RESENHA

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications**; text retrieval, extraction and categorization. Amsterdam / Philadelphia: John Benjamins Publ., 2002. 225p.

As pesquisas de Processamento de Linguagem Natural (PLN) têm sido produtivas há mais de quatro décadas, tanto no âmbito universitário, sobretudo em Departamentos de Informática ou Ciência da Computação, Engenharia de Software, Inteligência Artificial e Lingüística Computacional, bem como na indústria da Tecnologia da Informação. Uma vista de olhos pelas seções da Terceira Conferência Internacional sobre os Avanços em PLN, realizada em 2002 em Portugal (1), oferece um panorama diversificado, abarcando disciplinas e/ou áreas de pesquisa tais como: pragmática, discurso e léxico; interpretação e geração de linguagem falada e escrita; recuperação informacional, sistemas de perguntas e respostas, sumarização e extração informacional; aprendizagem automática baseada na linguagem; processamento multilíngue, tradução automática e tradução auxiliada por computador; sistemas de interfaces e diálogo em linguagem natural; desenvolvimento de ferramentas e recursos, além da avaliação de sistemas – panorama esse, no qual vêm sendo incorporados, em escala crescente, recursos direcionados aos desafios e às possibilidades da Web.

Uma contribuição relevante nesta área, ainda carente de visões de conjunto acessíveis sobre o estado-da-arte, pode ser considerada a obra em foco, centrada sobre algumas tecnologias da informação tidas como essenciais pelos autores. O livro compõe-se de cinco capítulos, a saber: 1. Processamento da linguagem natural; 2. Recuperação de documentos; 3. Extração informacional; 4. Categorização textual e 5. Mineração textual. Cada capítulo oferece uma apresentação introdutória da respectiva tecnologia da informação, seguindo-se uma exposição de tópicos mais complexos, com seções opcionais, notas explicativos em formato corrido e "dicas", além de referências selecionadas de fontes disponíveis até 2001, sob forma impressa ou virtual.

No primeiro capítulo, os autores situam a problemática geral do processamento informacional da linguagem natural (PLN), enfocando sucintamente alguns de seus aspectos teóricos e recursos práticos à disposição para a construção de *software*. Historicamente, nas pesquisas na área tentava-se solucionar os problemas referentes à manipulação de símbolos, processamento esse condicionado pela complexidade das regras gramaticais da linguagem humana. A partir da década de 1990, começaram a ser introduzidas sofisticadas abordagens estatísticas no processamento da linguagem natural, trabalhando–se com enormes quantidades

de dados lingüísticos, oriundos, por exemplo, de acervos de agências de notícias e páginas da Web. Os autores observam uma tendência para o desenvolvimento de programas capazes de executar automaticamente tarefas múltiplas *(supertasks)* tais como, por exemplo: selecionar documentos de uma base com enfoque no seu conteúdo, agrupá-los em categorias ou classes e deles extrair determinados conjuntos de informações. Cada tipo de tecnologia informacional descrito pressupõe o domínio de ferramentas lingüísticas específicas, tradicionalmente desenvolvidas na Lingüística Computacional e em projetos de pesquisa interdisciplinares de áreas afins. Dessas ferramentas fazem parte os chamados delimitadores de sentenças *(sentence delimiters);* itemizadores (*tokenizers*) identificadores de radicais de palavras *(stemmers)* e codificadores de partes do discurso oral ou escrito *(speech taggers)*. Os identificadores de sintagmas nominais *(noun phrase recognizers)* e de nomes próprios, bem como os analisadores gramaticais *(parsers)* pertencem hoje ao arsenal das ferramentas clássicas, utilizadas na implementação de *softwares* específicos, direcionados ao processamento da informação baseada na linguagem humana.

O capítulo dois trata da recuperação (automática) de documentos, uma vez que os autores preferem reservar o termo "recuperação informacional" para um contexto mais abrangente, que inclua também a recuperação de imagens e de áudio, além de documentos com características não-textuais, tais como notação musical, dados tabulados, equações e similares. No que diz respeito à recuperação informacional de textos, os fundamentos "lógicos" apresentados são do tipo booleano, além da técnica da recuperação ordenada *(ranked retrieval)* dos resultados, baseada na distribuição de frequência dos termos de busca. O capítulo contém uma breve revisão sobre resultados práticos gerados a partir das *Text Retrieval Conferences (TREC)*, co-patrocinadas, desde 1992, pela agência norte-americana de Normas Técnicas *(National Institute of Standards and Technology (NIST)* e a *Defence Advanced Research Projects Agency (DARPA)*, do Ministério da Defesa dos EUA. Os autores discorrem sobre avanços recentes nas pesquisas da área, que resultaram em sistemas comercializados, lembrando que as crescentes facilidades de acesso pelo usuário comum à informação veiculada na Internet só foram possíveis graças a uma progressiva sofisticação das tecnologias da informação alavancadas pelo mega-projeto TREC (2). Quanto ao processamento de questões *(query processing),* sob forma da chamada busca booleana, ainda a técnica de recuperação informacional mais difundida atualmente, os autores expõem alguns aperfeiçoamentos mais recentes sem, contudo, omitir as limitações inerentes a essa modalidade de recuperar informações. A tecnologia da recuperação probabilística, em uso a partir da década de 1970, merece uma descrição detalhada, assim como a chamada

"modelagem lingüística" *(language modeling approach)*, desenvolvida em 1998 no *Massachusetts Institute of Technology*. São abordadas, ainda, técnicas de avaliação de máquinas de busca, terminando o capítulo com uma visão prospectiva da recuperação informacional na Web. Os autores lembram que a Web constitui, de fato, um imenso laboratório para experimentos de recuperação da informação em grande escala, possibilitando o desenvolvimento de ferramentas destinadas a auxiliar no controle da relevância e o aperfeiçoamento de testes para a avaliação do desempenho das próprias máquinas de busca.

O capitulo três é dedicado à tecnologia da extração automática de informações a partir de textos, modalidade de PLN direcionada não à recuperação de documentos inteiros (tratada no capítulo anterior), mas de informações específicas contidas nos próprios documentos. Pela extração informacional recuperam-se, então, informações propriamente ditas, por exemplo, referentes a fatos ou eventos, em documentos textuais não estruturados. O recurso tem sido utilizado desde a década de 1970 por Agências de Inteligência para rastrear fluxos informacionais de entidades de notícias e de comunicações, em língua inglesa e em outros idiomas. Nesse contexto, os autores mencionam alguns resultados gerados pelas *Message Understanding Conferences (MUC)*, projeto fomentado também pelo Ministério da Defesa dos EUA desde 1989, destacando-se aqui aplicações de analisadores sintáticos *(parsers)* tradicionais e analisadores de contexto-livre, além da chamada técnica de autômatos finitos. Explicações teóricas são exemplificadas com aplicações em ambientes informatizados sobre o noticiário geral de imprensa e em bases de informação jurídica. No caso da extração informacional, trata-se de um processo complexo, sendo que nenhum dos programas disponíveis atinge 100 por cento de revocação ou de precisão, ao identificar os itens de interesse pesquisados pelo usuário. Tais sistemas, na visão dos autores, devem ser considerados recursos adicionais em sistemas de editoração manual. Teoricamente, um sistema deste tipo pode ser programado para "aprender" a incorporar automaticamente regras de extração informacional a partir de casos exemplares, do que resultaria a uma manutenção automática do mesmo. De acordo com os autores, será necessário aprofundar as pesquisas para chegar a resultados mais satisfatórios, embora alguns sistemas de extração informacional automática, desenvolvidos na seqüência do Programa MUC, tenham atingido uma fase madura. Já existem algoritmos suficientemente eficientes para processar grandes quantidades de textos, desde que o usuário queira analisar apenas determinadas porções discursivas em documentos, que contenham termos de busca por ele selecionados.

No capítulo quatro, destinado à tecnologia da classificação ou categorização automática de textos *(text categorization),* são analisados problemas básicos a serem

resolvidos pelos respectivos algoritmos. Os autores utilizam-se indiferentemente dos termos classificação e categorização de textos, embora na literatura corrente se faça uma diferenciação entre ambas. Essa modalidade de PLN, bastante complexa em sua construção, é de grande interesse comercial por parte de organizações privadas, órgãos do governo, editores de Internet etc. Na opinião dos autores, a integração desta tecnologia da informação representa um grande desafio, se as corporações quiserem alavancar melhor suas disponibilidades informacionais. Observam, contudo, uma certa indefinição na área, em relação às tarefas específicas a serem alcançadas. Freqüentemente, fala-se em categorização ou classificação, quando, o foco de interesse está realmente na indexação, na sumarização ou mesmo na extração automática de informações. Após descrever as diversas modalidades de categorização textual, dentre as quais, o direcionamento seletivo; o agrupamento; a suplementação de informações e as anotações (por exemplo, na área jurídica), os autores enfatizam a necessidade (óbvia) do correto entendimento prévio sobre os dados a serem categorizados, condição que muitos produtores deste tipo de *software* ainda não estariam considerando satisfatoriamente. Dentre as propriedades inerentes aos dados a serem levadas em consideração previamente figuram: o número total de categorias atribuíveis ao documento (granularidade); a quantidade de traços classificatórios (dimensionamento); se o mesmo documento pode pertencer a uma ou mais de uma categoria (exclusividade) e se o documento pode conter um só ou vários tópicos (topicalidade). Para a classificação automática de textos necessita-se também de taxonomias das respectivas áreas, por exemplo, levantamentos terminológicos de atividades industriais e comerciais, seguros, recursos humanos, saúde etc. As pesquisas de classificação automática têm-se concentrado, sobretudo, em noticiários de imprensa, e-mails e composição de conteúdos a partir de páginas da Web. Até certo ponto, de acordo com os autores, os problemas técnicos da categorização textual coincidem com aqueles da extração informacional, a qual, entretanto, já conta com cerca de 40 anos de experiência, enquanto a categorização textual apenas há uma década vem recebendo mais atenção. O capítulo conclui com a avaliação de alguns sistemas de categorização textual automática, estimulados pelos resultados das Conferências TREC.

O quinto e último capítulo, cautelosamente intitulado “Em direção à mineração de textos” (3), assunto que os autores optaram por não incluir no subtítulo do livro, trata de uma área de pesquisa mais recente do que a já estabelecida mineração de dados *(data mining)*. Direcionada à descoberta automatizada de conhecimento a partir de textos, a mineração textual difere basicamente tanto da recuperação como da extração informacional, tratadas nos capítulos anteriores. Nestas duas, a rigor, não são feitas “descobertas” nem são produzidas

informações novas. Numa base de noticiário de imprensa, por exemplo, "descobrir" novos tópicos informacionais, gerados automaticamente a partir de notícias contidas nos textos, poderá, sim, ser considerado o resultado de uma mineração de texto. Já numa classificação ou categorização textual automática de fatos ou eventos, conforme uma hierarquia conceitual previamente estabelecida, não se detectam padrões informacionais novos ou inéditos, como na mineração textual. Nesta modalidade de PLN, são duas as linhas de pesquisa referidas pelos autores (p. 173): a primeira tem como objeto a geração automática de metadados a partir dos documentos de uma coleção, em especial, referentes a nomes próprios e resumos; a segunda visa o processamento informacional além dos limites de documentos singulares, resultando o agrupamento de documentos *(document clustering)*, a sumarização transdocumentária *(cross-document summarization)* isto é, o resumo produzido com base em vários documentos; e a identificação de novos tópicos *(detection of new topics)* a partir de uma coleção. Incluiu-se este tipo de sumarização no presente capítulo porque, ao envolver a síntese de informações que não estejam presentes em um só documento, requer-se, para sua execução, um *rationale* semelhante ao da mineração textual. Como os demais capítulos do livro, também este finaliza com uma breve avaliação dos principais problemas abordados.

Ao concluir, os autores arriscam algumas previsões em relação à mineração textual e ao PLN em geral, conscientes de que uma pretensão desse tipo é quase sempre falha (p. 22). Quanto às tecnologias da informação por eles consideradas essenciais como a recuperação, extração e categorização informacional, constituem ferramentas úteis, desde que sejam entendidas suas atuais limitações. Contraproducentes seriam expectativas em relação a *softwares* que pretensamente "compreendem" significados ou produzem classificações "baseadas em conceitos" etc., freqüentemente alardeadas pelo *marketing*. Os aplicativos, sem tornar-se cada vez mais "inteligentes", na acepção de ficcionistas científicos, serão ferramentas úteis, seja para encontrar fontes relevantes e extrair-lhes as informações, seja para classificar os resultados dessa busca adequadamente, conforme as necessidades do usuário. Outros recursos, cada vez mais aperfeiçoadas, deverão auxiliar na navegação pela Internet, em intranets corporativas e em bibliotecas digitais. No entanto, a relação homem-máquina mais promissora, na visão dos autores, será aquela em que um sistema altamente sofisticado, capaz de processar a linguagem humana, servisse de complemento à nossa inteligência (4).

É lícito afirmar que o livro de Jackson e Moulinier constitui uma excelente introdução, embora pressupondo conhecimentos prévios, para as tecnologias de informação aqui tematizadas, que se baseiam no processamento da linguagem natural. A organização do conteúdo, cuidadosamente documentado, de estilo fluente e sóbrio, permite vários enfoques

de leitura, tornando a obra recomendável não só na área acadêmica, como também na prática profissional de engenheiros e desenvolvedores de software. Seu conteúdo é igualmente informativo para administradores, executivos e consultores, envolvidos com a informatização de empresas, numa época em que o agressivo *marketing* de *software* de tecnologias da informação leva organizações privadas e governos a fazerem investimentos de monta, freqüentemente sem propiciar o retorno almejado.

Notas:

(1) Os anais foram publicados sob o título de Advances in Natural Language Processing. Third International Conference, Faro, Portugal, June23-26, 2002. Ed. Nuno Mamede. Heidelberg; New York: Springer, 2002

(2) Em 2003, no programa Text Retrieval Conference (TREC), supervisionado pelo Governo dos EUA, além de representantes da indústria e de universidades, participaram 93 grupos de pesquisa , oriundos de 22 países http://trec@nist.gov (Acesso em 28 de julho de 2004)

(3) A mineração de texto na acepção dos autores, corresponde à descoberta de conhecimento *(knowledge discovery),* tal como definido em F.W.LANCASTER & A. WARNER, Intelligent technologies in library and information service applications. Medford, N.J.: information Today, 2001, p. 74

(4) Alguns dos trabalhos apresentados em um evento internacional recente, cujos anais foram publicados no Brasil, confirmam essa visão: Ver 8th ICCC – International Conference on Electronic Publishing. Brasília, DF, 23-26 June, 2004. Ed. By J. Engelen et al. Brasília: UnB/FALE, 2004 (346p.)

Ulf Gregor Baranow, Doutor
Professor-Adjunto IV do Departamento de Ciência e Gestão da Informação Universidade Federal do Paraná

Originais recebidos em 31/08/2004