhos futuros.

2. Reconhecimento de Entidades Nomeadas

Para contextualizar a tarefa de reconhecimento de entidades nomeadas no domínio do *e-commerce*, considere a sentença apresentada na Figura 1.

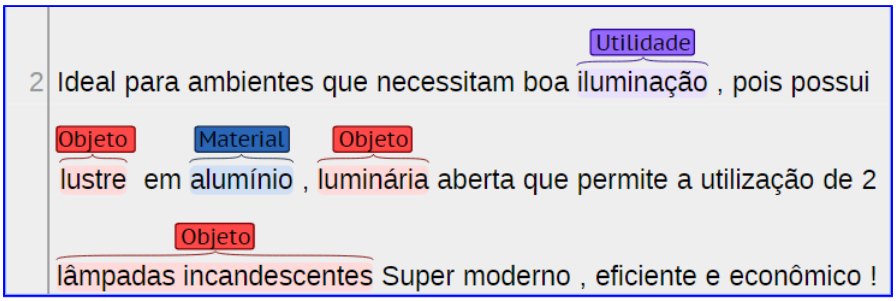


Figura 1. Exemplo de sentença do corpus de treinamento na qual três tipos de entidades nomeadas estão anotadas: Utilidade (1 ocorrência), Objeto (3 ocorrências) e Material (1 ocorrência).

Uma das técnicas mais utilizadas na atualidade para reconhecimento automático de entidades nomeadas é a *Conditional Random Fields* (CRF), um modelo probabilístico cujo objetivo é etiquetar e segmentar dados sequenciais, utilizando-se de uma abordagem condicional [Lafferty et al. 2001]. O CRF é descrito como uma união de representação gráfica, especificamente um grafo não direcionado, e de um modelo probabilístico condicional, que atribui uma distribuição logaritmicamente linear sobre conjuntos de sequências de rótulos, dada uma sequência de observação especificada.

Formalmente, a definição deste modelo é derivada de uma abstração de grafo e de seu modelo probabilístico. Seja $x = [x_1, ..., x_n]$ um conjunto de sequência de dados a serem etiquetados. Seja $y = [y_1, ..., y_n]$ um conjunto de sequência de etiquetas correspondentes a x. Neste contexto, x é o conjunto de atributos que define uma sequência a ser classificada. Tais atributos podem conter como valores, dados como todas as palavras da sentença, bem como sua etiquetação de *part-of-speech* e posição de tais dados na sentença. Em contrapartida, y é definido como um conjunto de etiquetações-alvo da sequência x; é o conjunto de combinações de etiquetação possíveis para x.

Neste âmbito, é preciso definir também a distribuição aplicada ao sistema. Assim, seja $\Phi = \{\phi_1(D_1), ..., \phi_k(D_k)\}$ um conjunto de distribuição, baseada numa distribuição de Gibbs [Geman e Geman 1984], onde ϕ representa o fator de distribuição e D o escopo destes fatores. Dessa forma, $P_\Phi(y|x)$ é a probabilidade P de y, dado o conjunto de atributos x, ou seja, $P_\Phi(y|x)$ é a probabilidade de y, dado x. Segundo CRF, $P_\Phi(y|x)$ está

definido como:

$$P_\Phi(y|x) = \frac{\tilde{P}_\Phi(x,y)}{Z_\Phi(x)}$$

em que $\tilde{P}_\Phi(x,y)$ representa a medida não normalizada, e $Z_\Phi(x)$ representa a função de normalização. $\tilde{P}_\Phi(x,y)$ é definido como:

$$\tilde{P}_\Phi(x,y) = \prod_{i=1}^{k} \phi_i(D_i)$$

e sua função de normalização como:

$$Z_\Phi(x) = \sum_y \tilde{P}_\Phi(x,y)$$

Nesta concepção, D é definido como o escopo de cada ϕ. Esse escopo é estabelecido segundo finalidades propostas ou por modelos já utilizados. A Figura 2 demonstra um exemplo de pesos de distribuição no grafo.

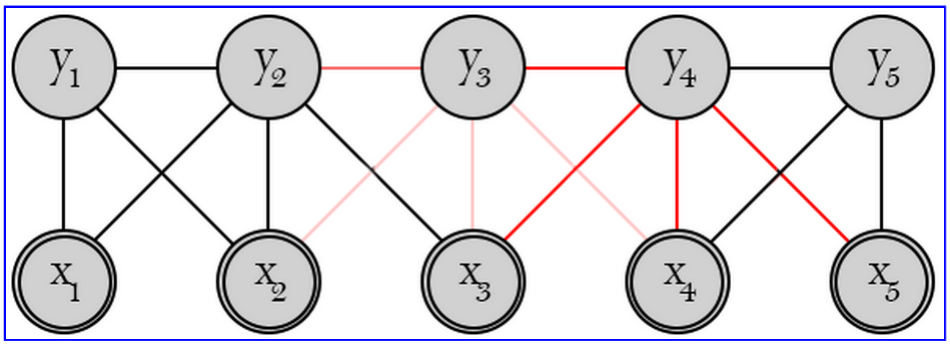


Figura 2. Exemplo de escopo de distribuição em y_4

Na Figura 2, a rotulação presente em y_4 é influenciada diretamente pelos atributos e rótulos aos quais ela está diretamente ligada, demarcados em vermelho. As relações mostradas em rosa são influências indiretas, também, na definição de y_4. O exemplo demonstrado na Figura 2 é apenas uma exemplificação de como podem ser definidas as distribuições ϕ, mas tais distribuições podem representar quaisquer valores, dependendo da finalidade da aplicação. Ou mesmo, podem ser definidos automaticamente segundo outro método de aprendizado de máquina.

A aplicação desse método deu-se por meio da ferramenta CRFSharp[2], uma implementação de CRF em C#. Atualmente, no treinamento com o corpus, o CRFSharp faz uso de processadores multinúcleo e uso eficiente de memória, especialmente para densos corpora de treinamento.

[2]Disponível em: `http://crfsharp.codeplex.com/documentation`. Acesso em: 01 jul. 2015.

3. Treinamento

O corpus de treinamento utilizado no experimento descrito neste artigo é composto por um conjunto de textos (com descrição de produtos) retirados de páginas de *e-commerce* (páginas de lojas *online*). As entidades nomeadas de interesse presentes neste corpus foram, então, anotadas por dois linguistas utilizando a ferramenta BRAT[3] [Stenetorp et al. 2012]. A concordância entre os anotadores foi computada por meio da medida Kappa (κ) [Carletta 1996], e o valor obtido foi de $\kappa = 0{,}73$.

Como resultado do processo de anotação manual do corpus obteve-se um conjunto de 295 arquivos anotados, com 5.884 ocorrências de entidades nomeadas, sendo a quantidade de instâncias em cada tipo de entidade como segue: 1.421 de Parte, 1.080 de Grandeza, 818 de Utilidade, 792 de Produto, 743 de Modelo, 579 de Objeto, 505 de Marca, 296 de Material, 221 de Dimensão e 221 de Cor.

No treinamento do modelo, para cada palavra, as seguintes 13 *features* foram utilizadas: a forma superficial (palavra como ocorre no texto) e *part-of-speech* (POS)[4] da palavra e das duas palavras anteriores e posteriores a ela, além da indicação se a palavra ou as palavras imediatamente anterior ou posterior a ela contêm número (Verdadeiro/Falso). Como exemplo, considere a sentença apresentada na Figura 3 para a qual as *features* geradas para a palavra "recebidas" são apresentadas em seguida.[5]

Registro de 20 chamadas **recebidas** (atendidas e não atendidas) e 20 chamadas realizadas

feature	**valor**
palavra	recebidas
palavra - 1	chamadas
palavra - 2	20
palavra + 1	(
palavra + 2	atendidas
POS da palavra	ADJ
POS da palavra - 1	N
POS da palavra - 2	NUM
POS da palavra + 1	PUN
POS da palavra + 2	ADJ
palavra contém número?	Falso
palavra - 1 contém número?	Falso
palavra + 1 contém número?	Falso

Figura 3. Sentença retirada do corpus de treinamento e conjunto de *features* para a palavra "recebidas"

A ferramenta CRFSharp avalia o conjunto de *features* para cada palavra do corpus e, a partir disso, gera o modelo de treinamento segundo tais especificações.

[3]Disponível em: `http://brat.nlplab.org/`. Acesso em: 09 jul. 2015.

[4]A etiquetação de POS foi realizada por outra equipe de linguistas em tarefa paralela à anotação de entidades nomeadas.

[5]As POS neste exemplo indicam: adjetivo (ADJ), substantivo (N), numeral (NUM) e símbolo de pontuação (PUN).

4. Resultados

O modelo treinado usando o CRF, o corpus e as *features*, como descritos anteriormente, foi aplicado a um corpus de teste. O corpus utilizado como teste, assim como o corpus de treinamento, é formado por textos retirados de páginas de *e-commerce*. As entidades nomeadas presentes no corpus de teste também foram anotadas pelos mesmos dois linguistas do corpus de treinamento gerando um corpus de referência utilizado na avaliação automática. Após a anotação manual do corpus, obteve-se um conjunto de 10 arquivos anotados, com 279 ocorrências de entidades nomeadas, sendo a quantidade de instâncias em cada tipo de entidade como segue: 58 de Utilidade, 55 de Parte, 48 de Grandeza, 28 de Produto, 28 de Objeto, 23 de Modelo, 12 de Marca, 10 de Dimensão, 9 de Material e 8 de Cor.

A avaliação da ferramenta aplicada ao corpus de teste foi realizada automaticamente por meio das medidas de precisão (P) e cobertura (C), calculadas segundo as fórmulas:

$$P = \frac{|W_a \cap W_c|}{|W_a|} \qquad C = \frac{|W_a \cap W_c|}{|W_c|}$$

em que W_a representa o conjunto de entidades anotadas automaticamente pelo modelo, e W_c o conjunto de entidades anotadas manualmente pelos linguistas (corpus de referência), sendo $|W_a \cap W_c|$ o conjunto de entidades anotadas corretamente segundo o modelo.

A Tabela 1 traz as quantidades de instâncias de cada entidade nomeada anotadas pela ferramenta, pelos linguistas e anotadas corretamente pela ferramenta, bem como mostra os resultados de precisão e cobertura obtidos para cada entidade em questão, e o resultado geral.

	Anotados			**Resultados**	
Entidade	**Pela ferramenta**	**Corretamente**	**Pelos linguistas**	**Cobertura**	**Precisão**
Marca	12	12	12	100,00%	100,00%
Grandeza	48	42	48	87,50%	87,50%
Material	8	7	9	77,78%	85,50%
Produto	25	20	28	71,43%	80,00%
Dimensão	7	7	10	70,00%	100,00%
Parte	70	36	55	65,45%	51,43%
Modelo	18	13	23	56,52%	72,22%
Utilidade	34	29	58	50,00%	85,29%
Cor	4	4	8	50,00%	100,00%
Objeto	16	5	28	17,86%	31,25%
S (Sem EN)	854	785	817	96,07%	91,92%
Total	1096	960	1096	**87,59%**	

Tabela 1. Resultados de cobertura para cada entidade obtidos pela anotação automática.

Como exposto na Tabela 1, a entidade Objeto foi a que obteve piores resultados, atingindo apenas 17,86% em cobertura e 31,25% em precisão, enquanto Marca, Dimensão

e Cor obtiveram os melhores resultados de precisão, atingindo 100%. O excelente resultado obtido por essas entidades pode ser explicado com base na uniformidade do contexto no qual elas ocorrem (os contextos de ocorrência de cada uma são muito semelhantes). De modo similar, os baixos resultados obtidos pela entidade Objeto podem ser justificados pela considerável variação de seu contexto de aparição, além da semelhança de seu contexto com o de outras entidades, tornando-a uma entidade mais complexa para ser classificada. Além disso, a Tabela 1 mostra os resultados gerais de cobertura e precisão obtidos pela ferramenta, sendo ambos 87,59%.

As principais avaliações da eficácia, bem como a comparação de sistemas de REN, ocorreram por meio de dois eventos: *Conference on Computational Natural Language Learning* [CoNLL 2002] e a Avaliação de Reconhecedores de Entidades Mencionadas (HAREM) [Mota e Santos 2008]. Para fins de comparação, no HAREM 2008 a ferramenta SeRELeP [Bruckschen et al. 2008] foi a que obteve maior precisão na tarefa de Reconhecimento de Entidades Nomeadas, atingindo 81,31%. Dessa forma, é possível observar que o resultado de 87,59% alcançado pela ferramenta treinada para o artigo em questão é altamente satisfatório.

A Figura 4 traz um exemplo de uma sentença do corpus de teste anotada automática e manualmente.

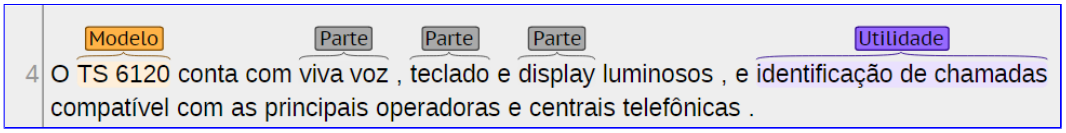
a	
b	O TS[Modelo] 6120[Modelo] conta com viva[Parte] voz[Parte], teclado[Parte] e display[Parte] luminosos, e identificação[Parte] de[Parte] chamadas[Parte] compatível com as principais operadoras e centrais telefônicas.

Figura 4. Sentença anotada pelos linguistas (a) e sentença anotada automaticamente (b).

5. Trabalhos Futuros

Como trabalhos futuros pretende-se investigar o uso de novas *features* na busca por um aumento de precisão e cobertura da ferramenta, em especial da entidade Objeto. Para tal, uma nova *feature* ou um relacionamento entre *features* pode ser considerado.

Agradecimentos

Esse trabalho foi realizado em parceria e com suporte da Boolabs (`http://boolabs.com.br/`) e faz parte do projeto de extensão "Processamento de texto e de imagem na descrição *online* de produtos" (UFSCar/FAI #23112.003944/2014-81).

Referências

Bruckschen, M., Vieira, R., e Rigo, S. (2008). Reconhecimento automático de relações entre entidades mencionadas em textos de língua portuguesa. In *Anais do CELSUL 2008*, pages 1–11.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

CoNLL (2002). Conference on Computational Natural Language Learning. Taipei, Taiwan.

Geman, S. e Geman, D. (1984). In *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*.

Jing, J. (2012). Information extraction from text. In *Mining Text Data*, pages 11–41.

Lafferty, J., McCallum, A., e Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ling, X. e Weld, D. S. (2012). Fine-grained entity recognition. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

Mota, C. e Santos, D. D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., e Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *EACL '12 Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.