PROCESSAMENTO NATURAL DA LINGUAGEM E RECUPERAÇÃO DA INFORMAÇÃO: uma caracterização dos softwares extratores de termos em sistemas automatizados

Juliana Rabelo do Carmo[1]
Cenidalva Miranda de Sousa Teixeira[2]
Valdirene Pereira da Conceição[3]

Resumo

Apresenta o Processamento da Linguagem Natural (PLN) como um dos eixos de estudo da área da Computação, em especial a IA, que centraliza-se em desenvolver métodos para que os computadores realizem tarefas de modo a simular a inteligência humana na resolução de problemas. As motivações e objetivos da pesquisa são originadas de ordem profissional, na Ciência da Informação, no sentido de identificar ferramentas que facilitem a recuperação e representação da informação, para identificação dos itens lexicais recorrentes em grandes volumes de textos. O objetivo da pesquisa consiste em analisar o cenário prático da recuperação da informação, visando à sistematização e organização de ferramentas de gestão terminológica, apontado para a interpretação correta dos termos tratados. Emprega a metodologia da pesquisa bibliográfica e documental sobre as temáticas das áreas de Ciência da Informação, Linguística e Computação, e por este motivo, assume o caráter interdisciplinar, ao aplicar estes campos no modelo de pesquisa em PLN. Apresenta o histórico dos estudos sobre PLN e seus resultados em ordem cronológica, bem como suas principais áreas de aplicação. Explica os níveis e limitações do PLN tais como interpretação e geração, onde o computador adquire a capacidade de traduzir a compreensão do sistema para a linguagem natural por meio de estruturações semânticas pré-determinadas, no caso dos resumos e palavras-chave. Caracteriza três categorias de softwares extratores, suas funcionalidades, e exemplos de sistemas baseados em conhecimento estatístico; sistemas baseados em conhecimento lingüístico e; sistemas híbridos. Conclui indicando que o PLN não é um modelo de recuperação da informação, e sim um método de interação que pode ser efetivado em sistemas de informação (ou banco de dados específicos) visando interpretar de forma mais precisa e possível a linguagem dos usuários, focando o texto, uma vez que as expressões utilizadas para busca da informação são constituintes dos objetos lingüísticos.

Palavras-chave: Processamento da Linguagem Natural. Recuperação da informação. Softwares extratores.

Resumen

Presenta el Procesamiento del Lenguaje Natural (PLN) como una de las áreas del ordenador de ejes de estudio, especialmente IA, que se centra en el desarrollo de métodos para los ordenadores para realizar tareas con el fin de simular la inteligencia humana en la solución de problemas. Las motivaciones y objetivos de la investigación se obtienen de asociación profesional, en la Ciencia de la Información, para identificar herramientas para facilitar la recuperación y representación de la información, para identificar los elementos léxicos recurrentes en grandes volúmenes de textos. El objetivo de la investigación es examinar la situación práctica de recuperación de información, orientada a la sistematización y organización de herramientas de gestión de terminología, se refirió a la correcta interpretación de los tratados términos. Emplea la metodología de la investigación bibliográfica y documental sobre las áreas temáticas de Ciencias de la Información, Lingüística y Ciencias de la Computación, y por esta razón, tiene el carácter interdisciplinario, para aplicar estos

[1] Bibliotecária do Instituto Florence de Ensino Superior, São Luís, Maranhão – Brasil.
[2] Professora Associada II do Curso de Biblioteconomia da Universidade Federal do Maranhão (UFMA).
[3] Professora Adjunta do Curso de Biblioteconomia da Universidade Federal do Maranhão (UFMA).

campos en el modelo de investigación en PLN. Muestra la historia de los estudios sobre el PLN y resultados, en orden cronológico, así como sus principales áreas de aplicación. Explica los niveles y limitaciones de PLN tales como la interpretación y generación, donde el equipo adquiere la capacidad de traducir el conocimiento del sistema de lenguaje natural a través de la estructuración semántica predeterminada en el caso de los resúmenes y las palabras clave. Cuenta con tres categorías de software extractor, sus características, y los ejemplos de los sistemas basados en el conocimiento estadístico; los sistemas basados en el conocimiento y lingüística; sistemas híbridos. Concluye que indica que el PLN no es un modelo de recuperación de información, sino un método de interacción que puede efectuarse en los sistemas de información (o base de datos específica) con el fin de interpretar y más exacta posible el idioma de los usuarios, centrándose en el texto, ya que las palabras clave utilizadas para buscar la información de los objetos de lenguaje son constituyentes.

Palabras-clave: Procesamiento del Lenguaje Natural. Recuperación de la información. Extractor de software.

1 INTRODUÇÃO

A década de 40 foi significativa para o desenvolvimento dos primeiros computadores, inicialmente utilizados para fins científicos e comerciais de comunicação e armazenamento de dados, haja vista os acontecimentos políticos e militares da época, tendo as suas funcionalidades evoluídas com o passar do tempo. (BRIGGS; BURKE, 2004).

Dentre os vários eixos de estudo da área de Computação, a IA centraliza-se em desenvolver métodos para que os computadores realizem tarefas de modo a simular a inteligência humana na resolução de problemas. A comunicação e o uso da linguagem em sistemas originaram a necessidade da tradução da linguagem humana para a linguagem de máquinas, utilizada pelos computadores, constituindo assim uma de suas bases de estudo.

As motivações e objetivos da pesquisa são originadas de ordem profissional, na Ciência da Informação, no sentido de identificar ferramentas que facilitem a recuperação e representação da informação, para identificação dos itens lexicais recorrentes em grandes volumes de textos.

O objetivo da pesquisa consiste em analisar o cenário prático da recuperação da informação, visando à sistematização e organização de ferramentas de gestão terminológica, apontado para a interpretação correta dos termos tratados. Emprega a metodologia da pesquisa bibliográfica e documental sobre as temáticas das áreas de Ciência da Informação, Linguística e Computação, e por este motivo, assume o caráter interdisciplinar, ao aplicar estes campos no modelo de pesquisa em PLN.

2 PROCESSAMENTO DA LINGUAGEM NATURAL: histórico e caracterizações

O PLN tem sido estudado pela área da CI na perspectiva teórica, em especial no campo da Indexação e Recuperação da Informação, por entender que os softwares baseados neste modelo propiciam a extração de termos com maior precisão semântica para recuperação da informação em sistemas de busca automatizados.

Analisando o PLN, Bobrow *et al.* (1967, p. 161) percebeu que, inicialmente, as preocupações eram direcionadas para o processamento analítico e não estatístico das linguagens naturais, excluindo assim a maioria dos trabalhos em indexação automática, sumarização, análise de conteúdo e de estilo. Desse modo, os modelos de PLN desenvolvidos tinham seus estudos voltados para a geração de textos e, conseqüentemente, o seu alcance visava contemplar gramáticas ou estudos de línguas específicas em bases de dados, por outro lado, a freqüência do uso dos termos não era considerada nesta etapa de estudo.

Somente na década de 80, a estruturação do PLN, com abrangência dos aportes computacionais – em especial, no que diz respeito ao uso de softwares que possibilitaram avanços como: analisadores (*parsers*) de linguagem, representação de significado por computador.

Chowdury (2003) corrobora ao considerar que PLN é “[...] uma área de pesquisa e de aplicação que explora como os computadores podem ser usados para processar e manipular texto ou discurso em linguagem natural para fazer coisas úteis.”. Em suma, a aplicação do PLN refere-se às áreas de: acesso a banco de dados; recuperação da informação; extração da informação; tradução automática e geração de resumos.

2.1 Níveis e limitações de PLN

O PLN subdivide-se em níveis de análise e/ou estudo que compreendem: a interpretação, onde são desenvolvidas questões relativas ao estudo da língua de modo que as palavras se tornem compreensíveis pelo computador e, conseqüentemente, o armazenamento para que ocorra a utilização destas palavras em sistemas, tomando como exemplo os tradutores (ou *chatterbots*); e de geração, que ocorre de forma inversa, a partir da inclusão de termos ou expressões, o computador adquire a capacidade de traduzir a compreensão do sistema para a linguagem natural por meio de estruturações semânticas pré-determinadas, no caso dos resumos e palavras-chave.

Tais estruturações fundamentam a arquitetura do PLN, apresentada por Nunes *et al.* (1999) e mostram que o banco de palavras, representado pelo Léxico, é acessado pelos analisadores Léxico, Sintático e Semântico, enquanto a Gramática serve ao analisador semântico para autenticar as palavras ou frases. Nesta perspectiva, o PLN enquanto Sistema baseado no Conhecimento utiliza-se de cinco alicerces: gramática, léxico e o modelo de discurso, ou seja, as informações sobre a língua; modelo de domínio, a ser aplicado; e modelo do usuário que utiliza o sistema (NUNES *et al*., 1999).

O nível morfológico consiste na definição da estrutura de palavras, bem como a significação e função de cada palavra na frase (adjetivo, substantivo, verbo, etc.); nível sintático, por meio da análise da construção gramatical, suas relações entre unidades linguísticas e sua colocação (sujeito, predicado verbal, etc.); nível semântico, onde as palavras são analisadas pelo seu significado, a partir da análise sintática; nível do discurso, compreensão do significado da palavra a partir do contexto em que ele está inserido; nível pragmático, onde ocorre a compreensão do conteúdo da frase ou texto, a partir da determinação de sua tipologia (pergunta, afirmação) (NUNES *et al.*, 1999).

Constituinte da principal dificuldade do PLN, a ambigüidade, ou seja, a pluralidade de sentidos de uma palavra tem sido uma das motivações para o aprimoramento dos modelos de aplicação do PLN, pois exige uma identificação das unidades gramaticais. Nesse sentido, as soluções para esta problemática da ambigüidade estão indicadas no contexto de uso dos termos para assim apreender a sua significação. Com base nisto, revela-se a necessidade da análise linguística em diferentes níveis nas bases de conhecimento, por meio de abordagens que podem ser aliadas ao PLN, tomando como referenciais teóricos metodológicos aspectos morfossintáticos, semânticos e lexicais.

Em contrapartida, dentre as vantagens do uso do PLN estão: a eliminação da necessidade de adaptação a formas inusitadas de interação, cuja construção gramatical costuma ser de difícil aprendizado e domínio, a exemplo das linguagens de consulta de bancos de dados (NUNES, 2007, apud NANTES, 2008, p. 26); é possível ainda, o entendimento de consulta com erros (termos digitados erroneamente) e incompletas, buscando por palavras próximas e pelo contexto da conversação (SILVA; LIMA, 2007, p. 2). Para tanto, basta que o usuário tenha um

conhecimento básico da área - e ainda, assunto ou domínio -, da especialidade da base de dados.

3 SOFTWARES EXTRATORES PARA PROCESSAMENTO DE *CORPUS*

Entende-se por softwares estatísticos, aqueles que empregam os dados de frequências de ocorrências de elementos lexicais, para assim extrair os termos que representam o documento em questão. Citamos na Tabela 1 os exemplos deste tipo, de natureza gratuita:

Tabela 1: Exemplos de softwares extratores estatísticos

SOFTWARES	CARACTERÍSTICAS/FUNCIONALIDADES
Pacote NSP (N-gram Statistics Pack-age)	Realiza a identificação e extração de termos (n-gramas), ou seja, de sequências de caracteres de comprimento, que podem ser unigramas, bigramas, trigramas e tetragramas. Com abordagem puramente estatística, para a sua execução é necessário do software Perl instalado, e por não possuir interface gráfica, funciona via linha de código.
Corpógrafo	Dentre suas funções, pode-se destacar: pesquisar nos corpora utilizando expressões regulares, criar listas de N-gramas, obter listas de candidatos a termos, criar novas classificações para domínios específicos, relações semânticas e conectores de discurso, visualizar as ocorrências de termos nos corpora, visualizar redes lexicais constantes numa base de dados. (LINGUATECA, 2014).
ZExtractor	O ZExtractor possibilita o ajuste parâmetros estatísticos; extração precisa de n-gramas; interface gráfica na qual é possível o usuário definir um número mínimo de ocorrências para que uma palavra seja candidata a termo; e o estabelecimento dos itens que devem ser excluídos, ou seja, os *stoplists*.

Fonte: adaptado de (TEIXEIRA, 2010)

Os softwares indicados da Tabela 3 apresentam abordagem estatística com base na extração de n-gram, ou seja, de acordo com a extensão de unidades lexicais. Pazienza *et. al*. (2005) explica que existem mais medidas estatísticas que podem denotar relações entre as que podem ser utilizadas na extração de termos, que representam a dimensão linguística dos termos extraídos. Outro método que visa aprimorar a extração de candidatos a termos tem como base a Linguística e suas propriedades de processamento.

Os softwares linguísticos possuem a funcionalidade de identificar informações linguísticas baseando-se em conjuntos de anotações linguísticas como análise morfológica, morfossintática, sintática, semântica e pragmática. Alguns exemplos de softwares linguísticos são apresentados na Tabela 2:

Tabela 2: Exemplos de softwares extratores linguísticos

SOFTWARES	CARACTERÍSTICAS/FUNCIONALIDADES
WebCorp	Consiste em um conjunto de ferramentas que permitem acesso a Web como um recurso linguístico, ou seja, realiza a extração de vários aspectos sobre línguas como se a Web fosse um *corpus*. Possui como público alvo linguistas, lexicógrafos, editores, jornalistas, pesquisadores, professores de língua que estudam o uso da língua, seus neologismos, entre outros.
Unitex	O software *Unitex* consiste em um conjunto de programas para processamento de *corpus* linguísticos com interface Java, que permite que a ferramenta não atrapalhe o desempenho de outras plataformas durante o processamento de *corpus*. Os principais recursos linguísticos do Unitex são: dicionários (para serem utilizados pela máquina e não para humanos), e tabelas do léxico-gramática, que consistem em matrizes binárias que mostram as propriedades de algumas palavras.
GATE (General Architecture for Text Engineering)	O software foi desenvolvido em código aberto (livre), baseado em Java, com a finalidade de solucionar as problemáticas que envolvem a análise e processamento de texto, como extração de informações por meio de construção de taxonomias via menus, etiquetagens morfossintáticas e anotações semânticas e tratamento de co-referência ou anáforas.
LácioWeb	Possui ferramentas como contador de frequência padrão; contador de frequência por palavra; concordanciador para *corpus* sem anotação; etiquetadores morfossintático; editor de cabeçalhos, entre outros. Em suma, o intuito do Lácio-Web é de representar corporas devidamente compilados, catalogados e codificados por um padrão que permita o intercâmbio, navegação e análise.
SYNTEX	Ferramenta chamada de analisador de corpus para extração de termos em corpus de língua francesa, que utiliza textos anotados por um *parser* para construção de um léxico específico do domínio e paralelamente realiza a análise sintática. A extração de termos é feita por meio de sintagmas nominais, levando em consideração as categorias morfossintáticas e as principais relações sintáticas como sujeito, objeto direto e complemento proposicional.

Fonte: adaptado de (ALUÍSIO; ALMEIDA, 2006; LÁCIO-WEB, 2014; LOPES; VIEIRA, 2010)

Pazienza *et. al*. (2005) indica que a abordagem linguística na extração de termos deve: analisar o corpus de um domínio e identificar a *Parts of Speech* (PoS) – que são as categorias sintáticas ou gramaticais -; identificar e extrair termos a candidatos conforme as regras linguísticas criadas; preservar os significados de acordo com o termo original e; implementar filtros linguísticos para refinar a terminologia, fatos estes que requerem um tempo maior para seu desenvolvimento, para que o refinamento seja mais preciso. E por fim, a integração das funções estatísticas e linguísticas para extração de termos são características dos softwares híbridos, que combinam os modelos de frequência de ocorrências, com a base

linguística, e por este motivo, tende a melhorar os resultados devido o equilíbrio entre cada abordagem.

Pazienza *et. al.* (2005) apresenta que esse tipo de abordagem tende a alcançar resultados mais corretos se comparado às abordagens puramente estatística ou puramente linguísticas. Deste segmento, destacamos de forma breve, um software híbrido que é voltado para construção de ontologias:

- **OntoGen:** é uma ferramenta para extração automática de candidatos de termos, que identifica os documentos que correspondem ao tema e a seleção pode ser refinada pelo computador do usuário, e realiza a determinação de hierarquia de conceitos. (LOPES; VIEIRA, 2010, p. 193).

A partir destas considerações iniciais, vale ressaltar que o PLN não é um modelo de recuperação da informação, e sim um método de interação que pode ser efetivado em sistemas de informação (ou bancos de dados específicos) visando interpretar de forma mais precisa possível a linguagem dos usuários, focando o texto, uma vez que as expressões utilizadas para busca da informação são constituintes dos objetos linguísticos.

4 CONCLUSÃO

Com relação à análise comparativa entre os processos manual e automático de extração de termos, a convergência entre estes dois métodos consiste na subjetividade humana para seleção e correção dos termos encontrados, porém, vale ressaltar que a intersecção entre as categorias obtidas por Indexação manual e as categorias geradas por extração automática alcançaram índices de frequências diferentes durante o processo. Isso significa que, nem todos os termos elencados no método manual foram extraídos automaticamente.

A principal problemática encontrada durante o processo de extração automática de candidatos a termos consiste na disponibilidade dos softwares (híbridos, linguísticos ou estatísticos) de forma gratuita, haja vista que tais softwares citados durante o referencial teórico encontram-se em teste nos Programas de Pós-Graduação voltados para a Computação/Inteligência Artificial e Linguística Computacional e, por este motivo, ainda não foram disponibilizados para a comunidade acadêmica.

Vale retomar algumas considerações a respeito da extração automática, e suas abordagens observadas no desenvolvimento da pesquisa: na Terminologia, a extração automática corresponde à aquisição de um produto terminológico que

representa os léxicos, a exemplo de dicionários, índices ou glossários; enquanto a Computação a entende como abordagem automática de reconhecimento e extração de termos de uma especialidade, geralmente realizada por meio das ferramentas de PLN.

O PLN mostra-se como uma ferramenta eficaz para processamento de grandes volumes de dados, com muito a contribuir no que diz respeito à redução do tempo de desempenho de tarefas de mineração de textos e ao possibilitar a identificação dos termos mais utilizados para representação de um domínio. Apesar destas contribuições, destaca-se que a intervenção humana ainda é necessária para a limpeza dos materiais obtidos e para a validação dos resultados.

REFERÊNCIAS

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus: lições aprendidas na compilação de vários corpora para pesquisa lingüística. **Calidoscópio** (UNISINOS), vol. 4, n. 3, p. 155-177, set./dez. 2006. Disponível em: <http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol 4
n3/art04_aluisio.pdf >. Acesso em 25 mar. 2014.

BOBROW, D. G.; FRASER, J. B.; QUILLIAN, M. R. Automated Language Processing, **Annual Review of Information Science and Technology**, v. 2, p. 161-186, 1967.

BRIGGS, Asa; BURKE, Peter. Uma história social da mídia: de Gutemberg à Internet. Rio de Janeiro: Jorge Zahar Ed., 2004.

CHOWDHURY, Gobinda C. Natural Language Processing, Annual Review of Information Science and Technology, v. 37, p. 51-89, 2003.

LINGUATECA. 2014. Disponível em: <http://www.linguateca.pt/>. Acesso em 14 abr. 2014.

LÁCIO-WEB. 2014. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/>. Acesso em: 05. mar. 2014.

LOPES, Lucelene; VIEIRA, Renata. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. In: PERNA, Cristina Lopes; DELGADO, Heloísa Koch; FINATTO, Maria José (Orgs.). **Linguagens especializadas em corpora:** modos de dizer e interfaces de pesquisa [recurso eletrônico]. Porto Alegre: EDIPUCRS, 2010.

NANTES, L. M. **Desenvolvimento de um sistema baseado em linguagem natural para consultas em banco de dados na Web.** 63 p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade do Oeste Paulista,

Presidente Prudente, 2008. Disponível em: <http://fipp.unoeste.br/~chico/FIPP/projetos/projeto2008/Monografia_Nantes_2008.pdf>. Acesso em: 20 ago. 2013.

NUNES, M. G. V.; DIAS-DA-SILVA, B. C.; RINO, L. H. M.; OLIVEIRA JR., O. N.; MARTINS, R. T.; MONTILHA, G. **Introdução ao processamento das línguas naturais**. Notas Didáticas do ICMC, n. 38. São Carlos/SP, 1999. p. 91.

PAZIENZA, M. T. *et. al*. *Terminology extraction: na analysis of linguistic and statistical approaches*. ***Studies in fuzziness and soft computing***, v. 185, p. 255-280, 2005.

RODRIGUES FILHO, Ilson Wilmar. Processamento de linguagem natural. 2004. Disponível em: <http://www.inf.ufsc.br/~ilson/slides.ppt>. Acesso em: 15 mar. 2014.

SILVA, Renato Rocha; LIMA, Sérgio Muinhos Barroso. Consultas em bancos de dados utilizando linguagem natural. **Revista Eletrônica da Faculdade Metodista Granbery**, Juiz de Fora, v. 7, n. 2, ago/dez. 2007. Disponível em: <http://re.granbery.edu.br/artigos/MjQ0.pdf>. Acesso em: 30 ago. 2013.

TEIXEIRA, Rosana de Barros Silva e. **Termos de (Onco)mastologia:** uma abordagem mediada por corpus. 2010. 392 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – Pontíficia Universidade Católica de São Paulo, São Paulo, 2010.

VIEIRA, R.; LIMA, V. L. S. Linguística computacional: princípios e aplicações. In: IX Escola de Informática da SBC-Sul. Luciana Nedel (Ed.). Passo Fundo, Maringá, São José. SBC-Sul, 2001.

WARNER, A. J. Natural Language Processing, **Annual Review of Information Science and Technology**, v. 22, p. 79-108, 1987.