

Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação

Marco Gonzalez, Vera Lúcia Strube de Lima

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16, PPGCC
90619-900 Porto Alegre, Brazil
{gonzalez,vera}@inf.pucrs.br

Resumo. Sintagma nominal (SN) é a parte de uma sentença com comportamento de sujeito ou objeto. Tratado como uma expressão referencial, o SN, neste artigo, tem seus componentes analisados. A função semântica de cada um é utilizada na construção de uma estrutura hierárquica temática para representar o SN. Neste formato, o SN pode ser tratado mais adequadamente em sistemas de recuperação de informação (RI), como termo de índices e de consulta. Com essa abordagem e com o uso de thesaurus, são obtidos benefícios na RI: (a) formulação de consulta em linguagem natural, (b) uso do mesmo procedimento para identificar temas na indexação e na consulta, e (c) aprimoramento do cálculo de similaridade na busca de documentos.

Abstract. Noun phrase (NP) is essentially the part of a sentence with subject or object behavior. NP had been viewed as a referring expression and in this paper its components are analyzed. The semantic function of each one is considered for constructing a thematic hierarchical structure for NP representation. This is a NP format that can be tractable by information retrieval (IR) systems as index and query term. With this approach and with thesaurus utilization, benefits are obtained in IR: (a) query in natural language, (b) use of the same process to identify themes on indexing and query, and (c) improvement on the similarity evaluation in the documents search.

1. Introdução

O sintagma nominal (SN), quando precedido de artigo definido, pode ser classificado como uma expressão referencial definida [Searle1981], mas ele também é usado como exemplo de referências indefinidas ou genéricas [Lyons1977]. Numa visão mais ampla, acredita-se que toda ocorrência de um SN seja uma expressão referencial [Thrane1980], sustentando-se sua aplicação para referenciar objetos, pessoas, ocorrências, etc. Assim, os SNs constituem uma classe de forma gramatical com comportamento sintático de sujeito, de objeto direto e, se precedidos de preposição, com comportamento de adjunto adnominal e de objeto indireto [Perini1986].

Em sistemas de recuperação de informação (RI), enquanto os termos de índice fazem referência aos documentos, os termos da consulta fazem referência àquilo que o usuário tenta encontrar nestes documentos. Assim, procuramos salientar, aqui, a importância de se examinar com detalhe o uso do SN como formador de termos de consulta e de índice de documentos. Sabe-se, por outro lado, que o uso de termos de índice baseados em expressões tem crescido muito entre os sistemas de RI [Evans1996]. Nesse sentido, utilizando as características do SN, propomos uma forma de representação de sentenças em linguagem natural, num formato que denominamos de estrutura hierárquica temática (EHT). O SN em EHT é adequado, como veremos, como forma lógica para consultas e índices de documentos na RI.

Na seção 2 deste artigo, são analisados os componentes do sintagma nominal e sua formação; na seção 3 é apresentada a estrutura hierárquica temática juntamente com o cálculo de similaridade que a mesma viabiliza; na seção 4 são discutidos alguns exemplos de aplicação da teoria, incluindo problemas e soluções; na seção 5 é dada notícia de ferramentas de apoio para esta proposta, desenvolvidas em forma de protótipo, e são tecidas considerações finais.

2. Componentes do sintagma nominal

Por apresentarem as características de expressões referenciais, é útil examinar as funções semânticas dos componentes de um SN, definidas por Liberato [Liberato1997] como sendo: classificador (CLA), subclassificador (SUB), qualificador (QUAL), recortador (REC), quantificador (QUAN) e partitivo (PART). A seguir formalizamos a definição de tais componentes.

Definição 1. Seja SN um sintagma nominal e E uma expressão componente de SN, tal que $E \subset \text{SN}$. Então:

- E é CLA se E delimita a classe mais ampla em que o referente é enquadrado numa determinada descrição.
- E é SUB se CLA engloba subclasses possíveis e E delimita uma subclasse de CLA, tendo E, portanto, função restritiva.
- E é QUAL se E tem função explicativa, fornecendo características do referente que, no entanto, não são utilizadas como delimitação de SUB ou de CLA.
- E é REC se E indica se o referente é constituído da totalidade ou de parte dos elementos de CLA.
 - (a) REC é universal se E engloba todos os elementos de CLA.
 - (b) REC é parcial se E não considera todos os elementos de CLA.
- E é QUAN se E indica a quantidade de elementos que constituem CLA.
- E é PART se E faz um recorte parcial sobre CLA, sendo esta uma classe de entidades identificáveis.

Exemplos: agenda é CLA, minha é SUB e lotada é QUAL em minha agenda lotada; os é REC universal em os telhados de Viena; um é REC parcial em um canto do jardim; poucas é QUAN em as poucas pesquisas existentes; a maior parte de é PART em a maior parte dos exemplares recuperados.

2.1. Formação do Sintagma Nominal

Com as funções recém definidas, os referentes de um enunciado podem ser encontrados e este pode ser reconstruído na forma de um SN, buscando preencher tais funções. Se tentarmos encontrar os referentes do enunciado (1), a seguir, argüindo sobre como seria uma consulta adequada para recuperar um documento que ele representasse, teríamos (2.a), (2.b) ou (2.c). Nota-se que o principal referente em (2.a) é o agente do processo de construção, em (2.b) é o próprio processo, e em (2.c) temos o produto desse processo.

- (1) A Edifica S.A. construiu edifícios grandes em 1999
- (2.a) construtora Edifica S.A. de edifícios grandes em 1999
- (2.b) construção de edifícios grandes pela Edifica S.A. em 1999
- (2.c) edifícios grandes construídos pela Edifica S.A. em 1999

As seqüências de passos que constituem possíveis funções de obtenção dos SNs (2.a), (2.b) e (2.c), a partir de (1), são apresentadas a seguir.

Obtenção de SN₁. Entrada: enunciado (1). Saída: SN₁ = (2.a).

- (1º) encontrar o CLA através do substantivo correspondente ao agente do verbo construir (obtendo CLA = construtor);
- (2º) encontrar o SUB₁ através do sujeito em (1) (obtendo SUB₁ = Edifica S.A.);
- (3º) encontrar outros SUBs, através do objeto direto (obtendo SUB₂ = de edifícios grandes) e do adjunto adverbial (obtendo SUB₃ = em 1999).

Obtenção de SN₂. Entrada: enunciado (1). Saída: SN₂ = (2.b).

- (1º) encontrar o CLA, através do substantivo correspondente ao processo relativo ao verbo construir (obtendo CLA = construção);
- (2º) encontrar outros SUBs, através do objeto direto (obtendo SUB₁ = de edifícios grandes), do sujeito (obtendo SUB₂ = pela Edifica S.A.) e do adjunto adverbial (obtendo SUB₃ = em 1999).

Obtenção de SN₃. Entrada: enunciado (1). Saída: SN₃ = (2.c).

- (1º) encontrar o CLA através do SN que compõe o objeto direto em (1) (obtendo CLA = edifícios grandes);
- (2º) encontrar um SUB através do efeito do verbo construir, enfatizando um estado (obtendo SUB₁ = construído);
- (3º) encontrar SUBs relacionados ao verbo, através do sujeito (obtendo SUB₂ = pela Edifica S.A.) e do adjunto adverbial (obtendo SUB₃ = em 1999).

Assim, sentenças em linguagem natural podem ser transformadas em SNs se contarmos com um léxico, onde relações semânticas sejam explicitadas [Gonzalez2000], e adotarmos a definição 1, apresentada anteriormente. Antes, porém, temos de decidir qual procedimento de transformação é mais adequado, dependendo do assunto principal (TEMA) da sentença, seguindo as seguintes instruções:

- (i) Se TEMA = agente do verbo principal
 - (i.a) Se sujeito não predomina como uma classe principal
CLA = agente¹ correspondente ao verbo principal

¹ O conceito de agente, utilizado nestas instruções, corresponde ao substantivo com sufixação geralmente em dor ou nite, associado ao verbo constituindo papel agenteivo [Basílio1995]. Como exemplo, temos: jogar – jogador, despachar – despachante, etc.

- SUB₁ = SN extraído do sintagma verbal
 SUB₂ = sujeito
- (i.b) Se sujeito predomina como uma classe principal
 CLA = sujeito
 SUB₁ = agente correspondente ao verbo principal
 SUB₂ = SN extraído do sintagma verbal
- (ii) Se TEMA = ação correspondente ao verbo principal
 CLA = nominalização do verbo principal
 SUB₁ = sujeito
 SUB₂ = SN extraído do sintagma verbal
- (iii) Se TEMA = objeto da ação correspondente ao verbo principal
 CLA = objeto
 SUB₁ = estado associado à ação correspondente ao verbo principal
 SUB₂ = sujeito
 SUB₃ = SN extraído do sintagma verbal, exceto o CLA

Por exemplo, considerando a sentença original (3), conforme (i.b), tendo o agente (gerador) como TEMA e o predomínio do sujeito (receita) como classe principal, teríamos o SN (3.a). Conforme (ii), tendo a ação (gerar) como TEMA, teríamos o SN (3.b). Conforme (iii), tendo o objeto (lucro) da ação como tema, teríamos o SN (3.c)

- (3) A receita gera lucro.
 (3.a) Receita [CLA] geradora [SUB₁] de lucro [SUB₂];
 (3.b) Geração[CLA] de lucro[SUB₁] pela receita [SUB₂];
 (3.c) Lucro[CLA] gerado[SUB₁] pela receita[SUB₂].

E considerando (4), conforme (i.a), tendo o agente (jogador) como TEMA, sem o predomínio do sujeito (Pelé) como classe principal, teríamos o SN (4.a).

- (4) Pelé joga futebol.
 (4.a) Jogador[CLA] de futebol [SUB₁] Pelé[SUB₂].

A definição de um tema (ou dos temas) de um texto parece tomar dois caminhos alternativos preferenciais: (i) com a utilização de PLN ou (ii) com apoio de procedimentos estatísticos. O primeiro caso passa, por exemplo, por técnicas de sumarização de texto [Nagao1998], como uma atividade intermediária, para posterior extração dos temas tratados. No segundo caso, a análise de frequência das palavras no texto, apoiada na teoria estatística Zipf-Luhn [Frants1997, Kowalski1997, Baeza-Yates1999], pode ser usada para se chegar ao tema, identificando-se a frequência de ocorrência dos itens lexicais no texto. Utilizando-se tanto as características gramaticais e funções dos itens lexicais no contexto das sentenças, quanto a frequência dos mesmos no texto, teríamos a seguinte abordagem:

- (i) extrair os SNs do texto a partir de cada sentença;
 (ii) identificar os termos com função de classificador (CLA) e subclassificador (SUB); e
 (iii) para cada sentença (ou SN) identificar seu tema como sendo o CLA (preferencialmente) ou o SUB presentes na sentença e com maior frequência no texto.

3. Estrutura Hierárquica Temática

Utilizando-se principalmente os conceitos de CLA, SUB e QUAL, pode-se assumir uma estrutura hierárquica, esquematizada na figura 1, de acordo com as funções dos componentes do SN. A hierarquia adotada, que passaremos a denominar de estrutura

hierárquica temática² ou simplesmente EHT, leva em conta a importância da delimitação da referência concebida por Liberato.

Conforme a EHT esquematizada na figura 1, a função de CLA confere, ao componente, uma posição mais elevada na hierarquia. Os componentes com função de REC, QUAN e PART não interferem na hierarquia, mas podem, se utilizados, qualificar semanticamente a representação. Os SUBs recebem uma posição de segundo nível, pois delimitam uma subclasse de CLA. E os componentes QUAL, pela função explicativa, assumem posição menos importante hierarquicamente.

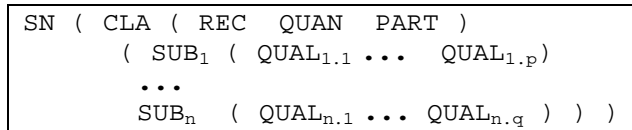


Figura 1. SN em EHT esquemática

Cada SUB, por sua vez, pode apresentar recursivamente seus próprios SUBs. Desse modo, é possível deduzir que termos de consulta e de índice podem ser similares em níveis hierárquicos diferentes.

3.1. Similaridade entre consulta e índice

A EHT, sendo utilizada para representar tanto a consulta como os documentos na RI, viabiliza o cálculo do grau de relevância dos documentos através da similaridade estrutural, que detalhamos na definição 2. Este critério pode associar-se àquele onde considera-se a similaridade direta entre os termos, independentemente da posição hierárquica.

Definição 2 Sejam as EHTs **C** e **D**, que representam, respectivamente, uma consulta e um documento, e sejam pf_i um par de termos que constitui um relacionamento “pai-filho”, e f_i um termo “filho”, nessas estruturas. E_{CD} é a similaridade estrutural entre uma consulta (representada por **C**) e um documento (representado por **D**), se:

$$E_{CD} = \frac{\sum_{i=1}^m \text{máx}(1, G_i)}{r}$$

onde:

r = número de pares do tipo pf_i de **C**;

m = número de pares do tipo pf_i de **C** encontrados em **D**;

$\text{máx}(1, G_i)$ = maior valor entre 1 e G_i ,

$$G_i = \frac{g(pf_i, C)}{g(pf_i, D)} = \text{grau de importância relativa de } pf_i \text{ de } C \text{ em } D;$$

$g(pf_i, C) = P_{t=1 \rightarrow u} N_{t,C}$ = grau de importância de pf_i em **C**, tal que:

$N_{t,C}$ = número de nós do nível t em **C**;

u = nível de f_i em **C**;

$g(pf_i, D) = P_{t=1 \rightarrow v} N_{t,D}$ = grau de importância de pf_i em **D**, tal que:

$N_{t,D}$ = número de nós do nível t em **D**;

v = nível de f_i em **D**.

•

² Esse tipo de estrutura não é uma novidade, já que as tabelas de conteúdos (ou sumários) e os índices remissivos a utilizam há bastante tempo. Abordando o problema de indexação para RI, Loukachevitch e seus co-autores [Loukachevitch1999] propõem, também, uma representação temática de cada documento como uma estrutura hierárquica de nodos conceituais.

Através deste cálculo, a similaridade estrutural será 1 (máxima), quando as estruturas apresentarem os mesmos pares “pai-filho” (pf_i) nos mesmos níveis; será zero, quando nenhum pf_i da consulta for encontrado no índice do documento. Os valores intermediários são obtidos da média dos graus de importância relativa (G_i) dos pf_i , de acordo com o número de nós “irmãos” e “antecessores” do nó “filho” de cada pf_i , nas duas árvores. Quanto mais subdividida a estrutura (ou seja, quanto maior o número de nós encontrados desde o nível 1 até o nível onde encontra-se o nó “filho” do par de termos analisado), menor o seu grau de importância. E quanto maior a média dos G_i s dos termos, maior será a similaridade estrutural entre as EHTs da consulta e do documento.

Se considerarmos, também, o cálculo da similaridade direta³ entre termos (definição 3), podemos utilizar a média entre os valores assim obtidos para chegar à similaridade geral (ou média) entre uma consulta e um documento, conforme a definição 4, a seguir.

Definição 3. T_{CD} é a similaridade de termos entre uma consulta C e um documento D se:

$$T_{CD} = NT_{CD} / NT_C$$

onde: NT_{CD} = número de termos de C em D ;
 NT_C = número de termos de C .

Definição 4. S_{CD} é a similaridade geral entre uma consulta C e um documento D se:

$$S_{CD} = (T_{CD} + E_{CD}) / 2$$

onde: T_{CD} = é a similaridade de termos entre C e D ;

E_{CD} = é a similaridade estrutural entre C e D .

4. Aplicando a teoria

4.1. Calculando a similaridade

Na figura 2 aparecem quatro árvores representando, de forma esquemática, da esquerda para a direita, uma consulta e três índices de documentos. As letras identificam os termos presentes em cada documento. De acordo com a definição 2, as similaridades estruturais entre a consulta C e os documentos $D1$, $D2$ e $D3$ são calculadas como:

$$E_{CD1} = (2/3 + 2/3) / 2 = 2/3$$

$$E_{CD2} = (2/(2 \times 3) + 2/(2 \times 3)) / 2 = 1/3$$

$$E_{CD3} = 0 / 2 = 0$$

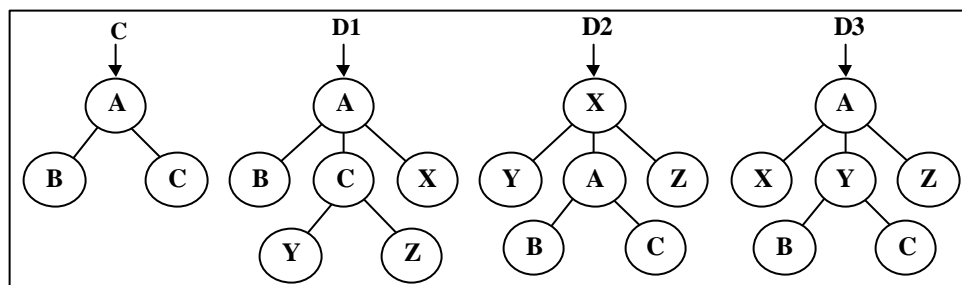


Figura 2. Consulta e índices de documentos em formato de árvore

³ Queremos classificar como cálculo de similaridade direta, aquele que é independente da posição hierárquica dos termos. Utilizamos na definição 3 um método simples apenas para efeitos didáticos. Não

A similaridade de termos T_{CD} será 1, pela definição 3, em qualquer um dos três casos, já que todos os termos A, B e C da consulta estão presentes nos índices dos documentos. De acordo com a definição 4, a similaridade geral será, respectivamente:

$$S_{CD1} = (1 + 2/3) / 2 = 5/6$$

$$S_{CD2} = (1 + 1/3) / 2 = 2/3$$

$$S_{CD3} = (1 + 0) / 2 = 1/2$$

4.2. Construindo a EHT

Retomando o exemplo da seção 2.1, e assumindo a prévia classificação de edifícios grandes como CLA = edifícios e SUB = grandes, a partir da obtenção de

SN1 = construtora Edifica S.A. de edifícios grandes em 1999,

SN2 = construção de edifícios grandes pela Edifica S.A. em 1999, e

SN3 = edifícios grandes construídos pela Edifica S.A. em 1999,

através das funções de obtenção do SN, poderiam ser geradas, respectivamente, as seguintes EHTs:

(5.a) EHT₁ (construtora (Edifica S.A.)
(de edifícios (grandes))
(em 1999))

(5.b) EHT₂ (construção (de edifícios (grandes))
(pela Edifica S.A.)
(em 1999))

(5.c) EHT₃ (edifícios (grandes)
(construídos (pela Edifica S.A.)
(em 1999)))

Vamos, agora, supor que o enunciado

(1) A Edifica S.A. construiu edifícios grandes em 1999

esteja contido num documento cujo contexto seja “construções do ano de 1999 na região norte”. Nesse caso, a representação do enunciado (1) deveria ser conseguida pela função de obtenção do SN que gera o CLA a partir do verbo, ou seja, que enfatiza o evento. Desse modo, teríamos a EHT₂, apresentada em (5.b), representando (1). Para analisar quais consultas teriam sucesso na recuperação de tal documento, vamos examinar as consultas em (6.a), (6.b) e (6.c), através de suas EHTs, na respectiva ordem, em (7.a), (7.b) e (7.c),.

(6.a) construção de edifícios

(6.b) construtora Edifica S.A.

(6.c) edifícios grandes

(7.a) EHT₄ (construção (de edifícios))

(7.b) EHT₅ (construtora (Edifica S.A.))

(7.c) EHT₆ (edifícios (grandes))

Calculando-se a similaridade, através das definições 2, 3 e 4, teríamos, em relação a EHT₂, respectivamente, para as consultas representadas por EHT₄, EHT₅ e EHT₆:

$$S_{42} = (1 + 1/3) / 2 = 2/3$$

$$S_{52} = (1/2 + 0) / 2 = 1/4$$

$$S_{62} = (1 + 1/3) / 2 = 2/3$$

Parece coerente o valor menor para a consulta (7.b), pois (1) enfatiza a construção e não a construtora. Mas, de qualquer forma, seria interessante detectar a relação semântica que existe entre construção e construtora, de forma a diminuir a diferença entre a consulta (7.b) e as outras. O modo como fazê-lo é discutido na seção que segue.

4.3. Ambigüidade e refinamento da consulta

Antes de analisarmos esse problema, cabe examinarmos os possíveis casos de consulta com certo grau de ambigüidade, como os seguintes:

- (8.a) construção da Edifica S.A.
- (8.b) edifícios grandes em 1999

Podemos entender que a intenção do usuário em (8.a) fosse, entre outras, construção da sede da Edifica S.A. ou construção de edifícios pela Edifica S.A. Já em (8.b) não conhecemos o evento associado, podendo ser construção, fiscalização, reforma, todos estes ou, ainda, outros. Visualizamos duas alternativas: (i) refinar (realimentar) a consulta, em interação com o usuário para reduzir a ambigüidade; ou (ii) assumir o risco e atribuir um grau de relevância menor ao documento que contém (1). Mas, o que fazer com consultas como (6.b) ou (9)?

- (9) prédios grandes

Em princípio, tais consultas teriam que gerar um grau de relevância para o documento que contém (1), de modo a não desprezá-lo totalmente.

4.4. Normalização Semântica dos termos da consulta

As EHTs para as consultas (6.b) e (9) – repetidas em (10.b) e (10.c), respectivamente – deveriam recuperar com algum grau de relevância o documento contendo (1) – representada pela EHT (10.a). Fundamentalmente, a normalização semântica, nesses casos, deveria levar o sistema de RI a “compreender” que os itens lexicais construtora e construção estão associados, e que edifícios são prédios.

- (10.a) EHT₂ (construção (de edifícios (grandes))
(pela Edifica S.A.)
(em 1999))
- (10.b) EHT₅ (construtora (Edifica S.A.))
- (10.c) EHT₇ (prédios (grandes))

No caso de (10.c), o sistema deve reconhecer que edifício e prédio relacionam-se por hiponímia, consultando uma estrutura hierárquica de classificação lexical como, por exemplo, um thesaurus com estruturação semântica que adote uma abordagem como a da teoria do léxico gerativo [Pustejovsky1995], ou uma base de dados como o WordNet [Fellbaum1998]. No caso de (10.b), o processamento semântico deve identificar a associação entre construtora (ou construtor) e construção. O mesmo deve ocorrer nos procedimentos para formação do SN, no exemplo desenvolvido neste artigo, onde são necessárias associações entre o verbo construir e um processo (construção), um estado (construído) e um produto (construção). É necessário que essas associações sejam deduzidas da representação de construir que, no léxico, deve fazer referência aos itens lexicais ora citados, constituindo, dessa forma, relações semânticas entre eles.

Assim, no processo de busca de documentos relevantes na RI, a pesquisa de Edifica S.A. em (10.b), entre os SUBs de construção de (10.a), fica justificada pela associação de construção com construtor (ou construtora). Mas, tal recurso pode penalizar a recuperação do documento com a diminuição do seu grau de relevância.

5. Considerações finais

Este artigo descreveu uma abordagem voltada à utilização de termos de consulta e de índices de documentos, na RI, constituídos por SNs em EHT. A seguir, apresentamos a situação da prototipação relacionada à idéia central do texto e tecemos algumas conclusões preliminares.

Na abordagem aqui descrita, são fundamentais tanto o reconhecimento dos SNs quanto suas estruturações em EHT. Nesse sentido, foi implementado o protótipo de um analisador sintático e funcional para reconhecer um SN e classificar seus componentes conforme a categoria gramatical e as funções apresentadas. Foi elaborada uma gramática para o SN e, em sua implementação, foi utilizado o gerador de analisador léxico e sintático Lex Yacc [Lenine1998]. O protótipo permite reescrever uma sentença etiquetada com classes gramaticais em um formato onde são identificadas as funções semânticas de cada item lexical, podendo futuramente ser incluído numa aplicação de RI, para compor a tarefa de indexação de documentos.

Também em fase de protótipo, foi construído um thesaurus com estruturação semântica projetada para identificar relacionamentos lexicais, levando em conta aspectos da teoria do léxico gerativo [Pustejovsky1995] e conceitos de modelagem de software orientada a objetos (MSOO) [Rumbaugh1991]. Na forma de descritores, foram introduzidas referências diretas aos itens lexicais relacionados ao termo descrito, tornando explícitas as relações semânticas nas representações lexicais.

Foi utilizado um corpus de teste constituído por resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. O protótipo desenvolvido utiliza formulário html, para entrada e visualização dos dados, e programação CGI em ambiente Unix, com linguagem C, para o gerenciamento da base de dados. Foram incluídas operações gerativas para obter um conjunto de termos relacionados a um dado termo original, possibilitando, por exemplo, expansão automática de consulta em sistemas de RI.

Neste artigo, pretendemos salientar a importância da análise do SN, como construtor de termos de consulta e de índice para RI. O conhecimento detalhado das funções dos componentes do SN é fundamental para sua utilização, assim como é imprescindível a presença de um thesaurus, para viabilizar a normalização semântica dos termos envolvidos.

O processamento da consulta, assim, a partir da necessidade de informação do usuário, permite a formulação desta última em linguagem natural. Na verdade, termos de consulta e de índice devem ser construídos por procedimentos semelhantes. A EHT, resultante de tal processamento, auxilia a representação semântica e os mecanismos de busca, dando subsídios ao cálculo da similaridade entre os termos e permitindo critérios precisos para a classificação dos documentos por relevância.

Bibliografia

- Baeza-Yates, R.; Ribeiro-Neto, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.
- Basílio, Margarida. O Fator Semântico na Flutuação Substantivo / Adjetivo em Português. In: HEYE, J. (Org.), Flores Verbais, Rio de Janeiro, 1995. 34ª edição, p.177-192.

- Evans, David A.; ZHAI, Chengxiang. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In: 34th Annual Meeting of the ACL, Univ. of California, Sta. Cruz, California, USA, junho 1996. Proceedings of the Conference, p.17-24.
- Fellbaum, Christiane. **Wordnet: An Electronic Lexical Database**. Cambridge: MIT Press, 1998. 423 p.
- Frants, V.; Shapiro, J.; Voiskunskii, V. **Automated Information Retrieval: Theory and Methods**. San Diego, CA: Academic Press, 1997. 365 p.
- Gonzalez, M. O Léxico Gerativo de Pustejovsky sob o Enfoque da Recuperação de Informações. Trabalho Individual I, PPGCC, Faculdade de Informática, PUCRS, maio 2000. 52 p.
- Kowalski, G. **Information Retrieval Systems: Theory and Implementation**. Kluwer Academic Publishers, 1997. 282 p.
- Lenine, J. R. **Lex Yacc**. Cambridge: O'Reilly, 1998. 364 p.
- Liberato, Yara G. A Estrutura do SN em Português: Uma Abordagem Cognitiva. Belo Horizonte: Tese de Doutorado em Letras, UFMG, 1997. 203 p.
- Loukachevitch, N. V.; Salli, A. D.; Dobrov, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. NLDB'99 – 4th Int. Conf. on Applications of Natural Language to Information Systems, 1999. OCG Schriftenreihe, Lecture Notes, v.129, p.203-208.
- Lyons, J. **Semantics**. Cambridge: Cambridge University Press, 1977. V. I e II.
- Nagao, K.; Kôiti, H. Automatic text Summarization Base on the Global Document annotation. Montreal, Quebec, Canadá: COLING, ACL'98 – 36th Annual Meeting of the Assoc. for Comp. Linguistics and 17th Intern. Conf. on Comp. Linguistics, 1998. Proceedings of the Conference, V. II, p. 917-921.
- Perini, Mário A. **Para uma Nova Gramática do Português**. São Paulo: Ed. Ática S.A., 3^a edição, 1986. 94 p.
- Pustejovsky, J. **The Generative Lexicon**. Cambridge: The MIT Press, 1995, 298 p.
- Rumbaugh, J; Blaha, M.; Premerlani, W.; Eddy, F.; Lorenzen, W. **Object-Oriented Modeling and Design**. Englewood Cliffs: Prentice-Hall, 1991. 500 p.
- Searle, John R. **Os Actos de Fala**. Coimbra: Livraria Almedina, 1981. 270 p.
- Thrane, Torben. **Referential-Semantic Analysis: Aspects of a Theory of Linguistic Reference**. London: Cambridge University Press, 1980. 256 p.