Extração de relações semânticas de textos em português do Brasil no domínio do *e-commerce*

Leonardo Henrique Tozzatto Volpe[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/`

`leo_volpe@hotmail.com, helenacaseli@dc.ufscar.br`

1. Introdução

O comércio eletrônico brasileiro figura entre os 10 maiores em *ranking* mundial, representando 3,8% de todas as vendas no varejo do país [Company 2015]. Apesar dessa relevância econômica, verificamos dificuldade em encontrar trabalhos na área de processamento de textos em português brasileiro no domínio do *e-commerce*. A maior parte da informação veiculada em páginas de *e-commerce* é apresentada na forma de texto em língua natural de modo não estruturado (como os textos que contêm as descrições de produtos) sendo necessário, portanto, coletá-las, organizá-las e compará-las por meio da aplicação de técnicas computacionais. Neste contexto, este trabalho apresenta uma ferramenta treinada para extração automática de relações semânticas binárias de textos puros no domínio do *e-commerce*.

Uma relação semântica é uma especificação de relacionamento entre termos[1] dentro de uma sentença. Assim como em [Taba 2013], as relações de interesse neste trabalho são binárias, representadas na forma `relação(termo1, termo2)`, por exemplo isa(maçã, fruta). Essas relações semânticas são muito úteis no processamento de textos, uma vez que são indicativas de similaridade semântica entre sentenças em língua natural. Neste trabalho, o intuito é verificar a ocorrência de tais relações em textos de um domínio bastante específico: o *e-commerce*.

Para contextualizar a tarefa de extração de relações semânticas no domínio do *e-commerce*, considere a sentença "Sua estrutura feita em aço inox" retirada do corpus de treinamento utilizado neste trabalho. Neste exemplo, "estrutura" e "aço inox" são termos e há uma relação semântica do tipo made-of (feito de) entre eles, evidenciada pelas palavras "feita em" entre os dois termos: made-of(estrutura, aço inox).

Assim, este artigo traz o resultado do treinamento de uma ferramenta automática para extração de cinco tipos de relações semânticas em textos no domínio do *e-commerce*. Das sete relações semânticas investigadas em [Taba 2013], cinco foram selecionadas após análise de ocorrência em textos de *e-commerce*:

1. property-of(algo, característica)
   *Ex.: "Sua **estrutura** feita em aço inox é **compacta**"*
   property-of(estrutura, compacta)

[1]No contexto deste trabalho, um termo é uma sequência de *tokens* (sequências de caracteres com exceção do espaço) que representa uma entidade com algum significado específico na sentença. O termo, portanto, não tem nenhuma relação direta com seu significado adotado na área de terminologia.

2. is-a(subclasse, superclasse)
   *Ex.: "(...) utilize alguma **fruta**, como a **maçã**"*
   is-a(maçã, fruta)
3. part-of(todo, parte)
   *Ex.: "A **pipoqueira** conta com os **cabos** em baquelite"*
   part-of(pipoqueira, cabos)
4. made-of(produto, material)
   *Ex.: "Sua **estrutura** feita em **aço inox** é compacta"*
   made-of(estrutura, aço inox)
5. used-for(entidade, função)
   *Ex.: "(...) ela tem uma **pá giratória** no fundo, que serve para **mexer os grãos de pipoca**"*
   used-for(pá giratória, mexer os grãos de pipoca)

Este artigo está estruturado como segue. Na seção 2 apresenta-se a tarefa de extração automática de relações semânticas, com uma breve descrição das principais estratégias, bem como os resultados obtidos para o português do Brasil em outros domínios. Em seguida, descreve-se o corpus de treinamento usado neste trabalho (seção 3), os resultados obtidos até então (seção 4) e algumas considerações e propostas de trabalhos futuros (seção 5).

2. Extração de relações semânticas

Para reconhecimento de relações semânticas, as ferramentas devem ser capazes de receber como entrada um texto com marcação de entidades (termos) e retornar a melhor relação entre esses termos utilizando, por exemplo, métodos mais simples baseados em padrões (como expressão regular), como proposto por [Hearst 1992], sacrificando cobertura para obter grande precisão [Yap e Baldwin 2009]; ou métodos mais novos baseados em aprendizado de máquina tendo como *features* (atributos) anotações produzidas por etiquetadores e analisadores sintáticos [Girju et al. 2003, Snow et al. 2005, Caraballo 1999, Taba 2013, Taba e Caseli 2014].

Um dos primeiros métodos a utilizar padrões textuais para extrair relações semânticas foi proposto por [Hearst 1992]. Nesse método, a identificação de relações se dava em textos com processamento sintático raso, utilizando diversos padrões construídos manualmente ou semiautomaticamente com a observação do corpus utilizado [Hearst 1992, Hearst 1998, Berland e Charniak 1999, Girju e Moldovan 2002]. Hearst utilizou um corpus jornalístico (*New York Times*) para avaliar a aplicação de seus padrões textuais comparando as relações extraídas com dados da WordNet[2] 1.5 [Fellbaum 1998] e relatou que, de 200 sentenças utilizadas como amostra, 166 relações válidas foram extraídas, sendo 104 dessas (em torno de 63%) compatíveis com dados da WordNet.

Métodos baseados em aprendizado de máquina utilizam classificadores como as *Support Vector Machines* (SVMs) [Vapnik 1995], que encontram, através de propriedades geométricas, um hiperplano que melhor separe amostras de treinamento entre duas classes [Taba 2013]. Esses métodos foram utilizados para extrair relações de

[2]WordNet é um banco de dados lexical do inglês. Disponível em `wordnet.princeton.edu`. Acesso em: 02 set. 2015.

meronímia [Girju et al. 2003] e hiperonímia [Snow et al. 2005, Caraballo 1999] em trabalhos baseados no procedimento desenvolvido por [Hearst 1992], entre outros.

A maioria dos trabalhos desenvolvidos para reconhecimento de relações semânticas foi feita com base no inglês como língua-alvo, assim como muitos dos recursos utilizados na área, mas existem alguns poucos trabalhos focados na língua portuguesa. Em [Freitas e Quental 2007], os padrões de [Hearst 1992, Hearst 1998] são adaptados e expandidos para a língua portuguesa, obtendo 73,4% de relações corretas em sua avaliação quando aplicados sobre um corpus de textos sobre saúde pública coletados da internet etiquetado pelo PALAVRAS[3] [Bick 2000].

Podem ser citados também três sistemas submetidos para o ReRelEM (Reconhecimento de Relações entre Entidades Mencionadas)[4] de 2008 para avaliar sistemas de reconhecimento de entidades nomeadas: REMBRANDT [Cardoso 2008], SEI-Geo [Chaves 2008] e SeRELeP [Bruckschen et al. 2008]. Dos sistemas supracitados, REMBRANDT obteve melhores resultados – 58,2%, 36,7% e 45% para precisão, cobertura e medida-F, respectivamente –, porém ainda não são resultados suficientemente bons, principalmente quando comparados aos sistemas utilizados para a língua inglesa.

Em [Taba 2013, Taba e Caseli 2014], duas abordagens foram investigadas para reconhecimento de relações semânticas (entre elas as 5 relações de interesse neste trabalho). A primeira delas está baseada em trabalhos de [Hearst 1992] e [Freitas e Quental 2007] e utiliza padrões textuais para encontrar relações de hiponímia (is-a). A segunda baseia-se em aprendizado de máquina utilizando corpora anotados[5] e diferentes classificadores[6].

Com os seis experimentos de Taba, dois utilizando como abordagem padrões textuais e quatro utilizando aprendizado de máquina, ficou evidente o melhor desempenho do aprendizado de máquina utilizando *Support Vector Machines* sobre os padrões textuais, que apresentam dificuldades pelas ambiguidades existentes na língua natural, além do alto custo para definição de novos padrões. Os valores de precisão encontrados por Taba são apresentados na Tabela 1.

3. Extração de relações semânticas no domínio do *e-commerce*

Para possibilitar a geração de uma ferramenta capaz de extrair automaticamente relações semânticas em textos no domínio do *e-commerce*, um corpus desse domínio foi montado com textos extraídos de páginas *web* com descrição de produtos.[7] As instâncias das cinco relações semânticas presentes nesse corpus de treinamento foram anotadas por dois linguistas utilizando a ferramenta BRAT[8] [Stenetorp et al. 2012]. Para tanto, uma pequena

[3]PALAVRAS é um analisador sintático automático (*parser*) para português [Bick 2000].

[4]O ReRelEm é uma tarefa para avaliação de sistemas que identificam relações semânticas dentro do HAREM (Avaliação de Reconhecedores de Entidades Mencionadas), um evento organizado pela Linguateca (`www.linguateca.pt`).

[5]Os corpora utilizados em [Taba 2013] foram: corpus da revista Pesquisa FAPESP composto por 646 artigos científicos e corpus CETENFolha, com milhões de palavras provenientes de diversos artigos do jornal Folha de São Paulo (1994).

[6]Os classificadores utilizados em [Taba 2013] foram: *Support Vector Machines* [Vapnik 1995] e árvores de decisão C4.5 [Quinlan 1993].

[7]Esses textos foram fornecidos pela empresa Boo, parceira desta pesquisa, e fazem parte do banco de dados gerenciado pela empresa.

[8]Disponível em: `brat.nlplab.org`. Acesso em: 01 jul. 2015.

parte do corpus foi separada para testar a precisão do modelo final e o corpus restante foi dividido em duas partes com uma intersecção entre elas para permitir o cálculo da concordância entre os anotadores, a qual foi de $\kappa = 0{,}63$ [Carletta 1996].[9]

Ao final do processo de anotação obteve-se um corpus de treinamento composto por 291 arquivos com 47.312 *tokens*, 9.922 termos e 5.924 instâncias de relações semânticas. As quantidades de instâncias para cada tipo de relação semântica são apresentadas na Tabela 1.

Em [Taba 2013], foram descritas 288 *features* utilizadas para a classificação das relações no corpus e, dessas, 216 são *features* sintáticas, que utilizam etiquetas do *parser* PALAVRAS [Bick 2000] para representar o contexto sintático e as características sintáticas de cada termo. Tais *features* não foram consideradas para gerar o modelo SVM deste artigo pois não é desejado que tal modelo dependa de ferramentas proprietárias como o PALAVRAS [Bick 2000]. Assim, a ferramenta descrita neste artigo foi construída com base em um modelo SVM treinado com as 72 *features* restantes de [Taba 2013]: 36 superficiais (como a distância entre termos, ordem dos termos e tamanho de cada termo) e 36 morfológicas (como a existência de artigo antes dos termos, verbo "ser" entre termos e as etiquetas POS ao redor dos termos).

Para utilizar *features* morfológicas no treinamento do aprendizado de máquina, anotações de *part-of-speech* (POS) realizadas em paralelo por outra dupla de linguistas foram unidas às anotações de relações semânticas. Após a união das anotações, os arquivos finais foram convertidos para o formato de entrada da ferramenta ARS [Taba e Caseli 2012].[10] Durante o processo de união e conversão das anotações, nos casos das intersecções entre os as partes do corpus anotadas por diferentes anotadores, foram priorizadas as anotações que continham mais relações semânticas entre termos presentes na mesma sentença.[11]

Utilizando a ARS, todo o corpus anotado foi convertido em formato de entrada para a LibSVM[12] [Chang e Lin 2011]. Com o arquivo de entrada pronto para a LibSVM, suas *features* foram escaladas no intervalo [-1, 1] utilizando o svm-scale e um modelo foi criado com o treinamento feito pelo svm-train.

4. Resultados

Após o treinamento do modelo usando a LibSVM, a ferramenta svm-predict foi utilizada para avaliar a precisão do modelo verificando quantas relações de um corpus de teste anotado (não incluso no utilizado para treinamento do modelo) seriam classificadas cor-

[9] As concordâncias entre os anotadores foram as seguintes para cada tipo de relação semântica: 48,15% para used-for (13 casos em 27), 57,56% para is-a (118 casos em 205), 71,79% para part-of (28 casos em 39), 78,26% para made-of (18 casos em 23) e 88,22% para property-of (292 casos em 331).

[10] A ARS (Anotador de Relações Semânticas), desenvolvida em Java, manipula sentenças em formato JSON, permitindo marcação visual de termos e relações, além de outras funcionalidades [Taba e Caseli 2012]. Disponível em: `www.lalic.dc.ufscar.br/portal`. Acesso em: 02 jul. 2015.

[11] Para converter o corpus, a ARS [Taba e Caseli 2012] divide os textos por sentenças e não aceita, portanto, relações entre termos que ocorram em sentenças diferentes.

[12] A LibSVM é uma biblioteca para ferramentas de classificação *Support Vector Machines*, permitindo treinar modelos de classificadores e estimar a precisão de um modelo utilizando *cross-validation*. A LibSVM está disponível em: `www.csie.ntu.edu.tw/~cjlin/libsvm`. Acesso em: 10 jul. 2015.

retamente pelo modelo treinado.[13] Os resultados para cada relação e para o teste em geral são apresentados na Tabela 1. Para efeito de comparação, essa tabela também traz os resultados obtidos com a aplicação do modelo de [Taba 2013] treinado com todas as *features* (incluindo as sintáticas) no corpus CETENFolha[14]. Vale ressaltar que o modelo treinado neste trabalho utiliza apenas *features* superficiais e morfológicas.

	Precisão	
Relação	**Este trabalho**	**Modelo de [Taba 2013]**
property-of	92,9% (566/609)	82,2% (501/609)
is-a	83,3% (10/12)	50,0% (6/12)
used-for	81,4% (44/54)	77,7% (42/54)
part-of	72,7% (48/66)	69,6% (46/66)
made-of	65,3% (17/26)	65,3% (17/26)
Geral	91,9% (983/1069)	85,5% (914/1069)

Tabela 1. Precisão obtida pelos modelos treinados neste trabalho e em [Taba 2013] aplicados ao corpus de teste de *e-commerce*

A Figura 1 traz um exemplo de uma sentença do corpus de teste (a) processada pelo modelo de [Taba 2013] (treinado com todas as *features* no corpus CETENFolha) e (b) pelo modelo treinado como descrito neste trabalho. Com base neste exemplo é possível ver que o retreinamento do modelo a partir de um corpus específico permitiu o reconhecimento de 2 instâncias de relações semânticas não reconhecidas pelo modelo de [Taba 2013].

a	O TCF 1000 é o aparelho telefônico com fio para quem busca objetividade.
b	O **TCF 1000** é o **aparelho telefônico** com fio para quem busca objetividade. is-a(TCF 1000, aparelho), property-of(aparelho, telefônico)

Figura 1. Sentença anotada com o modelo treinado em [Taba 2013] (a) e o modelo treinado com corpus de *e-commerce* (b)

5. Trabalhos futuros

Como trabalhos futuros pretende-se investigar novas *features* no treinamento do modelo, como o uso de entidades nomeadas que também estão presentes nos arquivos do corpus de treinamento. Também deve-se avaliar o desempenho de outros classificadores utilizando as mesmas *features* e corpus para comparação com os resultados deste trabalho.

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).

[13] Valores de cobertura e medida-F não se aplicam ao teste realizado, pois todas as relações anotadas no corpus de teste recebem uma classificação baseada no modelo treinado.

[14] Disponível em: `http://www.linguateca.pt/cetenfolha/`. Acesso em: 03 jul. 2015.

Referências

Berland, M. e Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64. College park, MD.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Denmark: Aarhus University Press.

Bruckschen, M., Muniz, F., Souza, J. G. C., Fuchs, J. T., Infante, K., Muniz, M., Gonçalves, P. N., Vieira, R., e Aluísio, S. M. (2008). Anotação Lingüística em XML do Corpus PLN-BR. Technical Report Série de Relatórios do NILC, NILC-ICMC-USP.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cardoso, N. (2008). Rembrandt - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In Mota, C. e Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 11, pages 195–211. Linguateca, Portugal.

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

Chang, C. C. e Lin, C. J. (2011). LIBSVM : a library for support vector machines. In *ACM Transactions on Intelligent Systems and Technology*, pages 2:27:1–27:27.

Chaves, M. S. (2008). Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo HAREM. In Mota, C. e Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, chapter 13, pages 231–245. Linguateca, Portugal.

Company, U. B. (2015). Brasil é o décimo melhor mercado de e-commerce do mundo. `http://www.profissionaldeecommerce.com.br/brasil-e-o-decimo-melhor-mercado-de-e-commerce-mundo/`. Acesso em: 02 set. 2015.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT press.

Freitas, M. C. e Quental, V. (2007). Subsídios para a elaboracão automática de taxonomias. In *V Workshop em Tecnologia da Informação e da Linguagem Humana - TIL 2007*, pages 1585–1594.

Girju, R., Badulescu, A., e Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Girju, R. e Moldovan, D. (2002). Mining answers for causation. In *Proceedings of American Association of Artificial Intelligence*, pages 15–25.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational linguistics*, volume 2, pages 539–545, Nantes, France. ACL.

Hearst, M. A. (1998). Automated discovery of wordnet relations. In *WordNet: An electronic lexical database*, chapter 5, pages 131–151. The MIT press.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Snow, R., Jurafsky, D., e Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, number 17, pages 1297–1304. MIT Press.

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.

Taba, L. S. (2013). Extração automática de relações semânticas a partir de textos escritos em português do brasil. Master's thesis, Universidade Federal de São Carlos.

Taba, L. S. e Caseli, H. M. (2012). Uma ferramenta para anotação de relações semânticas entre termos. In *Anais do XI Encontro de Linguística de Corpus – ELC 2012*.

Taba, L. S. e Caseli, H. M. (2014). Automatic semantic relation extraction from portuguese texts. In *Proceedings of the 9th Language Resources and Evaluation Conference*, pages 2739–2746, Reykjavik, Iceland.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer Verlag.

Yap, W. e Baldwin, T. (2009). Experiments on pattern-based relation learning. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1657–1660.