Recuperação de Informação e Processamento da Linguagem Natural

Marco Gonzalez, Vera L. S. de Lima

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16 - PPGCC
90619-900 Porto Alegre, Brasil

{gonzalez, vera} @inf.pucrs.br

Abstract. *Information retrieval (IR) systems essentially execute indexation, search, and classification of (textual) documents. Their purpose is to satisfy user information needs, expressed through queries. However, to find the relevant information is a difficult task. Natural language processing (NLP) is present in different levels of the many approaches that researchers have been trying to solve that problem. The linguistic knowledge can bring intelligent strategies for IR, mainly through syntactic processing and semantic interpretation. Although statistical methods or linguistic knowledge are used, NLP has still to surmount many difficulties in the area of IR but, of course, it has many benefits to offer.*

Resumo. *Sistemas de recuperação de informação (RI) tratam essencialmente de indexação, busca e classificação de documentos (textuais), com o objetivo de satisfazer necessidades de informação de seus usuários, expressas através de consultas. Encontrar a informação relevante é, entretanto, uma tarefa difícil. O processamento da linguagem natural (PLN) está presente em diferentes níveis nas diversas abordagens que os pesquisadores têm procurado para solucionar este problema. O conhecimento lingüístico pode, principalmente através de processamentos morfo-sintático e semântico, trazer estratégias inteligentes para a RI. Tanto através de métodos estatísticos quanto pela aplicação de conhecimento lingüístico, o PLN tem ainda muitos desafios a vencer na área de RI mas, por certo, tem muitos benefícios a oferecer.*

***Palavras-chave**: recuperação de informação, processamento da linguagem natural.*

1. Introdução

1.1. Motivação

O homem tem armazenado, catalogado e organizado a informação há aproximadamente 4000 anos, com o principal objetivo de recuperá-la para uso posterior. Atualmente, cresce de forma cada vez mais rápida a quantidade dos textos armazenados em formato digital, e a maioria deles é esquecida pois nenhum ser humano pode ler, entender e sintetizar toda esta informação. Isto tem incentivado os pesquisadores a explorar estratégias para tornar acessível ao usuário a informação relevante [Rijsbergen1979,

Sparck-Jones1997, Kowalski1997, Frantz1997, Baeza-Yates1999, Croft2000, Meadow2000].

A necessidade de informação (NI) é considerada uma das necessidades vitais do ser humano e quem a satisfaz "viabiliza sua adaptação às condições externas da existência" [Frantz1997]. Entre os métodos tradicionais utilizados para satisfazer a NI estão: a captura da informação a partir da natureza (por exemplo: pela dedução de fórmulas ou pela medição de eventos), e a obtenção da informação através de consultas a um conjunto de dados armazenados [Frantz1997].

Tais consultas constituem a entrada de um sistema de recuperação de informação (RI), que deve buscar, em uma coleção de documentos, aqueles que podem satisfazer à NI do usuário. Entretanto, é difícil encontrar a informação relevante, principalmente porque há muita informação (a maioria irrelevante) [Baeza-Yates1999]. Os pesquisadores têm procurado abordagens alternativas para solucionar este problema. Além da aplicação de métodos estatísticos, o processamento da linguagem natural (PLN), com motivação lingüística, é uma dessas alternativas [Sparck-Jones1997, Jacquemin2000].

1.2. Objetivos

Este curso tem, como objetivo geral, apresentar uma visão geral sobre a área da RI e relatar aspectos e experiências relacionadas ao PLN, nesta área. Neste sentido, tem, como objetivos específicos, os seguintes:

- relatar os fundamentos do PLN quanto aos processamentos morfo-sintático e semântico e à representação do conhecimento;
- discutir o problema da RI, apresentando conceitos, modelos, técnicas e recursos que têm sido adotados para resolvê-lo;
- discutir como tem sido utilizado o PLN, em níveis sintático e semântico, para beneficiar técnicas e recursos na resolução dos problemas da RI; e
- apresentar um estudo de caso para ilustrar o funcionamento de um sistema tradicional de RI.

1.3. Organização do texto

Nesta seção (Introdução) é explicada a motivação do mesmo, são definidos seus objetivos e a seqüência dos assuntos abordados.

Na seção 2 (Processamento da Linguagem Natural), é dada uma visão geral sobre o PLN, quanto à transformação das sentenças em linguagem natural nas correspondentes formas lógicas, através de processamentos morfo-sintático e semântico, discutindo representação do conhecimento e estratégias do PLN.

Na seção 3 (Ontologia e Thesauri), são discutidos os conhecimentos ontológicos e são apresentados conceitos, classificação, modos de construção e aplicações de thesauri, com aplicação em alguns sistemas de RI.

A seção 4 (Recuperação de Informação) é reservada para o problema da RI. São discutidos conceitos e classificações, é apresentado um breve histórico e são descritos modelos, componentes e etapas (indexação e busca) de execução dos sistemas de RI. Também são apresentadas métricas para avaliação dos mesmos.

Na seção 5 (RI e PLN), o PLN é discutidos no contexto da RI. São discutidas abordagens e relatadas experiências.

Na seção 6 (Estudo de Caso) é apresentado um estudo de caso sobre um sistema de RI, com o objetivo de examinar seus componentes e recursos.

Para finalizar, a seção 7 (Considerações Finais) tece comentários sobre os pontos abordados, discutindo dificuldades e possibilidades do PLN na RI.

2. Processamento da Linguagem Natural

2.1. Introdução

O processamento da linguagem natural (PLN) trata computacionalmente os diversos aspectos comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos. Em sentido bem amplo, podemos dizer que o PLN visa fazer o computador se comunicar em linguagem humana, nem sempre necessariamente em todos os níveis de entendimento e/ou geração de sons, palavras, sentenças e discursos. Estes níveis são:

- fonético e fonológico: do relacionamento das palavras com os sons que produzem;
- morfológico: da construção das palavras a partir unidades de significado primitivas e de como classificá-las em categorias morfológicas;
- sintático: do relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças;
- semântico: do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças; e
- pragmático: do uso de frases e sentenças em diferentes contextos, afetando o significado.

A representação do significado de uma sentença, independente de contexto, é obtida através de sua forma lógica [Allen1995, Franconi2001] (ver Figura 1). A forma lógica codifica os possíveis sentidos de cada palavra e identifica os relacionamentos semânticos entre palavras e frases. Uma vez que os relacionamentos semânticos são determinados, alguns sentidos para as palavras tornam-se inviáveis e, assim, podem ser desconsiderados.

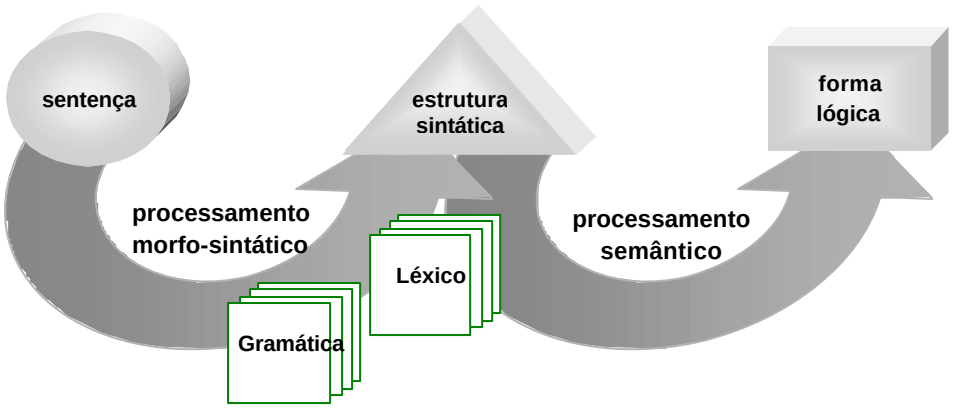

Figura 1. Transformações da sentença na estrutura sintática e na forma lógica

A estrutura sintática de uma sentença é obtida através do processamento morfo-sintático, sendo a representação desta estrutura é regidas por leis gramaticais – definidas em uma gramática. Outras informações necessárias a esta etapa, como as categorias morfológicas das palavras, são encontradas em um léxico.

O mapeamento da estrutura sintática da sentença em sua forma lógica é realizado pelo processamento semântico e, nele, o léxico também exerce papel fundamental, com informações sobre o significado dos itens lexicais.

Vimos que a gramática e o léxico são recursos indispensáveis para a transformação[1] da sentença em sua forma lógica. Vamos examiná-los um pouco mais de perto.

2.2. Gramática

Uma gramática é constituída por um conjunto de regras de boa formação das palavras e das sentenças de uma língua. Essas regras permitem dupla função para as gramáticas [Bouillon1998]: a função normativa, que define regras de combinação das palavras, gerando sentenças corretas; e a função representativa, que associa a uma ou mais frases suas representações sintáticas.

Uma boa gramática deve ser [Allen1995] (a) suficientemente genérica, aceitando o maior número possível de sentenças válidas; (b) seletiva, reconhecendo os casos identificados como problemáticos; e (c) inteligível, favorecendo o entendimento de suas regras, principalmente, pela simplicidade das mesmas.

Uma gramática é dita gerativa quando consegue traduzir os fatos lingüísticos (inclusive os aspectos criativos) da linguagem por meio de regras e processos explícitos, precisos e de aplicação automática, obedecendo a condições específicas [Lobato1986].

Diversos formalismos de representação computacional podem ser usados para representar uma gramática [Nunes1999]. Um destes formalismos é o da gramática de constituintes imediatos (*phrase-structure grammar* – PSG), que é definida como uma quádrupla

<T,N,P,S>,

onde: T representa o conjunto das palavras da língua,
N representa o conjunto das categorias funcionais e das categorias lexicais,
P representa o conjunto de regras de produção, e
S representa o símbolo inicial pertencente a N.

Não há um formalismo eleito como o melhor. Os modelos que se situam entre as gramáticas livres de contexto e aquelas sensíveis ao contexto têm sido propostos pelos pesquisadores como os mais indicados [Vieira2001]. De qualquer forma, quanto ao PLN, é indispensável o uso de critérios formais para a construção das regras gramaticais. Estes vão se aliar a outro recurso do PLN que é o léxico.

2.3. Léxico

De forma genérica, o termo “léxico” significa uma relação de palavras com suas categorias gramaticais e seus significados. Em relação a uma determinada língua, um

[1] Esta visão adota uma abordagem chomskyana, justificada pela ampla e fundamental influência exercida por Chomsky e seus discípulos nesta área.

léxico é o universo de todos os seus itens lexicais, que seus falantes utilizam, já utilizaram ou poderão vir a utilizar [Scapini1995].

Alguns autores argumentam que o termo "dicionário" carrega tipicamente impresso o significado de vocabulário (*wordbook*) para leitores humanos [Guthrie1996]. Em alguns casos, utiliza-se o termo "léxico" para identificar o componente de um sistema de PLN com informações semânticas e gramaticais sobre itens lexicais. Também, usa-se a expressão "base de dados lexical" como sendo uma coleção de informações lexicais, apresentadas em formato estruturado e acessível a sistemas de PLN.

De qualquer forma, o propósito dos dicionários (ou léxicos) é prover uma grande gama de informações sobre palavras, como etimologia, pronúncia, morfologia, sintaxe, entre outras. Eles fornecem definições de seus sentidos e, em decorrência disso, produzem conhecimento não apenas sobre a linguagem, mas sobre o próprio mundo [Guthrie1996].

Quanto ao conteúdo, podemos classificar os dicionários em cinco categorias [Wertheimer1995]: (a) convencionais, com verbetes em ordem alfabética; (b) analógicos, que organizam os itens lexicais de acordo com seu significado; (c) etimológicos, que se ocupam exclusivamente da origem das palavras; (d) morfológicos, que apresentam as formas flexionais dos lexemas; e (e) de sinônimos e antônimos, com listagens de palavras semelhantes ou opostas em significado.

Quanto ao objetivo a que se destinam, os dicionários podem ser classificados, também, em cinco tipos [Wilks1966]: (a) dicionários padrão, que explicam os significados das palavras; (b) thesauri, que apontam relacionamentos entre os itens lexicais; (c) dicionários bilíngües, que buscam relacionar dois idiomas em nível de equivalência de sentidos das palavras; (d) dicionários de estilo, que dão orientações sobre o bom uso das regras gramaticais; e (e) dicionários de concordância, que são essencialmente ferramentas escolares.

No contexto do PLN surgem ainda os dicionários (ou léxicos) com capacidade de serem legíveis e tratáveis por máquina [Wilks1996]. Espera-se que informações lexicais em larga escala possam ser extraídas automaticamente através do que tem sido denominado de "dicionário legível por máquina" (*machine-readable dictionarie* — MRD), melhorando, assim, a uniformidade e a consistência da informação. A capacidade das máquinas de tratar dicionários, entretanto, vai além dos MRDs, com o surgimento dos "dicionários tratáveis por máquina" (*machine-tractable dictionaries* — MTDs). Os MTDs, contendo um grande conjunto de informações lingüísticas, viabilizam a conversão de um dicionário existente em uma forma apropriada para PLN.

Entre os modelos de dicionários com potencial para processamento pelo computador, encontra-se o *Explanatory Combinatorial Dictionary* (ECD) [Mel'cuk1992], que adota o modelo *Meaning-Text Model* (MTM). Este modelo descreve uma linguagem natural como um dispositivo lógico que associa os significados aos textos, com quatro níveis lingüísticos de representação: (a) semântico; (b) sintático; (c) morfológico; e (d) fonético ou ortográfico.

Outro exemplo é o WordNet [Fellbaum1998], descrito por seus autores como uma base de dados lexical legível por máquina e organizada por significado. Ela está dividida em grupos de substantivos, verbos, adjetivos e advérbios. Os itens lexicais são

apresentados através de suas definições, seus diversos sentidos e suas relações com outros itens lexicais. Usa o conceito de *synset*, ou seja, conjunto de sinônimos, para construir o relacionamento semântico básico no WordNet, que é a sinonímia (relação entre sinônimos). Através de *synsets* relacionados é formada uma hierarquia lexical, pela hiponímia (relação entre um hiperônimo, mais genérico, e um hipônimo, mais específico) entre eles, como {*robin*, *redbreast* } → {*bird*} → {*animal*, *animate_being*} → {*organism*, *life_form*, *living_thing*} ({tordo, pisco-de-peito-ruivo} → {pássaro} → {animal, ser_animado} → {organismo, forma-de-vida, ser-vivo}).

Contando, então, com bases de dados lexicais e regras gramaticais, a transformação da sentença em sua forma lógica é iniciada pelo processamento morfo-sintático, discutido a seguir.

2.4. Processamento

2.4.1. Morfo-sintático

Fazem parte do processamento morfo-sintático, a análise morfológica e a análise sintática. A morfologia e a sintaxe tratam da constituição das palavras e dos grupos de palavras que formam os elementos de expressão de uma língua. Enquanto o analisador léxico-morfológico lida com a estrutura das palavras e com a classificação das mesmas em diferentes categorias, o analisador sintático trabalha em nível de agrupamento de palavras, analisando a constituição das frases.

A análise sintática (*parsing*) é o procedimento que avalia os vários modos de como combinar regras gramaticais, com a finalidade de gerar uma estrutura de árvore que represente a estrutura sintática da sentença analisada. Se a sentença for ambígua, o analisador sintático (*parser*) irá obter todas as possíveis estruturas sintáticas que a representam.

O papel do processamento sintático varia em importância [Nunes1999]. Ele tradicionalmente ocupa posição de destaque, com a semântica sendo considerada uma interpretação da sintaxe. Mas, também, pode ser considerado em posição secundária, de acordo com os pesquisadores denominados semântico-gerativistas. Neste último caso, a sintaxe é uma projeção da semântica. Entretanto, qualquer que seja a visão adotada, o processamento sintático é uma etapa indispensável para viabilizar o processamento semântico, que passamos a discutir.

2.4.2. Semântico

Enquanto a sintaxe corresponde ao estudo de como as palavras agrupam-se para formar estruturas em nível de sentença, a semântica está relacionada ao significado, não só de cada palavra, mas também do conjunto resultante delas. O processamento semântico é considerado um dos maiores desafios do PLN, pois se vincula, de um lado, com a morfologia e a estrutura sintática e, de outro lado em alguns casos, com informações da pragmática [Saint-Dizier1999].

Segundo o princípio da composicionalidade, o significado de qualquer construção em uma linguagem depende do significado de cada um dos seus componentes [Allen1995]. Assim, o significado de uma frase, por exemplo, origina-se do significado de cada palavra. Este princípio revela a importância das relações que ocorrem entre os itens lexicais. Quando essas conexões ligam elementos de domínios semânticos, tem sido usual denominá-las “relações semânticas”, enquanto as ligações

entre itens lexicais são tratadas como “relações lexicais”. Entretanto, quando não é possível ou é desnecessária a distinção, é adotado o termo “relações semânticas lexicais” [Evens1992].

As palavras podem se associar através de dois tipos de relações: paradigmáticas [Lyons1979, Evens1992, Pustejovsky1995, Scapini1997, Yule1998, Sacconi1999] e sintagmáticas [Lyons1979, Evens1992] ou colocações [Yule1998]. Entre as relações paradigmáticas estão: sinonímia, antonímia, hiponímia, hiperonímia (em sentido contrário da hiponímia), meronímia (relação entre um holônimo, que representa o todo, e um merônimo, que representa a parte), holonímia (em sentido contrário da meronímia), implicatura e pressuposição. A implicatura é a relação entre A e B, quando B só é verdadeiro se A também for. A pressuposição é a relação entre A e B, quando B é verdadeiro se A ou a negação de A forem verdadeiros.

As relações paradigmáticas associam palavras através do significado, como “nadar” e “água”. As relações sintagmáticas conectam palavras que são freqüentemente encontradas no mesmo discurso, como “água” e “poça”.

As associações de termos [Ruge1999] englobam diferentes tipos de relações semânticas lexicais, como a sinonímia (exemplo: “recipiente” e “receptáculo”), a hiponímia (exemplo: “reservatório” e “tanque”), a meronímia (exemplo: “carro” e “tanque”), a antonímia (exemplo: “aceleração” e “desaceleração”) e a compatibilidade (exemplo: “carro” e “dirigir”), entre outras. Estão incluídos nestas classes de relacionamentos tanto os paradigmáticos quanto os sintagmáticos.

O processo composicional é um dos principais problemas do processamento semântico. Assim, como o significado de um constituinte de uma sentença depende dos significados dos seus sub-constituintes, os significados destes podem, por sua vez, ser determinados por regras gramaticais.

Como alternativa, entretanto, a complexidade do processamento pode ser deslocada das regras gramaticais para o léxico. Um exemplo deste último caso é o conceito de léxico hierárquico [Allen1995]. Nele, os sentidos dos verbos são organizados de tal forma, permitindo que sejam herdados de umas classes para outras.

Outra abordagem neste sentido é a teoria do Léxico Gerativo [Pustejovsky1995], que introduz um conjunto de recursos para análise semântica de expressões em linguagem natural. Esta teoria considera a semântica das palavras isoladas e, também, a capacidade de composição entre elas. O Léxico Gerativo procura abordar (a) explicação da natureza polimórfica da linguagem; (b) a caracterização da semanticalidade (*semanticality*) das expressões em linguagem natural; (c) a captura do uso criativo das palavras em novos contextos; e (d) o desenvolvimento de uma representação semântica co-composicional mais rica.

Todos esses esforços são realizados com o objetivo de processar os prováveis significados de palavras, frases, sentenças e textos. Mas, o que é o significado?

2.5. Questões sobre o significado

2.5.1. O significado do significado

O termo “significado” pode ser utilizado como sendo o sentido da linguagem corrente, como sentido intuitivo, pré-teórico, e podem ser relacionadas três funções da informação semântica codificável em enunciados lingüísticos [Lyons1977]: (a) o

significado descritivo, que pode ser objetivamente verificado; (b) o significado social, que serve para estabelecer e manter relações sociais; e (c) o significado expressivo, que depende do locutor.

Para discutir a formação (ou a descrição) do significado, tem sido usado o conceito de "primitivas" [Wilks1996]. Numa analogia com o conceito de átomo, há a identificação de entidades primitivas, não subdivisíveis, e a partir das quais outras, mais complexas, são formadas. Palavras e primitivas não seriam elementos distintos, mas as primitivas, como característica, têm o propósito principal e específico de definir. Neste sentido, um conjunto de primitivas deve ser entendido como mais um recurso (juntamente com processamento sintático) a serviço da representação semântica de expressões em linguagem natural, através de algoritmos que associem palavras a primitivas, em busca do significado.

As teorias propostas para explicar o que é o significado podem ser classificadas em três grandes grupos de visões [Wilks1996]:

- Visão não-simbólica: o significado pode ser visto como uma coleção de objetos do mundo, ou como algo que se mostra mas não se explica com palavras ou, ainda, como um procedimento de verificação do valor-verdade.
- Casos intermediários: o significado pode ser visto como uma descrição formal, como um mapeamento funcional, associado a atividades sub-simbólicas do cérebro, como um agente seletivo, ou como um estereótipo.
- Visão simbólica: o significado pode ser visto como um conjunto de condições, como o resultado de implicação ou dedução ou, simplesmente, como um símbolo.

Verifica-se que definir o que é o significado é uma tarefa tão difícil quanto estabelecer precisamente a noção de semântica, já que estes termos são usados em contextos e propósitos diversos. O processamento do significado de um item lexical ou de uma sentença enfrenta diversos obstáculos. Entre eles está o problema das variações lingüísticas.

2.5.2. Variações lingüísticas e ambigüidade

As variações lingüísticas [Jacquemin1997, Arampatzis2000] podem ser classificadas em:

- morfológica, quando processos flexionais ou derivacionais criam palavras diferentes, como em "lobo" e "lobos";
- lexical, quando diferentes palavras são usadas para representar o mesmo significado, como "calçado" e "sapato";
- sintático-semântica, quando a posição relativa das palavras determinam frases com significados diferentes, como "biblioteca da ciência" e "ciência da biblioteca";
- morfo-sintática, quando variações morfológicas não impedem a manutenção do significado essencial da frase, podendo ser:
  - variações substantivo-substantivo, como resultado/agente em "fixação de nitrogênio" e "fixador de nitrogênio", ou recipiente/conteúdo em "reservatório de água" e "reserva de água";
  - variações substantivo-verbo, como processo/resultado em "fixação de nitrogênio" e "fixar nitrogênio"; e

    - variações substantivo-adjetivo, onde um modificador preposicional é substituído por modificador adjetival, como em "variação do clima" e "variação climática"; e
- semântica, quando diversos significados são possíveis para o mesmo objeto lingüístico, como "palmas" e "queda da bolsa".

Portanto, objetos lingüísticos de diferentes tipos, como palavras e frases, podem não ter o mesmo significado em cada ocorrência. Isto causa a ambigüidade, que é a propriedade que faz com que uma sentença, por exemplo, possa ser interpretada através de dois ou mais modos diferentes [Lyons1977]. Esta propriedade pode ser atribuída a qualquer objeto lingüístico, seja ele uma palavra, uma frase, ou todo um texto.

Quanto ao nível de processamento do PLN, temos os seguintes tipos de ambigüidade [Jurafsky2000]:

- sintática, quando a ambigüidade é encontrada já em nível sintático; e
- semântica, quando a ambigüidade aparece somente em nível semântico.

A ambigüidade sintática ocorre quando um item lexical pode pertencer a mais de uma classe gramatical, como "casa" em "a bela casa", que pode ser substantivo ou verbo. Outras causas da ambigüidade sintática são [Smeaton1997]: (a) mais de uma ligação possível do sintagma preposicional, como em "comprei um cofre com dinheiro"; (b) mais de uma coordenação ou conjunção possíveis, como em "tenho parentes e amigos gremistas"; ou (c) mais de uma combinação possível para substantivos compostos, como em "lareira da casa de pedras".

Um exemplo de ambigüidade semântica é a que ocorre com o verbo "passar", que pode apresentar mais de um significado, como em "passar a ferro" e em "passar no exame".

Quanto à causa da ambigüidade, podemos encontrar os seguintes tipos [Beardon1991]:

- lexical, que ocorre quando uma palavra possui múltiplos significados; e
- estrutural, quando é possível mais de uma estrutura sintática para a sentença, podendo ser:
    - local, quando a ambigüidade pode ser resolvida em nível de sentença, dispensando o conhecimento do contexto onde ela ocorre; ou
    - global, quando exige análise do contexto para sua resolução.

Há ambigüidade estrutural local em "ele olhou o computador com esperança", e há ambigüidade estrutural global em "ele olhou o colega com esperança". No segundo caso, é possível construir duas associações diferentes: "olhou com esperança" e "colega com esperança". No primeiro caso, a vinculação "computador com esperança" pode, em princípio, ser descartada.

Em relação à ambigüidade lexical temos dois fenômenos a ressaltar: a homonímia e a polissemia.

A homonímia ocorre entre itens lexicais com significados diferentes, que (a) possuem o mesmo som e a mesma grafia (homônimos perfeitos: como substantivo alvo e adjetivo alvo), ou (b) apenas o mesmo som (homônimos homófonos: como acento e assento), ou (c) apenas a mesma grafia (homônimos homógrafos: como verbo "seco" e adjetivo "seco") [Sacconi1999].

Os homônimos homógrafos podem existir (a) por possuírem origem comum (o adjetivo "triangular" e o verbo "triangular"), (b) por coincidência ("vogal", a letra, e "vogal", um membro de júri) ou (c) por derivação (substantivo "procura", derivado do verbo "procurar") [Santos1996].

A polissemia, ao contrário da homonímia, seria o resultado do processo que ocorre, onde diferentes significados vão sendo adquiridos por uma mesma palavra com o tempo, como é o caso das palavras "filme" e "banco".

A normalização lingüística e, mais especificamente, a resolução da ambigüidade, qualquer que seja ela, são essenciais para uma representação eficiente do conhecimento.

2.6. Representação do conhecimento

Duas formas de conhecimento são consideradas cruciais para os sistemas que utilizam representação do conhecimento (RC) [Allen1995]: o conhecimento geral do mundo e o conhecimento específico da situação corrente. Este último pode ser subdividido em: (a) conhecimento semântico lexical, que associa as palavras e suas propriedades sintáticas a estruturas conceituais; e (b) conhecimento de domínio (ou de contexto), que agrega significado aos conceitos [Franconi2001].

Grande parte do conhecimento que se tem sobre aquilo que tratamos e sobre seu ambiente é descritiva, podendo ser expressa de forma declarativa [Genesereth1988]. A tarefa da RC, entretanto, não é simples, pois depende justamente do processamento do significado. A RC pode ser vista como [Davis1993]:

- um *surrogate*: a RC substituiria aquilo que representa, possibilitando que se argumente sobre o mundo através de inferências;
- um conjunto de compromissos ontológicos: a RC estabeleceria um ponto de vista em relação ao mundo, afetando decisões e posicionamentos que, por sua vez gerariam novos compromissos ontológicos;
- uma teoria fragmentária sobre o raciocínio inteligente, que estabelece (a) o modo como o raciocínio é representado, (b) o que podemos deduzir do que conhecemos e (c) o que devemos deduzir do que conhecemos;
- um meio para a computação eficiente em termos pragmáticos: a RC seria um ambiente computacional onde o pensamento se realiza, como um guia que organiza a informação, para facilitar inferências recomendadas; ou
- um meio de expressão humana: a RC seria uma linguagem através da qual discorremos sobre o mundo.

No contexto da inteligência artificial, o objetivo da RC é expressar o conhecimento em uma forma tratável pelo computador [Russel1995], de tal modo que o resultado possa ser útil à comunicação entre as pessoas e a máquina. Tal comunicação está fundamentada em um conjunto de objetos com os quais o conhecimento é expresso e que é denominado universo do discurso [Genesereth1988]. Este conjunto deve atender à conceitualização do mundo por parte de quem expressa o conhecimento, considerando a formalização do mesmo.

2.6.1. Forma lógica

Diversos fenômenos lingüísticos, como o polimorfismo, a composição, a homonímia, entre outros, fazem com que a transformação de uma sentença em sua correspondente forma lógica não seja um processo trivial.

O ideal é que, ao ser adotado um formalismo para representação, este formalismo obedeça aos seguintes critérios [Beardon1991]: (a) prover distintas representações para distintos significados que qualquer conjunto de palavras possa ter; (b) prover uma única representação para dois conjuntos diferentes de palavras que tenham o mesmo significado; (c) ser completo, de forma a representar o significado de qualquer conjunto de palavras; e (d) produzir somente representações que correspondam a significados possíveis.

O cálculo de predicados de primeira ordem (CPPO) é adequada para a RC, pois é tratável computacionalmente e flexível [Jurafsky2000]. É dito de primeira ordem porque os predicados estabelecem relações entre termos. Seria de segunda ordem se os predicados estabelececm relações também entre predicados.

Um CPPO possui os seguintes elementos: predicados e termos. Os predicados representam relações entre objetos. Os termos podem ocorrer na forma de (a) variáveis, que representam classes de objetos, (b) constantes, que representam objetos específicos, ou (c) funções, que, ao serem aplicadas sobre objetos, representam outros objetos resultantes desta aplicação. Outros elementos também podem ser inseridos, como conectivos e quantificadores.

Um CPPO pode combinar unidades primitivas de significado (denominadas “sentido”) para formar o significado de uma expressão mais complexa [Russel1995]. Isto é útil porque o significado de uma sentença, como vimos, é composicional, ou seja, é o resultado do significado de suas partes.

A lógica aplicada às sentenças em linguagem natural consiste [Russel1995] de um sistema formal, para descrever os estados dos acontecimentos, e de teoria de prova. O sistema formar inclui a sintaxe da linguagem, que descreve como devem ser construídas as sentenças, e a semântica, que estabelece as restrições sistemáticas de como as sentenças estão relacionadas aos estados dos acontecimentos. A teoria de prova é um conjunto de regras para deduzir as implicações de um conjunto de sentenças.

Descrições lógicas têm sido usadas para codificar os elementos sintáticos e semânticos necessários em uma base de conhecimento, para orientar o processamento semântico [Franconi2001]. Neste sentido, o PLN busca estabelecer, através de descrições lógicas, as relações entre os objetos lingüísticos e esclarecer a correspondência dos mesmos a situações do mundo real. O que nos leva, por generalização, à representação do próprio conhecimento ontológico, que discutiremos mais adiante.

2.7. Estratégias de processamento

Na busca do significado, no tratamento da ambigüidade e no enfrentamento de outros desafios, por exemplo, para obter uma forma lógica adequada, o PLN pode se apoiar no conhecimento lingüístico e em métodos estatísticos, não necessariamente de forma excludente. Têm sido, inclusive, apontados benefícios quando há a associação de ambos [Bod1995].

2.7.1. Aplicação de conhecimento lingüístico

São citados, a seguir, algumas estratégias de PLN que envolvem conhecimento lingüístico.

A) Etiquetagem de texto

Quando algum conhecimento lingüístico é considerado, a etiquetagem gramatical do texto é um dos passos iniciais. Um etiquetador gramatical (*part-of-speech tagger*) é um sistema que identifica, através da colocação de uma etiqueta (*tag*), a categoria gramatical de cada item lexical do texto analisado [Bick1998]. Enquanto, u etiquetador morfológico inclui informações sobre categorias morfológicas, como substantivo e adjetivo, um etiquetador sintático acrescenta etiquetas indicando as funções sintáticas das palavras, como sujeito e objeto direto.

Além da etiquetagem ou marcação gramatical, existe a etiquetagem semântica [Vieira2000], que anexa informação relacionada ao significado, podendo indicar os papéis dos itens lexicais na sentença, como agente, processo e estado.

B) Normalização de variações lingüísticas

O reconhecimento de variações lingüísticas encontradas em um texto permite, por exemplo, o controle de vocabulário [Jacquemin1997]. A normalização lingüística pode ser subdividida em três casos distintos [Arampatzis2000]: morfológica, sintática e léxico-semântica.

A normalização morfológica ocorre quando há redução dos itens lexicais através de conflação[2] a uma forma que procura representar classes de conceitos. Os procedimentos mais conhecidos para conflação são:

- *stemming*, que reduz todas as palavras com mesmo radical a uma forma denominada *stem* (similar ao próprio radical) [Orengo2001], sendo eliminados afixos oriundos de derivação ou de flexão (em alguns casos, apenas os sufixos são retirados); e
- redução à forma canônica (tratada por alguns autores como *lemmatization*), que, geralmente, reduz os verbos ao infinitivo e os adjetivos e substantivos à forma masculina singular [Arampatzis2000].

No caso da forma canônica, não há perda da categoria morfológica original, ao contrário de um *stem* que pode ser oriundo de palavras de categorias diferentes. Por exemplo, "contruções" e "construiremos" seriam reduzidas ao *stem* "constru", no processo de *stemming*. Por outro lado, ao ser adotada a redução à forma canônica teríamos, respectivamente, "construção" e "construir".

A normalização sintática ocorre quando há a normalização de frases semanticamente equivalentes mas sintaticamente diferentes, em uma forma única e representativa das mesmas, como "processo eficiente e rápido" e "processo rápido e eficiente".

A normalização léxico-semântica ocorre quando são utilizados relacionamentos semânticos (como a sinonímia, hiponímia e meronímia) entre os itens lexicais para criar um agrupamento de similaridades semânticas, identificado por um item lexical que representa um conceito único.

---

[2] Algoritmos de conflação (*conflation*) são aqueles que combinam a representação de dois ou mais termos num único termo, ou seja, reduzem variantes de uma palavra numa forma única [Sparck-Jones1997].

Em nível de item lexical, portanto, podemos encontrar dois extremos de normalização. De um lado está a normalização morfológica através de *stemming*, que explora similaridades morfológicas, talvez, inferindo proximidades conceituais. Em outro extremo está a normalização léxico-semântica, por exemplo, atrvés de busca de sinônimos em thesauri, considerando informações terminológicas [Jacquemin1999].

C) Eliminação de stopwords

A eliminação de *stopwords* pode ser, também, uma estratégia adotada no PLN. *Stopwords* são palavras funcionais, como artigos, conetivos e preposições [Baeza-Yates1999]. Com tal eliminação, corre-se, entretanto, o risco de perder a estrutura composicional de expressões. As preposições, por exemplo, podem exercer papel composicional significativo [Gamallo2002], entretanto, como termos isolados perdem significado ao contrário de outras categorias gramaticais como o substantivo.

As estratégias mencionadas podem se socorrer de gramáticas e/ou bases de dados lexicais ou, também, podem ser executadas com o auxílio de métodos estatísticos.

2.7.2. Aplicação de métodos estatísticos

Métodos estatísticos têm dado grande contribuição ao PLN, como são os casos da lei de Zipf e do gráfico de Luhn.

Zipf, em 1949, estabeleceu o que ficou conhecida como "*constant rank-frequency law of Zipf*" [Moens2000]. Esta lei define que, tomando um determinado texto, o produto $\log(\mathbf{f}_t) \times \mathbf{k}_t$ é aproximadamente constante, onde $\mathbf{f}_t$ é o número de vezes que o termo **t** ocorre no texto e $\mathbf{k}_t$ é a posição deste termo em uma relação de todos os termos daquele texto, ordenados pela freqüência de ocorrência.

Por outro lado, Luhn sugeriu, em 1958, que a freqüência de ocorrência das palavras em um texto pode fornecer uma medida útil sobre a expressividade das mesmas [Frants1997, Moens2000], pois "o autor normalmente repete determinadas palavras ao desenvolver ou variar seus argumentos e ao elaborar diferentes enfoques sobre o assunto que trata". As palavras com maior freqüência de ocorrência deveriam ser consideradas pouco expressivas porque este conjunto de palavras é composto normalmente por artigos, preposições e conjunções. Também as palavras que muito raramente ocorrem deveriam ser consideradas pouco expressivas justamente em razão da baixa freqüência. Sobram como expressivas as palavras com freqüência de ocorrência intermediária, como mostra a **Figura 2.**

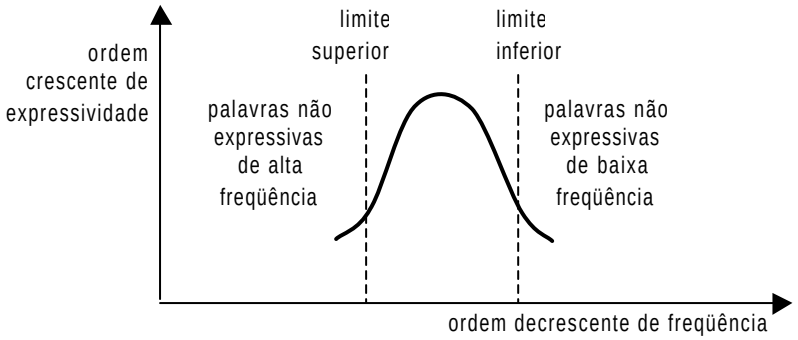

Figura 2. Gráfico de Luhn para relacionar expressividade e freqüência de palavras

A utilização, no PLN, da teoria de probabilidade e das abordagens estatísticas em geral apontam caminhos interessantes para o processamento do significado [Bod1995]. A teoria da probabilidade, por gerar modelos matematicamente precisos para freqüências de ocorrência, e as abordagens estatísticas, por permitirem suposições valiosas em casos de incertezas.

Regida pela teoria da probabilidade, o estudo das variáveis estocásticas introduz diversos tipos de medidas, como distribuição, freqüência, esperança e variança, e também modelos [Krenn1997]. Um modelo probabilístico, útil para o PLN, é o que analisa seqüências de palavras (modelo N-grama ou modelo de Markov de orden n-1) supondo que as n-1 palavras antecedentes afetam a probabilidade da próxima palavra [Jurafsky2000]. Diversas aplicações deste modelo podem ser encontradas no PLN, para criar *clusters* de palavras [Brown1992] ou para extrair unidades lexicais complexas [Dias2000].

No âmbito da teoria da informação, a estatística provê mecanismos para indicar quanta informação ou quanta incerteza temos em relação a um evento, ou, ainda, qual o grau de associação de eventos co-ocorrentes [Krenn1997]. Um desses mecanismos, a informação mútua (IM), leva em conta a probabilidade de ocorrência dos eventos isolados e em conjunto. Quanto maior a probabilidade dos termos ocorrerem juntos em relação às probabilidades de suas ocorrências isoladas, maior a informação mútua. Podem ser encontradas aplicações desta medida, por exemplo, para calcular o grau de similaridade entre palavras [Gauch1994, Mandala1999].

Os métodos estatísticos podem ser utilizados para auxiliar o PLN em diversas situações. Eles têm sido utilizados na etiquetagem gramatical, na resolução de ambigüidade e na aquisição de conhecimento lexical [Krenn1997], entre outras aplicações.

3. Ontologia e thesauri

3.1. Introdução

Em razão da existência da polissemia e da sinonímia, uma palavra pode denotar mais de um conceito e um conceito pode ser representado por palavras distintas [Pustejovsky1995, Scapini1997, Sacconi1999]. Não há dúvidas, também, que são os conceitos, e não as palavras, que representam sem ambigüidade as entidades do mundo real [Clark2000]. Portanto, na RC, é indiscutível a vantagem de se trabalhar com ênfase nos conceitos e não nas palavras. Então, um dos desafios a serem vencidos é o de construir uma ontologia apropriada, ou seja, um vocabulário conceitual, e os thesauri têm sido usados para tal tarefa [Sparck-Jones1986, Loukachevitch1999, Clark2000].

Vamos discutir o conhecimento ontológico e, posterioremente, os thesauri.

3.2. Conhecimento ontológico

Pode-se entender que ontologia é o estudo do ser enquanto ser, ou seja, é o estudo da própria categorização do ser ou, ainda, que é uma espécie de modelagem do conhecimento do mundo, envolvendo objetos, relações e propriedades [Bouillon1998]. Uma ontologia pode se referir a um conjunto de objetos distintos que resultam da análise de um domínio específico ou de um micro-mundo [Jurafsky2000] ou, ao contrário, pode ser construída com abrangência geral, sendo, neste caso, mais difíceis de elaborar que as especializadas [Russel1995].

A elaboração de uma ontologia, em sua essência, consiste numa estruturação dos elementos envolvidos. Tal estruturação, que se assemelha ao formato de árvore, recebe o nome de taxonomia [Jurafsky2000]. Numa taxonomia, normalmente há um conjunto de restrições de formação, que estabelecem quais os relacionamentos válidos, sendo adotada a noção de hierarquia de objetos, com determinados critérios para heranças de características de ancestrais para descendentes.

As palavras, contidas nos léxicos de diversas línguas, tem a função de referenciar os conceitos apresentados numa ontologia [Bouillon1998]. Assim, no PLN, uma ontologia pode dar maior qualidade e mais generalidade ou abrangência ao processamento. A representação do texto ganha profundidade e abstração

Entre os aspectos considerados numa ontologia, para que produza esse conhecimento do mundo, destacam-se principalmente informações sobre categorias e composição dos objetos, mas também outras, como medidas, eventos, processos, tempo e espaço [Russel1995]. Através da classificação dos objetos em categorias seriam estabelecidas classes, coleções, espécies, tipos ou conceitos que incluem componentes com propriedades comuns e associados numa hierarquia taxonômica. Em alguns casos, a composição, mais do que outra característica qualquer, permitiria que objetos pertençam a determinadas categorias. A consideração do conceito de medida possibilitaria associar, aos objetos, propriedades como comprimento e idade e, ainda, quantificadores como massa e número. Poderiam ser, também, reconhecidos eventos e processos, sendo estes últimos considerados eventos contínuos e homogêneos. Eventos ocorreriam em tempos e locais específicos e poderiam ser decompostos em subeventos. Também, ao serem assumidas dimensões temporais e espaciais, poderiam ser tratadas durações diferentes e ocorrência simultânea de eventos.

É através das palavras (e de seus sentidos) que representamos e registramos todas essas noções. A esse registro, na forma de conceitos refletidos nos sentidos das palavras, damos o nome de léxico. Um léxico, então, não contém apenas um conjunto de conhecimento sobre a linguagem mas, mais que isso, armazena informações sobre o mundo [Guthrie1996]. Um dos tipos de léxicos, como vimos, são os thesauri, que passamos a discutir.

3.3. Thesauri

A origem do termo “thesaurus” deve-se à obra elaborada por Roget [Roget1958], com primeira versão editada em 1852. Há diversas definições na bibliografia para thesauri [Gonzalez2001], sendo que a maioria enfatiza o uso de relacionamentos lexicais e a prioridade de manipulação do conceito e não da palavra em si. Um thesaurus $\boldsymbol{Th}(\boldsymbol{C},\boldsymbol{R})$ pode ser definido como um grafo dirigido composto por:

(i) um conjunto de nós finito não-vazio

$\boldsymbol{C} = \{\boldsymbol{c}_k \mid \boldsymbol{c}_k$ é um conceito representado por um item lexical$\}$; e

(ii) um conjunto de arcos finito

$\boldsymbol{R} = \{\boldsymbol{r} = (\boldsymbol{c}_i,\boldsymbol{c}_j) \mid \boldsymbol{r}$ é uma relação semântica lexical$\}$.

O grafo que implementa um thesaurus deve ser dirigido para alcançar seu objetivo: a partir de qualquer conceito, representado por um item lexical, poder chegar a todos os outros itens lexicais relacionados. Apesar de existirem relações simétricas, como sinonímia ou antonímia, também há relações assimétricas, como hiponímia ou meronímia.

Embora a relação se sinonímia seja mais tradicionalmente utilizada em thesauri, são aceitas tanto as relações paradigmáticas, quanto sintagmáticas.

3.3.1. Classificações

Os thesauri podem ser classificados de diversas formas.

Quanto aos recursos utilizados para formação das associações lexicais, podem ser: (a) estatísticos, baseados na co-ocorrência dos termos, em n-gramas, em janelas de texto, ou em similaridade de indexação de textos, (b) sintáticos, baseados no cálculo de similaridade sintática ou em padrões léxico-sintáticos, ou (c) semânticos, baseados na captura do conhecimento semântico lexical;

Quanto ao grau de automatização da construção, podem ser: (a) manuais, assumindo características de uma ontologia de abrangência geral e sendo tipicamente semânticos, ou (b) de geração automática, geralmente estatísticos e de domínio específico;

Quanto à abrangência das informações armazenadas, podem ser: (a) de domínio específico, geralmente dependentes de *corpus*[3], ou (b) genéricos;

Quanto à composição de cada item lexical, podem ser: (a) baseados em palavra, sendo cada termo uma palavra, ou (b) baseados em sintagma, sendo cada termo um item lexical composto por uma ou mais palavras; e

E, finalmente, quanto ao idioma, podem ser: (a) monolíngües, em um único idioma, ou (b) multilíngües, que adotam dois ou mais idiomas.

3.3.2. Construção

Os thesauri podem ser construídos através de métodos orientados a estatística, a sintaxe ou a semântica [Gonzalez2001].

Os principais métodos estatísticos para construção de thesauri utilizam mapeamento de co-ocorrência, basicamente considerando a freqüência de ocorrência dos termos em coleções de documentos. Entre os principais, encontramos os baseados em n-gramas, baseados em similaridade entre palavras encontradas em janelas de texto e, ainda, baseados em similaridade de termos indexação de textos.

Os métodos orientados a sintaxe também utilizam mapeamento de co-ocorrência, como os estatísticos. No entanto, necessitam geralmente de etiquetagem do *corpus* com categorias gramaticais e, assim, levam em consideração não apenas a freqüência de ocorrência entre os termos, mas também o comportamento sintático dos mesmos. Os principais métodos são os baseados em cálculo de similaridade e em padrões léxico-sintáticos.

Os métodos semânticos de construção de thesauri são, geralmente, manuais e utilizam mapeamento de vocabulário controlado ou semântico. Podem ser usados diversos recursos e técnicas, como *text mining*, bases de conhecimento, redes semânticas e outros, envolvendo PLN. Entretanto, por serem preferencialmente manuais, os thesauri com motivação semântica são construídos através de métodos que consistem em preencher campos de informação como aqueles estabelecidos pela norma

[3] Conjunto de documentos sobre um determinado assunto ou com uma finalidade comum.

ISO-2788, como termo preferencial e relacionamentos como USE (termo usado), UF (use por), BT (termo genérico) e NT (termo específico).

3.3.3. Aplicações

Um thesaurus pode ter as seguintes aplicações [Soergel1997, Sparck-Jones1997]: (a) apoio para classificação de documentos, na caracterização de temas e categorização de conceitos; (b) apoio à produção e à tradução de textos, principalmente na seleção de vocabulário; (c) comunicação e aprendizado, na geração da estrutura conceitual; (d) base conceitual para projetos, na produção do contexto conceitual; (e) apoio à tomada de decisão, na classificação de assuntos; (f) apoio à sumarização de textos, na identificação e associação dos principais conceitos desenvolvidos; ou (g) apoio à recuperação de informação.

Neste último caso, um thesaurus pode ser utilizado na geração de uma base de conhecimento para consulta por navegação em tópicos ou na associação de termos, em expansão automática ou manual de consultas [Gonzalez2001b], ou na estruturação da apresentação dos resultados da pesquisa [Soergel1997], ou na normalização do vocabulário para a indexação [Soergel1997, Sparck-Jones1997].

Os thesauri podem ser considerados [Soergel1997]: (a) dicionários analógicos para uso humano, quando possibilita a pesquisa do significado e não diretamente do item lexical; (b) bases conceituais para sistemas baseados em conhecimento em geral, quando provê a construção de ontologias e taxonomias; ou (c) bases de conhecimento especificamente para PLN, quando se constitui em dispositivo de compreensão da linguagem natural para extração de informação, sumarização ou indexação automatizadas de textos.

4. Recuperação de Informação

4.1. Introdução

Desde que o termo *information retrieval* surgiu em 1950 tem gerado muita polêmica [Swanson1988]. Hoje, entretanto, é largamente aceito e utilizado pela comunidade científica, ainda que existam diversos pontos de vista do que seja “recuperação de informação” (RI). Algumas delas serão discutidas a seguir.

Entre as tarefas de um sistema de banco de dados, além da inserção, da atualização e da eliminação de dados, encontra-se a recuperação de dados [Date1991]. Entretanto, é necessário que os dados armazenados estejam agrupados por propriedades específicas e que sejam consideradas categorias conceituais bem definidas para as consultas [Lewis1996]. Logo, RI não é recuperação de dados. Ou seja, ao entender que o foco de um sistema de RI é a informação textual, é válido supor que um sistema de RI seja diferente de um sistema de gerenciamento de banco de dados [Sparck-Jones1997].

Um sistema de RI tem como meta encontrar a informação exigida para satisfazer a necessidade de informação (NI) do usuário [Frantz1997]. Para tanto, além da recuperação propriamente dita, um sistema de RI deve ser capaz de realizar armazenamento e manutenção de informação [Kowalski1997]. Ou seja, além do procedimento de busca, pode incluir catalogação, categorização e classificação de informação, particularmente na forma textual [Strzalkowski1999]. Em outras palavras, deve representar, organizar e dar acesso a itens de informação (documentos) [Baeza-Yates1999].

Neste caminho, logo se constata que um sistema de RI não recupera informação, pois a informação consiste no relacionamento que ocorre entre o usuário e os sinais que recebe [Frantz1997]. Assim, pelo caráter extrínseco da informação enfocado pela teoria dos sistemas, quem a pode recuperar informação é o usuário e não o sistema.

Freqüentemente a expressão "recuperação de informação" é tratada como um sinônimo para "recuperação de documentos" ou "recuperação de textos" [Sparck-Jones1997], acreditando-se que sua tarefa essencial seja recuperar documentos ou textos com informação.

Vamos considerar, aqui, que sistemas de RI são sistemas que tratam essencialmente de indexação, busca e classificação de documentos (textuais), com o objetivo de satisfazer NIs expressas através de consultas.

Tendo em mente esta visão, vamos analisar como estes sistemas surgiram e evoluíram.

4.2. Histórico

Podemos considerar a existência de três gerações de sistemas de RI [Baeza-Yates1999]: (a) primeira geração, quando os sistemas de RI consistiam basicamente de catálogo de cartões, contendo principalmente nome do autor e título do documento; (b) segunda geração, quando ocorreram acréscimos nas funcionalidades de busca, permitindo pesquisa por assunto, por palavras-chave e outras consultas mais complexas; e (c) terceira geração, quando o foco é o uso de interface gráfica, de formulários eletrônicos, de características de hipertexto e de arquiteturas de sistemas abertos, como ocorre atualmente.

Para tentar satisfazer as NIs mais adequadamente, o foco na sintaxe tem sido desviado para a semântica. Este fato pode ser evidenciado pela evolução dos sistemas de RI [Schatz1997], sendo constatadas grandes fases por que passaram as estratégias concebidas para RI: soluções universais, busca de textos, busca de documentos e busca de conceitos.

Soluções universais foram consideradas até meados dos anos 60. cogitava-se a ficção de que os sistemas de RI tratariam de coleções universais de documentos, onde o usuário navegaria buscando informações de todo o tipo e de diferentes fontes.

Na fase de busca de textos, inicialmente, a ênfase consistia em utilizar bases de dados onde eram pesquisados dados de referências bibliográficas. Houve a introdução dos operadores lógicos nas consultas. Posteriormente passaram a ser utilizados índices invertidos, indexação automática, recuperação full-text, redução das palavras ao seu radical, estatísticas de co-ocorrência de termos, técnicas de probabilidade e busca por proximidade de palavras. A sintaxe era o alvo fundamental nas pesquisas.

Na fase de busca de documentos, nos anos 80, consolidou-se a tecnologia full-text e, nos anos 90, surgiu a Web e chegaram os *browsers* multimídia. O modelo de grandes computadores compartilhados evoluíram para estações de trabalho pessoais distribuídas. Múltiplas coleções de documentos passaram a ser armazenadas em locais fisicamente dispersos. Estilos diferentes de interação tornaram-se possíveis.

Na atual fase de busca de conceitos, as abordagens para a captura da informação semântica contida nos textos, com o objetivo de construir índices, ainda envolve intermediários humanos, exigindo tarefas como a etiquetagem de termos. Entretanto, a

utilização cada vez mais freqüente de expressões como “normalização semântica” e “expansão de consulta baseada em conceitos”, evidencia a preocupação e a tendência corrente.

Tabela 1. Resumo histórico da pesquisa em RI

Alguns fatos e momentos históricos	**Local**	**Autor**	**Data**
Uso, pela primeira vez, do termo “*information retrieval*”	EUA	Calvin Mooers	1950
Elaboração da primeira abordagem estatística para indexação	EUA	H. P. Luhn	1957
Início da pesquisa sistemática em RI	EUA - International Conference on Scientific Information		1958
Introdução de thesaurus na RI	Inglaterra	T. Joyce e R. M. Needham	1958
Popularização do termo “*information retrieval*”	Inglaterra	Robert A. Fairthorne	1961
Publicação de estudos sobre RI automática (principalmente indexação automática): pós-coordenação, estatística de associação de termos, PLN, peso de termos, realimentação de relevância	EUA	Lauren B. Doyle	1961
	EUA	H. P. Luhn	1961
	EUA	Gerard Salton	1968
Surgimento do SMART, primeiro sistema de RI com indexação automática	EUA	Gerard Salton	1965
Elaboração dos primeiros testes para linguagens de indexação manual	Inglaterra	Cyril Cleverdon	1967
Publicação de artigos que impulsionam a recuperação automática	EUA	T. Saracevic	1970
Surgimento do DIALOG e do MEDLINE, primeiros sistemas de RI interativos	EUA e Inglaterra	F. W. Lancaster e E. G. Fayen	1973
Incentivo aos métodos estatísticos na RI, devido às dificuldades da lingüística computacional, no final dos anos 60, em tradução automática de textos			anos 70
Introdução dos “*online public access catalogues*” (OPACs)			
Elaboração dos princípios da classificação por probabilidade	EUA	S. E. Robertson	1977
Ocorrência da primeira Text Retrieval Conference (TREC-1)	EUA		1992
Aplicações de técnicas oriundas das ciências comportamentais e cognitivas	EUA	D. Ellis	1992
Surgimento dos primeiros *sites* de busca	*Web*		1994

Para que esta evolução tenha ocorrido e continue a ocorrer, os pesquisadores da área têm trabalhado desde os meados do século XX. A Tabela 1 apresenta um resumo do caminho percorrido até agora pela pesquisa em RI. Mais informações podem ser obtidas em [Sparck-Jones1997] e [Baeza-Yates1999]. Este histórico determinou as características dos atuais sistemas de RI.

4.3. Classificação dos sistemas de RI

Os sistemas de RI podem ser classificados quanto ao modo operacional e quanto às tarefas do usuário [Baeza-Yates1999].

Quanto ao modo operacional, temos sistemas de RI convencionais ou ad-hoc e de filtragem (filtering). No primeiro caso, os documentos da coleção permanecem relativamente estáticos enquanto novas consultas são submetidas ao sistema. Exemplo: sistemas de pesquisa de documentos em bibliotecas digitais. Nos sistemas de filtragem,

as consultas permanecem estáticas enquanto novos documentos entram no sistema. Exemplo: sistemas de classificação de documentos.

Quanto às tarefas do usuário, temos: (a) sistemas de RI propriamente ditos, onde o usuário propõe consultas que orientarão as pesquisas; ou (b) browsing, onde o usuário navega através de "páginas" selecionando links em busca de documentos.

Um outro tipo de classificação está relacionada ao ambiente de uso do sistema de RI. Embora estes sistemas tenham sido usados inicialmente em bibliotecas, podem hoje ser classificados, quanto ao ambiente onde são utilizados, em três níveis diferentes [Sparck-Jones1997]: (a) nível 1, com sistemas constituídos por um conjunto de rotinas, tipicamente desenvolvidos para serem usados por um grupo de pesquisa com a finalidade de facilitar novas pesquisas e desenvolvimentos adicionais; (b) nível 2, com sistemas associados a coleções de documentos e a ambientes de software/hardware de uma instituição; e (c) nível 3, com sistemas comerciais amigáveis destinados a uma ampla gama de usuários com perfis variados, utilizando diversos tipos de coleções de documentos em diferentes plataformas.

4.4. Componentes

São componentes iniciais de um sistema de RI: o usuário, sua NI e a coleção de documentos disponíveis para pesquisa. São componentes adicionais: a consulta, que traduz a NI do usuário; os índices, que representam os documento da coleção; e a referência (ou *surrogate* [Sparck-Jones1997, Meadow2000]) a cada documento da coleção, que pode ser constituída por: título, resumo, nomes dos autores, trecho do texto que contém os termos da consulta, etc. Pode ser também um componente adicional, em alguns sistemas, a consulta expandida por realimentação (*feedback*).

A Figura 3 mostra o esquema geral de um sistema de RI típico, com os relacionamentos entre seus componentes. Na Figura 3 se observa que, além da interface com o usuário e de um sistema de gerenciamento de base de dados (SGBD) para a coleção de documentos, existem alguns processos necessários à RI: indexação, construção da consulta (que pode incluir expansão) e, finalmente, busca e classificação.

Então, considerando uma necessidade de informação N de um usuário, em um dado momento, e uma coleção de documentos

$D = \{ d_i \mid d_i$ é um documento contendo texto em linguagem natural $\}$,

a resolução do problema da RI consiste em especificar:

- funções de representação para obter uma estrutura de indexação I, que represente D, e a consulta q, que represente N; e
- a função de recuperação para obter o conjunto

  $D_q = \{ d_k \mid d_k$ é um documento relevante à consulta q, sendo $D_q \subseteq D \}$.

O conceito de relevância é bastante subjetivo [Saracevic1975]. Entretanto, se considerarmos N expressa da seguinte forma: "Quais são os documentos que tratam de um determinado tema com maior abrangência e profundidade?", D_q seria o conjunto de documentos esperado como resposta a esta questão.

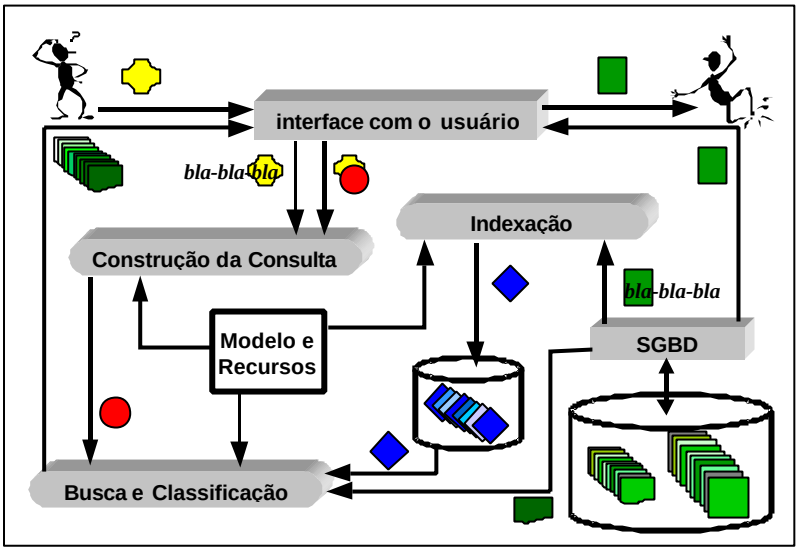

Figura 3. Esquema geral de um sistema de RI

Os elementos presentes na Figura 3 são:

- NI do usuário;
- consulta;
- realimentação da consulta;
- documento;
- índice; e
- *surrogate.*

Um sistema de RI deve adotar um modelo (ou características de alguns modelos – discutidos na próxima seção) e necessita de recursos, ou seja um conjunto **L** de estruturas de dados de apoio, como um thesaurus ou, simplesmente, uma *stoplist* (uma lista de *stopwords*). Tais recursos são opcionais, sendo constituídos conforme a estratégia adotada para a função de recuperação.

Podemos, então, identificar dois momentos distintos durante a execução de sistema de RI: de indexação e de busca [Sparck-Jones1997, Baeza-Yates1999].

Em tempo de indexação, a estrutura **I** é construída através da função de representação, e em tempo de busca, que inclui consulta, busca e classificação, a função de recuperação é executada. A indexação é a etapa em que os documentos são representados para propósitos de recuperação. A busca tem a ver com o momento em que os arquivos são examinados e os itens neles contidos são comparados, de algum modo, aos itens da consulta. A distinção entre estes dois momentos é crucial. Os procedimentos em tempo de indexação estão fortemente comprometidos com a seleção dos termos mais representativos dos documentos – os termos de índice –, o que não ocorre em tempo de busca, quando se parte do princípio que esta representatividade é correta, e que a pesquisa no índice tem resultado similar à pesquisa feita diretamente no documento.

4.5. Modelos clássicos

Em geral, a classificação de um sistema de RI em um determinado modelo é incerta e complicada, já que, geralmente, diversos dispositivos e estratégias são combinados.

Entretanto, é possível considerar três modelos clássicos [Sparck-Jones1997, Baeza-Yates1999]: Lógico, Vetorial e Probabilístico.

4.5.1. Modelo Lógico

Neste modelo, um documento é representado como um conjunto

$\mathbf{d}_i = \{ \mathbf{t}_j \mid \mathbf{t}_j \in \mathbf{I} \}$.

Uma consulta seria uma expressão lógica, como

$\mathbf{q} = (\mathbf{t}_1 \vee \mathbf{t}_2) \wedge \neg \mathbf{t}_3$.

O mecanismo de busca retorna os documentos que possuem combinações dos termos que satisfazem à construção lógica da consulta.

Nesses sistemas são utilizadas a teoria dos conjuntos e a álgebra booleana. Assume-se que todos os termos possuem peso 1 ou 0, dependendo de estarem presentes ou ausentes na consulta, respectivamente. Ou seja, no modelo Lógico tipicamente não há classificação de documentos, já que o cálculo da similaridade ($\mathbf{S}$) de uma consulta em relação a um documento será:

$\mathbf{S}$ ($\mathbf{q}$, $\mathbf{d}_i$) = 1 ou 0.

Os principais problemas [Sparck-Jones1997] do modelo Lógico são: (a) normalmente, o usuário não possui treinamento apropriado, tendo dificuldade em formular consultas usando operadores lógicos; (b) há pequeno controle sobre o tamanho da saída produzida por uma determinada consulta; e (c) a recuperação lógica resulta em uma simples partição da coleção de documentos em dois subconjuntos discretos: os registros que satisfazem a consulta e os que não a satisfazem.

Estas limitações têm incentivado o desenvolvimento de modelos de conjuntos *fuzzy*, que são mais flexíveis em termos de pertinência estrita a determinada classe.

4.5.2. Modelo Vetorial

Além do modelo Lógico, o que mais tem influenciado o desenvolvimento de sistemas de RI e, consequentemente, o desenvolvimento de sistemas de RI operacionais, é o modelo Vetorial. Um documento é representado como um conjunto

$\vec{\mathbf{d}}_i = (\mathbf{p}_{1i}, \mathbf{p}_{2i}, \mathbf{p}_{3i}, \ldots)$,

onde $\mathbf{p}_{ji}$ é o peso (entre 0 e 1) do termo $\mathbf{t}_j \in \mathbf{I}$, referente ao documento $\mathbf{d}_i$.

A idéia essencial é que os termos de índice são considerados como coordenadas de um espaço multidimensional de informação. Da mesma forma uma consulta $\mathbf{q}$ é visualizada como um vetor

$\vec{\mathbf{q}} = (\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots)$,

onde $\mathbf{p}_j$ é o peso (entre 0 e 1) do termo $\mathbf{t}_j \in \mathbf{N}$, ou seja, do contexto da NI do usuário.

O conjunto completo de valores dos vetores dos documentos e da consulta, consequentemente, descreve a posição dos mesmos no espaço. Portanto a indexação, neste modelo, também pode ser visualizada como a distinção, entre os documentos, pela separação dos mesmos em um espaço de termos multidimensional.

A similaridade entre documento e consulta (isto é, sua distância no espaço) pode ser calculada pela comparação de seus vetores, usando uma medida de similaridade tal como coseno. A similaridade ($\mathbf{S}$) de uma consulta em relação a um documento será:

$$\mathbf{S}(\mathbf{q}, \mathbf{d}_i) = \cos(\theta),$$

onde θ é o ângulo entre os vetores que representam o documento $\mathbf{d}_i$ e a consulta $\mathbf{q}$. Esta interpretação geométrica provê fundamento para uma ampla série de operações de recuperação, incluindo indexação, realimentação de relevância e classificação de documentos. A realimentação de relevância pode ser vista tanto como um processo que avalia novamente os pesos dos termos existentes na consulta, quanto como uma alteração da composição da consulta pela adição ou eliminação de termos.

É provável que documentos similares, isto é, que encontram-se em uma mesma parte do espaço multidimensional de termos, tenham a mesma relevância em relação às mesmas consultas, devendo ser armazenados e recuperados juntos. Esta observação, que forma a base para a "hipótese de *cluster*", sugere que a comparação de uma consulta com grupos de documentos resultará em altos níveis de efetividade na recuperação.

O uso de termos de índice para definir as dimensões do espaço onde ocorre a recuperação envolve pressupor que os termos são ortogonais, o que não é correto. Podem ser consideradas duas limitações práticas a essa idéia. A primeira é a necessidade de vários termos de consulta (se uma classificação seletiva precisa ser obtida), neste modelo, enquanto dois ou três termos ligados por operadores lógicos E seriam suficientes, no modelo Lógico, para obter uma saída de alta qualidade. A Segunda limitação é a dificuldade de explicitar relacionamentos específicos de sinonímia, e também relacionamentos entre termos compostos por mais de uma palavra, tarefas facilmente realizáveis, no modelo Lógico, através de operadores lógicos OU e E.

4.5.3. Modelo Probabilístico

Pode-se entender um sistema de RI como tendo a função de classificar documentos de uma coleção em ordem decrescente de probabilidade de relevância a uma NI de usuário. Esta observação é freqüentemente referenciada como "princípio da classificação por probabilidade", sendo utilizada no modelo Probabilístico.

A idéia básica é usar os dados obtidos sobre a distribuição dos termos da consulta nos documentos acessados e tidos como relevantes. Estas informações permitem o cálculo dos pesos dos termos da consulta que definem a probabilidade de serem relevantes ou não os documentos que não foram ainda analisados.

O modelo Probabilístico usa a especificação das propriedades do conjunto ideal de resposta (CIR), com realimentação da consulta. O cálculo da similaridade ($\mathbf{S}$) entre um documento $\mathbf{d}_i$ e a consulta $\mathbf{q}$ seria

$$\mathbf{S}(\mathbf{q}, \mathbf{d}_i) = \frac{\text{probabilidade de } d_j \in \text{CIR}}{\text{probabilidade de } d_j \notin \text{CIR}}$$

4.6. Pesos dos termos

O estabelecimento de pesos para os termos é um dos mais importantes fatores a determinar a efetividade de um sistema de RI [Sparck-Jones1997], podendo ser usado tanto na linguagem de indexação quanto na de consulta.

O cálculo do peso de um termo, como um indicador da importância do mesmo para o texto onde está presente, pode levar em conta diversos parâmetros [Moens2000], entre eles podemos citar: a categoria gramatical do termo; o número de palavras (ou de

termos) diferentes no texto onde o termo ocorre; a freqüência de ocorrência, a localização, os relacionamentos com outros termos e o contexto onde ocorre; e a freqüência de ocorrência no *corpus*, ou seja, na coleção de documentos.

Algumas estratégias principais são relacionadas a seguir para o cálculo do peso dos termos de índice [Frants1997, Moens2000]. Em geral, os mesmos critérios podem ser aplicados à consulta. Variações das abordagens apresentadas aqui são largamente encontradas na bibliografia.

Consideremos que

$\mathbf{p}_{ji}$ = peso do termo $\mathbf{t}_j$ para o documento $\mathbf{d}_i$;

$\mathbf{f}_{ji}$ = número de ocorrências (ou percentagem: número de ocorrências normalizado pelo número total de termos do documento sendo, em alguns casos, excluídas as *stopwords*) do termo $\mathbf{t}_j$ no documento $\mathbf{d}_i \in \mathbf{D}$;

$|\mathbf{D}|$ = número de documentos de $\mathbf{D}$;

$|t_j d|$ = número de documentos em $\mathbf{D}$ que contêm o termo $\mathbf{t}_j$; e

$\mathbf{F}_j = \Sigma_{1 \le i \le |\mathbf{D}|} \mathbf{f}_{ji}$ = freqüência de ocorrência do termo $\mathbf{t}_j$ no *corpus*;

então podemos calcular o peso de um termo $\mathbf{t}_j$ em relação a um documento $\mathbf{d}_i$ das seguintes maneiras:

- Freqüência simples: quanto mais freqüente o termo, maior seu peso, sendo
  $\mathbf{p}_{ji} = \mathbf{f}_{ji}$
- Freqüência normalizada pelo número de documentos: quanto menor o *corpus*, maior o peso do termo, sendo
  $\mathbf{p}_{ji} = \mathbf{f}_{ji} / |\mathbf{D}|$
- Freqüência normalizada pela ocorrência na coleção de documentos: quanto mais raro o termo no *corpus*, maior seu peso, sendo
  $\mathbf{p}_{ji} = \mathbf{f}_{ji} / \mathbf{F}_j$
- IDF (*inverse document frequency*): quanto maior o *corpus* e quanto menor o número de documentos em que o termo ocorre, maior seu peso, sendo
  $\mathbf{p}_{ji} = IDF_j = \log (|\mathbf{D}| / |t_j d|)$
- tf.IDF: quanto mais freqüente o termo e maior seu IDF, maior seu peso, sendo
  $\mathbf{p}_{ji} = \mathbf{f}_{ji} \times IDF_j$

4.7. Arquivos de indexação

As três principais técnicas de construção de arquivos de indexação são: as árvores de sufixos, os arquivos de assinatura e os arquivos invertidos [Baeza-Yates1999].

Árvores de sufixos são mecanismos que permitem rapidez nas operações de pesquisa mas são de difícil construção e manutenção. As árvores Patrícia (*Patricia tree*) [Frakes1992] são um tipo de árvore de sufixo compactada. São construídas a partir de todas as subseqüências possíveis de caracteres do texto (iniciadas em um determinado ponto do texto). Para um texto com n caracteres, a árvore terá n nós folha e n-1 nós internos. Os arcos indicam o caractere representado. Os nós internos indicam o número de posições entre o caractere anterior e o posterior. Cada nó folha indica a posição no texto da subseqüência representada no caminho do nó-raiz ao próprio nó-folha.

Para a construção de arquivos de assinatura (*signature files*) [Frakes1992] os documentos são divididos em blocos lógicos, contendo um número determinado de palavras (sendo eliminados as *stopwords*). Cada palavra possui sua "assinatura", um padrão de bits obtido através de uma função *hash*. Os blocos de assinaturas são concatenados formando o arquivo de assinatura, que é utilizado para a busca das palavra contidas no texto do documento. Encontrada a palavra no arquivo de índice, um ponteiro a localiza no documento. Arquivos de assinatura foram utilizados com freqüência nos anos 80, mas atualmente perdem, em preferência, para os arquivos invertidos.

Um arquivo invertido (ou índice invertido), a técnica mais utilizada atualmente, é um mecanismo orientado à palavra para indexar uma coleção de textos com o objetivo de agilizar a tarefa de busca. Arquivos invertidos normalmente contêm dois componentes principais [Sparck-Jones1997]: dicionário e endereçamentos. O dicionário é constituído por uma lista de todas as palavras-chave, todos os termos classificados, títulos, etc, em uma base de dados que pode ser usada como chave de recuperação. Os endereçamentos são constituídos por uma série de listas, uma para cada entrada do dicionário. Cada uma destas listas possui identificadores de todos os documentos que contêm o termo corrente. É comum não apenas armazenar a presença do termo no documento mas, também, sua localização, para permitir maior eficiência na implementação de busca por proximidade ou viabilizar a consulta de termos compostos.

Com os arquivos de índice estruturados, a atenção se concentra nos procedimentos de busca.

4.8. Procedimentos de busca

Sob o ponto de vista da busca, temos as operações de inspeção, verificação de similaridade (*matching*), classificação (*scoring*) e saída, com requisitos para as representações de documentos e consultas, critérios e componentes específicos.

A fim de otimizar a busca de similaridade, um sistema de RI trabalha com representações de documentos e de consultas através de linguagens de representação (de indexação e de consulta), como vimos. Estas representações devem ser projetadas para atender dois requisitos [Sparck-Jones1997]: (a) que seja assegurado que as correspondentes relações de relevância sejam mantidas, quando houver similaridade entre as representações dos documentos e das consultas; e (b) que os meios para alcançar tal objetivo também não permitam ou incentivem similaridades onde a relação de relevância não seja mantida.

Diversos recursos são utilizados nos procedimentos de busca [Smeaton1997]: (a) utilização de *clusters* de documentos, usualmente pré-classificados, para agilizar a busca e compor o conjunto de documentos recuperados; (b) utilização de diversas estratégias de classificação consolidadas, geralmente envolvendo mais de uma versão da mesma consulta, executada sobre a coleção de documentos; (c) utilização de indexação semântica latente (*latent semantic*), que basicamente reduz o número de índices ao considerar interdependências e relacionamentos termo a termo; e (d) utilização de trechos de texto (*passages*), onde os itens recuperados são seções internas aos documentos.

4.9. Avaliação

A avaliação de sistemas de PLN exige um elevado conhecimento sobre o problema que se deseja resolver e, também, o desenvolvimento de metodologias próprias [Santos2000]. Para bem avaliar é necessário que o problema seja quantificado e que as vantagens de uso sejam identificadas [Santos2001].

Um sistema de RI pode ser analisado não apenas pelo mecanismo de recuperação usado para comparar consultas com conjuntos de documentos, mas também [Sparck-Jones1997]: (a) pelo modo como a NI do usuário pode ser formulada; (b) pela interação usuário-computador necessária para realizar o processamento apropriado para a pesquisa; e (c) pelo ambiente social e cognitivo no qual esta interação se dá.

Entretanto, as evidências maiores, quanto a dados de avaliação quantitativos, são oriundas das *Text REtrieval Conferences* (TRECs), conforme é relatado a seguir.

As TRECs foram projetadas para incentivar a pesquisa em RI, levando em conta aplicações reais envolvidas com grandes coleções de texto, procedimentos de avaliação uniformes e um fórum para organizações interessadas em comparar resultados [Voorhees1999, Voorhees2000]. São conferências anuais que acontecem desde 1992 com o objetivo de incentivar a interação entre os grupos de pesquisadores nas empresas e no ambiente acadêmico. As TRECs são promovidas pela DARPA (Defense Advanced Research Projects Agency) e pelo NIST (National Institute of Standards and Technology) dos Estados Unidos. São considerados importantes exercícios de avaliação de sistemas de RI, com muitos participantes de diversos países e grande variedade de técnicas. São utilizadas grandes coleções de documentos de diversas fontes, como Wall Street Journal, AP Newswire, documentos do Federal Register e do setor de Patentes dos EUA, além de artigos e *abstracts* de diversas publicações.

Basicamente os sistemas de RI são avaliados conforme suas performance através das métricas precisão (*precision*) e resposta (*recall*), conforme as seguintes fórmulas:

$$\text{precisão} = \frac{|\mathbf{D_r}| \cap |\mathbf{D_q}|}{|\mathbf{D_r}|} \quad \text{e} \quad \text{resposta} = \frac{|\mathbf{D_r}| \cap |\mathbf{D_q}|}{|\mathbf{D_q}|}$$

onde

$|\mathbf{D_r}|$ = número de documentos recuperados pelo sistema para uma consulta $\mathbf{q}$;
$|\mathbf{D_q}|$ = número de documentos relevantes a uma consulta $\mathbf{q}$; e
$|\mathbf{D_r}| \cap |\mathbf{D_q}|$ = número de documentos relevantes recuperados.

Os melhores experimentos observados nas TRECs têm utilizado estabelecimento de pesos, cuidados com a correta formulação de consulta, técnicas para expansão de consulta, realimentação da consulta a partir dos documentos recuperados com maior relevância, uso de termos compostos para índices e uso de *passages*, ou seja, pesquisa em trechos de texto homogêneos em conteúdo [Voorhees1999].

As conclusões sobre a avaliação, nas TRECs, dos sistemas de RI com PLN incluído [Strazalkowski1999b] são resumidas a seguir.

As técnicas de PLN que aplicam conhecimento lingüístico possuem grande potencial, sendo aparentemente superiores aos métodos puramente quantitativos, entretanto as evidências para tal afirmação ainda são pequenas. Ainda assim, é possível observar algo de positivo. A comparação entre a abordagem booleana e as que utilizam

linguagem natural indicou, com relação a técnicas interativas de busca, que a performance da busca com linguagem natural é superior. A construção de índices com técnicas lingüísticas sofisticadas e de alto custo também é promissora, embora sua potencialidade quanto à consulta pareça ser maior, principalmente com a utilização de estratégias de expansão.

5. RI e PLN

5.1. Introdução

Vimos que os sistemas de RI, em resumo, tratam da indexação de textos e da seleção daqueles que são relevantes a uma determinada consulta de usuário. Também vimos que o PLN trata de diversos aspectos da língua, incluindo palavras e sentenças (ou seja, texto), e seus significados. Agora verificaremos como o PLN é aplicado à RI. Podemos relacionar, inicialmente, algumas possibilidades [Smeaton1997]: (a) na representação de consultas e documentos, o PLN pode contribuir para identificar termos compostos que constituam bons descritores, representativos do conteúdo dos textos; (b) na formulação das consultas, pode auxiliar o usuário a refinar sua NI; e (c) na busca, pode incorporar procedimentos de análise que envolvam inferências semânticas.

Um sistema de RI possui, entre seus componentes, dois conjuntos compostos de representações lógicas: um, formado pelos índices, representando os documentos, e outro, formado pelas consultas, representando as NIs do usuário [Baeza-Yates1999]. Portanto, constitui-se em um campo de aplicação propício para a RC. Porém, mesmo que algumas técnicas de PLN sejam aplicadas a muitos dos sistemas de RI, métodos mais avançados só algumas vezes são utilizados [Allen2000].

A RC está fortemente relacionada à construção de bons descritores na indexação, assim como tem papel crucial na formulação correta da consulta. Isto, em conseqüência, afeta a busca de similaridade entre tais representações. A RC utilizada em sistemas de RI, entretanto, é muito precária [Sparck-Jones1999]. Conceitos não são normalizados e descrições são meramente conjuntos de termos sem estrutura e sem economia. Pesos são adicionados apenas como refinamentos para esquemas básicos e, assim mesmo, refletem o significado muito grosseiramente, tanto em relação aos termos individualmente, quanto em relação ao todo. Conceitos e tópicos são incluídos de forma implícita quando deveriam estar explícitos, por exemplo, num formato proposicional.

Para a resolução destes e de outros problemas, entre os principais desafios da aplicação do PLN na RI, estão os seguintes [Lewis1996]: (a) a combinação apropriada de recursos não estatísticos e estatísticos; (b) o uso de recursos terminológicos, garantindo o cálculo de similaridade correto entre termos de índice e de consultas; (c) o tratamento de consultas, que envolvem pequena quantidade de texto mas grande variedade em formas e conteúdos; e (d) a definição de estratégias para análise semântica a ser realizada sobre grande quantidade de texto.

Ao se voltarem para o PLN, já há algum tempo, os pesquisadores de RI encontram uma série de questões a serem resolvidas nesta área [Lewis1996]. Na indexação de documentos, por exemplo, pode haver benefícios ao ser utilizado vocabulário controlado, mas isto implica utilização de técnicas avançadas. Também há evidências que sugerem que a utilização combinada de termos simples e compostos, representando conceitos complexos, pode ser útil. O PLN pode ser usado com o objetivo de capturar palavras ou sentenças como descrições contidas em documentos,

expressando relações sintagmáticas e reconhecendo a composição dos termos. Relações paradigmáticas entre itens lexicais também podem ser identificadas, permitindo a substituição controlada de termos na indexação e na consulta.

5.2. PLN básico

5.2.1. Indexação

Um fator que deve ser levado em conta, na indexação, é a definição quanto à coordenação dos termos de índice, sendo possível duas abordagens: pré-coordenação e pós-coordenação. Com as tentativas de recuperação automática, passaram a ser usadas técnicas para pós-coordenação (*postcoordination*) no lugar de pré-coordenação (*precoordination*) [Sparck-Jones1997]. Na pós-coordenação, os termos de índice podem ser combinados em tempo de busca para satisfazer condições lógicas absolutas ou relativas, permitindo que documentos que compartilhem uma determinada combinação de itens de busca possam ser encontrados. Na pré-coordenação, são utilizadas descrições fixas e completas dos documentos e é através destas descrições, sem combinações, que os documentos podem ser encontrados.

A coordenação de conceitos separados, sendo feita em tempo de indexação, implica que as entradas no índice deverão mostrar esta coordenação; por outro lado, a coordenação de conceitos sendo feita em tempo de busca implica que as entradas no índice serão associadas somente aos elementos atômicos.

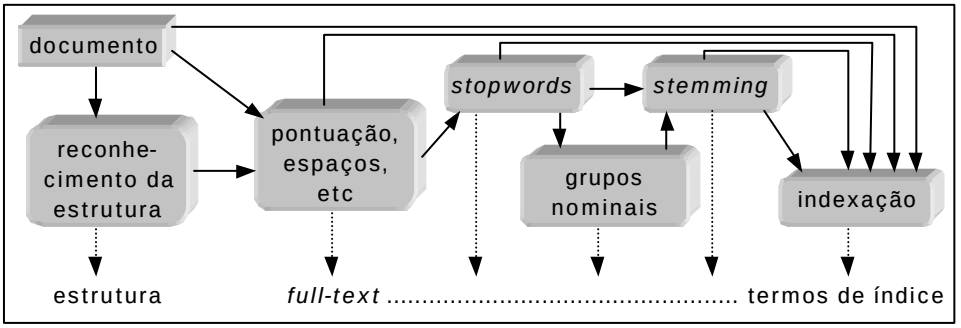

Figura 4. Construção de termos de índice (adaptada de [Baeza-Yates1999])

Esta atividade de construção da visão (ou forma) lógica dos documentos, que é a indexação, também, pode ser entendida a partir da Figura 4, onde observamos as diferentes etapas envolvidas no tratamento da estrutura do texto e na construção dos índices. Há um crescente detalhamento necessário ao considerarmos termos de índice selecionados, para representar o texto dos documentos; mas os passos básicos, comumente utilizados para a indexação de um documento, são: análise léxica, eliminação de *stopwords*, *stemming e* seleção de termos [Baeza-Yates1999].

O reconhecimento da estrutura é um processo que objetiva identificar características da estrutura do documento de interesse, para pesquisas que a utilizam. O termo *full-text* refere-se à representação do texto completo do documento. À medida que temos uma "normalização" seletiva do texto do documento, os termos que comporão o índice têm de ser proporcionalmente mais representativos do que aqueles que são deixados de lado, para que o índice cumpra com seu objetivo.

Entre os processos seletivos estão a eliminação de *stopwords*, a identificação de grupos nominais, que eliminam adjetivos, advérbios e verbos, e o uso de *stemming*.

A eliminação de *stopwords* (e, em geral, a seleção dos termos de índice) tem como principais objetivos diminuir o tamanho do índice e desconsiderar termos sem valor direto de significado, nas operações de busca.

Por outro lado, termos reduzidos por *stemming* podem possibilitar níveis de performance de recuperação comparáveis ou superiores aqueles obtidos por indexação com vocabulário controlado ou por indexação baseada em sintagmas [Sparck-Jones1997].

5.2.2. Formulação da consulta

Uma consulta é a formulação da NI do usuário. As consultas podem ter a seguinte classificação: consulta por palavra-chave, por padrões de comparação (*pattern matching*), orientadas à estrutura, ou em linguagem natural [Baeza-Yates1999].

Na forma mais simples de RI, uma consulta é composta de palavras-chave. Estas palavras-chave serão usadas para assinalar como relevantes os documentos que as contêm. As consultas mais simples e intuitivas são mais utilizadas, por serem mais fáceis de expressar e permitir rápida classificação dos documentos. Podem ser compostas por uma simples palavra ou, em geral, por uma combinação mais complexa de operações envolvendo diversas palavras. Entretanto, a forma mais elementar é consulta por palavra única. Muitos sistemas complementam as consultas por palavra única buscando palavras em um dado contexto ou pela proximidade a outras palavras. Também usam as consultas por frase que, na verdade, são seqüências de consultas por palavra única.

Tabela 2. Tipos de padrão de comparação nas consultas

Tipos de padrão	Observação
palavra	–
prefixo	–
sufixo	–
substring	–
range	um par de palavras que recupera todas aquelas que estão entre elas na ordem lexicográfica, por exemplo: medo – mero iria considerar meio
palavras similares	com pequenas diferenças permitidas
expressão regular	usa união, concatenação e repetição de termos
padrões estendidos	usa classes de caracteres, expressões condicionais, "coringas" (*wild characters*) e combinações

Nas consultas por padrões de comparação, um padrão é definido como um conjunto de características sintáticas que precisam ocorrer em um segmento de texto. Se um segmento de texto satisfaz as especificações do padrão, então ele coincide (*matches*) com o padrão. Os principais tipos de padrão são apresentados na Tabela 2.

As consultas orientadas à estrutura necessitam que os textos dos documentos apresentem alguma estrutura. Exemplo: se o alvo da RI é um conjunto de correspondências, cada uma com destinatário, remetente, data, assunto e corpo de texto, então uma consulta poderia ser feita buscando as correspondências com uma dada pessoa, como remetente, e cujo assunto fosse “futebol”.

Considera-se que as consultas booleanas são uma abstração simplificada das consultas que usam linguagem natural. Novas questões, entretanto, aparecem, quando adota-se realmente a linguagem natural, especialmente em relação ao modo como classificar os documentos com respeito à consulta. O critério de busca pode ser redefinido usando um modelo diferente, permitindo aplicação de PLN. Quando a consulta usa linguagem natural, ou é formulada através de grupamentos de palavras ou frases, seu processamento, de forma semelhante ao que foi visto quanto à indexação (ver Figura 4), envolve as seguintes etapas [Baeza-Yates1999]: análise léxica, eliminação de *stopwords* e *stemming*.

Na etapa da análise léxica, ocorre a conversão de uma cadeia de caracteres (o texto da consulta) em uma cadeia de palavras. Assim, o principal objetivo desta etapa é a identificação das palavras que constituem a consulta. Na fase seguinte, artigos, preposições e conjunções são candidatos naturais à lista de *stopwords*, ou seja, a serem eliminados. Então, pode ser executada a normalização lexical através de *stemming*, pois freqüentemente o usuário especifica uma palavra na consulta, mas somente variações desta palavra estão presentes em um documento relevante.

No contexto da formulação da consulta, uma estratégia que visa otimizá-la é a expansão de consulta, que destacamos a seguir.

5.2.3. Expansão da consulta

O recurso de expansão de consulta passou a ser usado na RI há cerca de 30 anos. Pode ser adotado tanto para tornar maior o conjunto de documentos recuperados, quanto para aumentar a precisão [Moldovan2000]. No primeiro caso, os termos expandidos são selecionados entre aqueles similares aos originais encontrados na consulta. Seriam considerados similares aqueles que possuem significado semelhante, mas nem sempre são sinônimos, como "casa" e "prédio". No segundo caso, os termos adicionados não são similares, porém apresentam algum tipo de relacionamento (como o que ocorre entre "casa" e "morar") com os termos originais, deduzido por motivação lingüística ou através de dados estatísticos.

As abordagens para melhorar a formulação da consulta original através da inclusão de novos termos com reavaliação de seus pesos podem ser classificadas em três categorias [Baeza-Yates1999]: (a) abordagens baseadas em realimentação (feedback) da consulta, com intervenção do usuário; (b) abordagens baseadas em análise local, ou seja, que leva em conta o conjunto de documentos inicialmente recuperados; e (c) abordagens baseadas em análise global, ou seja, que leva em conta a coleção de documentos, tipicamente utilizando thesauri.

Tendo sido formulada e otimizada a consulta, vamos passar aos procedimentos de busca, que tratamos na próxima seção.

5.3. Outras abordagens

Técnicas de indexação também podem considerar expressões complexas com sintaxe e estrutura semântica internas ricas, ou podem utilizar conjuntos de termos compostos ou frases [Sparck-Jones1997]. Também podem ser definidos por métodos que usam análise de proximidade, especialmente de adjacência estrita, apoiados apenas implicitamente em estrutura lingüística. Por outro lado, métodos que utilizam conhecimento semântico também são encontrados, utilizando, por exemplo, a WordNet para melhorar a performance da indexação [Gonzalo1998].

Seja por motivação lingüística, seja por aplicação de métodos estatísticos, há indicações que o uso de vocabulário controlado e a combinação de termos simples e compostos, representando conceitos complexos, tenham utilidade para o processamento do significado.

5.3.1. Representação do conhecimento na RI

A indexação é um processo que produz índices para representar documentos. Em sentido restrito, um termo de índice é uma palavra-chave (ou grupo relacionado de palavras) que tem algum significado próprio, tendo geralmente a semântica de um substantivo [Baeza-Yates1999]. A representação dos documentos através de índices pode acontecer em dois níveis [Smeaton1997]: (a) em nível de palavra, com equivalência entre os termos de índice e as palavra do texto; e (b) em nível conceitual, onde o mapeamento é realizado entre os conceitos, que as palavras ou as frases carregam, e os índices.

Captar esta semântica é a tarefa principal de uma indexação inteligente. Em busca desta qualidade, métodos estatísticos são bastante utilizados, mas a análise lingüística também tem sido indicada para indexação automática de textos já há algum tempo [Haller1986].

Os termos da consulta, por sua vez, tentam representar a comunicação do usuário com o sistema, descrevendo sua NI, de forma que ocorra um mapeamento entre esta NI e a informação contida nos documentos recuperados [Kowalski1997].

5.3.1.1. Algumas experiências com captura de relações sintagmáticas

Levando em conta padrões léxico-sintáticos, Byrd e Ravin [Byrd1999] apresentam um método para identificar e extrair relações lexicais que ocorrem entre conceitos (incluindo entidades e termos técnicos. Primeiramente são identificadas seqüências de palavras que coincidam com estruturas gramaticais específicas. Após, são extraídos e agrupados substantivos, e cada um destes grupos recebem uma referência única. Por exemplo, "*Treasury Secretary Nicholas Brady*", "*Secretary Brady*" e "*Mr. Brady*" são variantes da mesma forma canônica "*Nicholas Brady*".

Jacquemin e Tzoukermann, por sua vez, apresentam o sistema FASTR, uma ferramenta para extração de termos complexos e suas variantes sintáticas e morfo-sintáticas [Jacquemin1999]. A ferramenta utiliza normalização morfológica, resolução da ambigüidade, itiquetagem de categorias gramaticais e análise de sufixos para produzir uma lista de termos compostos para a indexação.

Em outra abordagem, Pearce utiliza restrições na possibilidade de substituição de sinônimos para capturar colocações [Pearce2001]. De acordo com esta abordagem, os sinônimos de uma palavra *w* podem ser classificados em três grupos diferentes, quanto à capacidade de substituir *w* em um determinado contexto: (a) os que substituem normalmente; (b) os que podem eventualmente substituir; e (c) os que não podem substituir, constituindo anti-colocações. São utilizados, por Pearce, os *synsets* da WordNet para esta classificação.

Kahane e Polguère propõem funções lexicais para representar colocações [Kahane2001]. Consideradas, nesta abordagem, relações orientadas, as colocações são modeladas através de funções lexicais, que denotam um conjunto de pares de itens lexicais, associados por uma correspondente relação lexical. São propostas funções

como “Caus(arg1,arg2)”, que significa que “arg1” causa “arg2”, e “Fact(arg)”, que significa que “arg” funciona.

5.3.1.2. Algumas experiências com captura de relacionamentos paradigmáticos

Berland e Charniak propõem a captura de relacionamentos de meronímia entre termos considerando diversos padrões [Berland1999], como “S_1 *of the* S_2”, e “S_1 *in a* S_2”, onde S_i é substantivo.

Morin e Jacquemin [Morin1999] capturam relacionamentos de hiperonímia entre termos considerando o padrão “SN *such as* LISTA”, onde “SN” é um sintagma nominal[4] e “LISTA” é uma lista de sintagmas nominais. Neste caso, cada componente de “LISTA” seria um hipônimo e “SN” seria o hiperônimo.

Hearst, além do padrão “*such as*”, já havia indicado o uso de outros padrões léxico-sintáticos que incluem expressões como “*or other*”, “*and other*”, “*including*” e ”*specially*”, para capturar a hiponímia [Hearst1992].

Sanderson e Dawn apresentam princípios básicos para a construção de hierarquia de conceitos, resumidos a seguir [Sanderson2000a]. Os termos A e B, participantes da hierarquia a ser construída, são aqueles que melhor refletem os tópicos cobertos pelos documentos que constituem o *corpus*. Se os documentos onde B ocorre constituem um subconjunto dos documentos onde A ocorre, então A subsume B. Isto quer dizaer que A está relacionado a um conceito mais geral, sendo considerado hierarquicamente superior a B.

Bouillon e co-autores utilizam a teoria do Léxico Gerativo (mais especificamente as relações paradigmáticas viabilizadas pela estrutura Qualia) para controlar a captura de informação relacionada a pares de itens lexicais do tipo substantivo-verbo [Bouillon2000]. Por exemplo, o substantivo “disco” estaria relacionado, através da estrutura Qualia, aos verbos “acessar”, “apagar”, “inicializar” e “gravar”. É proposto um método de extração automática de tais informações.

5.3.1.3. Algumas experiências com a Teoria do Léxico Gerativo

O CORELEX (*Core Lexical Engine*) [Pustejovsky1997, Buitelaar1998] é um thesaurus fundamentado na teoria do léxico gerativo (LG) [Pustejovsky1995]. Oriunda do CORELEX, a Lexical Web [Pustejovsky1997], uma estrutura interligada de entradas de índice geradas automaticamente, é utilizada no sistema TexTract [Pustejovsky1997, Cooper2000]. Este sistema é uma ferramenta para indexação semântica automática, projetada para executar análise sintática, indexação e geração de descritores (*hyperlinking*) de documentos em formato digital.

Com uma abordagem voltada para a língua Portuguesa e usando especificamente a estrutura Qualia da teoria do Léxico Gerativo, Santa Maria Fº., Chishman e Lima [SantaMaria1999] propõem um modelo para indexação de documentos que busca resolver problemas não solucionáveis em redes IS-A. É utilizada uma hierarquia de conceitos como estrutura de índices, mantendo relações semânticas entre si.

---

[4] Frase que tem valor de substantivo.

5.3.2. Normalização lingüística

A normalização de termos pode contribuir significativamente para RI, pois as abordagens lingüísticas têm levado vantagem sobre as técnicas de *stemming* [Voorhees1999]. Uma linguagem de indexação com motivação lingüística deve incluir técnicas sintáticas e semânticas para identificar termos chave, e deve considerar especificamente uma linguagem de controle para gerar índices normalizados [Sparck-Jones1999a]. Neste sentido, os analisadores sintáticos na RI são geralmente utilizados para extrair estruturas com elevado conteúdo de informação [Jacquemin1999].

Em virtude da diversidade de abordagens, Krovetz experimentou diferentes rotinas relacionadas à normalização de variantes de itens lexicais [Krovetz1997]: (a) agrupamento somente de variantes por flexão (plurais e flexões de formas verbais); (b) agrupamento de variantes por flexão e por derivação (com sufixos "ado" e "ente", por exemplo); e (c) uso do *stemmer* Porter [Frakes1992]. Estas estratégias alternativas para normalização trazem, para a RI, benefícios significativos mas, entretanto, eles dependem da coleção de documentos. Os benefícios são maiores para documentos com tamanho menor porque há maior probabilidade de poucas variantes morfológicas de um item lexical estarem presentes em cada documento. Neste caso, outras variantes presentes em uma consulta não seriam encontradas. Quanto ao algoritmo de Porter, os benefícios são muito pequenos se ele não tiver o apoio de um léxico para eliminar falsos agrupamentos morfológicos ao reduzir palavras em radicais.

A normalização de itens lexicais está relacionada a duas preocupações da RI: a determinação de similaridade entre termos, ao comparar variações lingüísticas dos mesmos, e a economia de índices.

5.3.2.1. Algumas experiências com normalização

Storb e Wazlawick enfrentam o problema da variação morfológica de termos, ao propor um modelo de RI para a língua Portuguesa [Storb1998]. Estes autores consideram que a similaridade entre os significados das palavras, através da comparação de possíveis radicais, é determinada pelo reconhecimento correto de radicais e sufixos. Neste sentido, para cada par radical-sufixo, no modelo proposto, é calculado um grau de certeza (entre 0 e 1). Este cálculo leva em conta a combinação das certezas obtidas para cada componente do par, ou seja, para o radical e para o sufixo.

Corston-Oliver e Dolan propõem um método para eliminar termos de índices que ocorrem como "cláusulas subordinadas" [Corston-Oliver1999]. É executada análise de dependência sobre a estrutura do texto, buscando identificar a idéia central das proposições contidas nele, reconhecendo a porção marginal como cláusulas subordinadas. Com o método utilizado, os autores relatam a redução de aproximadamente 30% do tamanho dos índices, sem prejuízo da performance do sistema de RI.

5.3.3. Composição de termos

O uso de termos compostos é uma estratégia importante, porém pouco entendida, com benefícios não consistentes à RI [Krovetz1997]. A expressão "termo composto" é utilizada, nesta seção, para identificar indistintamente frases ou termos constituídos por mais de um item lexical (*multi-word terms*).

Em geral, os termos compostos englobam seqüências de palavras ricas em significado, com menor ambigüidade que os itens lexicais simples e, por isso mesmo, permitem uma aproximação maior com o conteúdo do texto onde ocorrem [Dias2000]. Esta representação mais rica permite uma RI mais efetiva [Smeaton1997].

Algumas vezes os termos formados por itens lexicais que co-ocorrem em um *corpus*, sejam ou não vizinhos imediatos, são incluídos na classe dos termos compostos. Entretanto, duas categorias de termos compostos estão mais fortemente relacionados aos procedimentos com motivação lingüística [Krovetz1997]: (a) os *syntactic phrases*, cujos componentes possuem algum tipo de relacionamento sintático, exigindo processamento sintático para serem tratados, como é o caso de "ações contra o terrorismo"; e (b) os *lexical phrases* (um subconjunto dos anteriores), que podem ser exemplificados tipicamente através de nomes próprios ou conceitos técnicos, e seus componentes ocorrem em ordem fixa, como é o caso de "Estados Unidos".

É necessário salientar que os nomes próprios exigem um tratamento especial, inclusive no cálculo de similaridade entre termos [Friburger2002].

Arampatzis e co-autores defendem a importância de se utilizar termos compostos na RI, mais especificamente os sintagmas nominais [Arampatzis1998]. Justificam-no pelo papel central que os sintagmas nominais desempenham nas descrições sintáticas em linguagem natural, e por serem considerados, na IA, como referências ou descritores de conceitos complexos.

Em razão da maior complexidade em relação aos termos atômicos, os termos compostos requerem cuidados especiais. Enquanto a normalização dos termos atômicos é feita usualmente através de *stemming*, a normalização dos termos compostos pode ser [Smeaton1997]: (a) ignorada, com a manutenção do formato original, o que obriga a inclusão de variantes; (b) realizada pela comparação com versões de conjuntos de palavras contidos em léxico, o que possibilita a diminuição do vocabulário e, sendo utilizado o mesmo procedimento para consultas e documentos, permite a comparação palavra-a-palavra; ou (c) obtida através da representação dos termos compostos em formato estruturado, que permite a derivação de múltiplas interpretações semânticas.

Enfrentar o problema dos termos compostos ou, mais especificamente, dos substantivos compostos, através do PLN aplicado com conhecimento lingüístico, permite algumas vantagens para o correto entendimento do significado que eles carregam. Estas vantagens são [Clark2000]: (a) melhor normalização de variações de itens lexicais (por exemplo, ao reconhecer "antisubmarine", "anti-submarine" e "anti submarine", em inglês, como variantes do mesmo conceito); (b) auxílio à resolução da ambigüidade de um componente, baseada nos outros componentes do termo; (c) auxílio à identificação apropriada de sub-grupos de palavras incluídos em um substantivo composto; e (d) refinamento do papel desempenhado pelos componentes (por exemplo, "casa de madeira" e "casa de campo", onde "madeira" é um constituinte, enquanto "campo" indica localização ou tipo).

5.3.3.1. Algumas experiências com composição de termos

Kaji e co-autores propõem uma estratégia para a extração de substantivos (compostos ou não) [Kaji2000]. Um sintagma nominal, algumas vezes, pode ser interpretado como $W_1(W_2W_3)$ ou como $(W_1W_2)W_3$. Exemplo: casa de bairro grande, pode ter o adjetivo grande modificando casa ou bairro. Se o componente W_2W_3 ocorre

mais freqüentemente, então a estrutura $W_1(W_2W_3)$ será a preferida; em caso contrário $(W_1W_2)W_3$ será a escolhida.

Dias e co-autores apresentam um método para extrair unidades lexicais complexas utilizando a medida de expectativa mútua e o algoritmo LocalMarxs [Dias2000]. É analisada cada seqüência de texto A com n unidades lexicais, que contém a seqüência B, com n-1 unidades, e está contido na seqüência C, com n+1 unidades. A é selecionada se apresentar grau de associação entre seus constituintes superior aos de B e C. O objetivo desta abordagem é melhorar a indexação através de aquisição de informação mais eficiente.

5.3.4. Resolução de ambigüidade

Têm sido defendidas como verdadeiras três hipóteses quanto à ambigüidade lexical na RI [Krovetz1997]:

- Hipótese 1. A resolução da ambigüidade lexical beneficia a performance da RI.
- Hipótese 2. Os significados das palavras determinam uma separação efetiva entre os documentos relevantes e não relevantes.
- Hipótese 3. Mesmo em coleções pequenas de documentos de domínio específico, há uma proporção significante de ambigüidade lexical.

A resolução automática de ambigüidade constitui um problema reconhecidamente difícil [Smeaton1997]. As abordagens para a resolução de ambigüidade na RI podem ser divididas em duas categorias principais: baseadas em regras de co-ocorrência ou de padrões sintáticos, ou baseadas em informações oriundas de *corpora* ou dicionários (ou thesauri) [Sanderson2000]. Exemplos são apresentados a seguir.

5.3.4.1. Algumas experiências com resolução de ambigüidade

Gauch e Futrelle usam uma combinação de informação mútua e de informações de contexto para estabelecer similaridades entre itens lexicais e definir classes de palavras [Gauch1994]. Estas classes são utilizadas para resolver ambigüidades de palavras da língua inglesa terminadas em *ed*, indicando se são verbos no particípio passado ou adjetivos.

Krovetz [Krovetz1997] considera informações provenientes de dicionários, como morfologia, categoria gramatical e composição de termos, como fontes múltiplas de evidência para a resolução de ambigüidade. As diferenças das palavras na forma são consideradas associadas às diferenças em significados e, em virtude disto, deve-se estabelecer associações entre tais variações. Para atacar o problema, é explorada a presença de variantes de um termo na definição deste termo no dicionário, além de serem utilizadas sobreposições de palavras em definições supostamente variantes.

5.3.5. Identificação de temas

Um tema (assunto ou tópico) é uma proposição tratada ou discutida em um texto. Um documento possui informação contida em mais de um tópico e apenas um deles ou a combinação deles pode tornar o documento relevante a uma dada consulta [Smeaton1997]. Portanto, identificar temas deveria ser uma tarefa inerente à RI.

Sob a expressão “identificação de tema” caracterizamos, aqui, o problema de definir hierarquias para os assuntos (ou tópicos) tratados em um texto (de um

documento ou de uma consulta) [Gonzalez2001a]. Portanto, este problema engloba não apenas o estabelecimento do tema, mas também dos sub-temas. Algumas estratégias de RI envolvidas com a identificação de tema são relatadas a seguir.

5.3.5.1. Algumas experiências com identificação de temas

Abordando o problema de indexação na RI, Loukachevitch e co-autores [Loukachevitch1999] propõem uma representação temática de cada documento como uma estrutura hierárquica de nodos conceituais. A abordagem pressupõe que cadeias lexicais, que caracterizam o tema principal de um texto, normalmente possuem elementos que ocorrem juntos nas sentenças, com mais freqüência que outros elementos de outras cadeias lexicais. Também é considerado que temas mais gerais podem ser descritos em função de temas mais específicos, recursivamente.

Shatkay e Wilbur apresentam uma estratégia probabilística baseada em temas (ou tópicos) para RI [Shatkay2000]. É levada em conta a dualidade que estabelece que um tema pode ser caracterizado tanto através dos documentos que o discutem, quanto através dos termos que o descrevem.

Kim, Lu e Raghavan propõem uma abordagem chamada *Rule Based Information Retrieval by Computer* (RUBRIC) [Kim2000]. São produzidas regras para capturar os tópicos da consulta, representando-a no formato de árvore. Três tipos de regras são projetados para especificar conceitos específicos, genéricos e possíveis, com grau de confiança associado. Um thesaurus clássico é utilizado para prover conceitos e relacionamentos como “termo específico”, “termo genérico” e “termo relacionado”.

Contreras e Dávila formulam uma proposta para estabelecer descritores de documentos, tomando como referência uma gramática de estilos e formas lógicas [Contreras2001]. O tópico (ou tema) comum mais específico é sempre considerado como descritor relevante. As regras de extração destes tópicos expressam recomendações de estilo, onde são identificados agentes e ações (verbos) e são extraídos os tópicos.

5.3.6. Associação de termos

Nesta seção, utilizaremos a expressão “associação de termos” para definir o relacionamento que existe entre um termo original de uma consulta e um termo agregado a ela por expansão, visando melhor performance na RI. Discutiremos, a seguir, algumas abordagens sintáticas, sintático-semânticas e semânticas para expansão de consulta, com o uso de associação de termos. Basicamente, as abordagens sintáticas enfatizam os relacionamentos sintagmáticos entre as palavras e as abordagens semânticas levam em consideração, principalmente, os relacionamentos paradigmáticos. As abordagens sintático-semânticas consideram esses dois tipos de relacionamentos.

5.3.6.1. Algumas experiências com associação de termos

A) Abordagens sintáticas

Jing e Croft propõem a construção automática de um thesaurus de associação dependente de *corpus* [Jing1994]. Esta abordagem permite o uso de consultas em linguagem natural no sistema de RI chamado INQUERY. A associação entre termos e sintagmas é obtida pelo programa chamado PhraseFinder, que constrói automaticamente o thesaurus de associação, a partir de textos etiquetados com categorias gramaticais. A expansão de consulta é feita obtendo-se, através do PhraseFinder, uma lista classificada

de sintagmas, identificados como duplicatas ou não-duplicatas. São considerados duplicatas os sintagmas cujos termos também ocorrem na consulta. Um exemplo de duplicata é o sintagma veículo de fibra-de-vidro em relação à consulta construção de veículos usando fibra-de-vidro. Podem ser adicionados à consulta original somente os sintagmas duplicatas, somente os não duplicatas, ou todos os sintagmas.

Mandala, Takunaga e Tanaka procuram enriquecer o WordNet para realizar expansão de consulta [Mandala1999]. Eles usam, além da WordNet, o thesaurus de Roget, e um thesaurus dependente de *corpus*, baseado em co-ocorrência e em propriedades sintáticas, para o cálculo de similaridade entre duas palavras w_1 e w_2. Em relação a WordNet, a similaridade depende do número de nós no caminho de w_1 a w_2, e da profundidade máxima da taxonomia. Em relação ao Thesaurus de Roget, a similaridade depende do número de palavras comuns pertencentes às categorias onde estão w_1 e w_2. Em relação ao thesaurus de domínio específico, é usada uma medida de informação mútua para calcular a similaridade. É adotada a média das similaridades obtidas com os diversos thesauri.

A abordagem de Strzalkowski e Carballo baseia-se na co-ocorrência, em contexto sintático [Strzalkowski1999a]. As consultas são comparadas diretamente no texto do documento, analisando sentenças inteiras, ou parágrafos ou outras seqüências de itens lexicais. São considerados experimentos com expansão manual e automática. No procedimento automático, os 100 documentos do topo da lista dos recuperados pela consulta original são utilizados. No procedimento manual são considerados os 10 primeiros documentos. Eles são pesquisados procurando parágrafos que possam conter ocorrências (ou variantes) dos conceitos presentes na consulta original, não importando se são documentos relevantes ou não. Os trechos selecionados são copiados literalmente na consulta, com pesos diferenciados para os termos originais.

B) Abordagens sintático-semânticas

O sistema experimental IRENA é utilizado para melhorar os resultados de precisão e resposta na RI [Arampatzis1997]. O sistema de busca recebe um consulta na forma de sintagmas nominais. A partir de palavras-chave (extraídas por análise sintática) normalizadas são geradas variantes morfológicas. É adotada uma base de dados de sinônimos para melhorar a expansão lexical. O léxico utilizado foi construído a partir do WordNet.

No sistema de RI proposto por Moldovan e Mihalcea, o módulo de expansão de consulta possui duas funções principais: a construção de listas de similaridades usando a WordNet e a geração da consulta expandida [Moldovan2000]. Entre os sinônimos, são identificadas palavras semanticamente similares àquelas da consulta original. São analisadas alternativas de sentidos para pares verbo-substantivo encontrados na consulta. Usando operadores relacionais, é construída a consulta expandida, sendo que o operador OR liga as palavras similares a um termo da consulta e o operador AND conecta as listas obtidas para cada termo da consulta.

C) Abordagens semânticas

ExpansionTool [Järvelin1996] é uma ferramenta para expansão de consulta baseada em um modelo de dados baseado em três níveis de abstração, usados para representar o vocabulário do *corpus*: (a) nível conceitual, onde são representados conceitos e relacionamentos; (b) nível lingüístico, onde são associadas expressões em

linguagem natural a conceitos; e (c) nível de ocorrência, onde cada ocorrência de expressão no texto está associada a um ou mais modelos de comparação (por exemplo: com ou sem termos compostos, ou com ou sem stemming).

Gonzalez e Lima utilizam o thesaurus T-Lex, baseado na estrutura Qualia da teoria do Léxico Gerativo, como recurso para expansão de consulta [Gonzalez2001c, Gonzalez2001d]. Itens lexicais de categorias gramaticais diferentes possuem estruturas Qualia específicas. A estruturação semântica do thesaurus e o uso de operações gerativas permitem a expansão automaticamente. A seleção dos novos termos e o cálculo de seus pesos dependem da sobreposição da expansão e do nível de profundidade em que se avança, no thesaurus, na busca dos termos descritores.

6. Estudo de caso

Neste estudo de caso é apresentado um mecanismo de busca batizado informalmente como “Yahinho”.

O Yahinho foi projetado para pesquisar um *corpus* constituído por 34 resumos de dissertações do Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Faculdade de Informática da PUCRS. Estes documentos, ao todo, contêm 7095 palavras e, em média, apresentam 208 palavras cada um.

6.1. Indexação

A coleção de documentos a ser pesquisada foi indexada utilizando-se a estratégia de arquivo invertido. A Figura 5 apresenta um trecho do arquivo invertido gerado.

Na seleção dos termos foram eliminadas as *stopwords*. O dicionário é formado por termos de índice em *stem*. Uma versão alternativa foi construída com termos no formato canônico, ou seja, normalizados morfologicamente no infinitivo (no caso de verbos) ou masculino singular (no caso de substantivos e adjetivos).

```
abandon 3100
abert 1900
abertur 0700
abord 0102050015001602180019002000240331003201
abrang 0700
acab 010210011201
acadêm 04001101
aceit 23013300
acerv 0502
acess 050120013402
acompanh 04002500
acord 0102090114011700
adapt 060015001800200022012500 2800
adequ 07001700
adicion 23013400
adjacente 1900
adjun 0201
administr 0700
adot 100131003400
afet 1600
```

Figura 5. Trecho do arquivo invertido

Os endereçamentos contêm a identificação do documento, onde o termo ocorre, e seu peso. O peso de um termo é a sua freqüência (percentagem de ocorrência,

excluídas as *stopwords*) neste documento. Na Figura 5 aparece, por exemplo, o termo "aceit" e ao lado dele o endereçamento "23013300". Isto significa que este termo ocorre no documento 23 com uma freqüência de 01% (de 1 a menos que 2%) e no documento 33 com uma freqüência de 00% (menos que 1%). Nos outros documentos o termo "aceit" não ocorre.

O mecanismo de busca foi desenvolvido em duas versões (0.0 e 1.0), respectivamente, sem e com expansão de consulta, e podem ser encontrados nos seguintes endereços:

- www.inf.pucrs.br/~gonzalez/ri/yahinho/ e
- www.inf.pucrs.br/~gonzalez/ri/yarinho1.0/

6.2. Busca e classificação

O Yahinho, na versão 0.0, utiliza o seguinte procedimento para selecionar e classificar os documentos relevantes a uma dada consulta. Vamos considerar, como exemplo, a consulta "orientação a objetos".

(i) Eliminação de *stopwords*: a consulta é reduzida a "orientação objetos";

(ii) *Stemming*: a consulta é reduzida a "orient objet";

(iii) Busca: cada termo da consulta é pesquisado no índice;

(iv) Classificação: é calculada a similaridade do documento em relação à consulta, para determinar sua classificação por relevância;

(v) Apresentação dos resultados: é apresentada a relação dos documentos selecionados e classificados, para que o usuário escolha o(s) que deseja acessar.

Para selecionar e classificar os documentos recuperados, é calculada a similaridade entre a consulta corrente e cada documento. A similaridade (**S**) de um documento $\mathbf{d}_i$ em relação a uma consulta **q** é calculada da seguinte maneira:

$$\mathbf{S}(\mathbf{q}, \mathbf{d}_i) = \sum_{0 \le j \le |\mathbf{q}|} \left(\sum_{0 \le i \le |\mathbf{D}|} (\mathbf{p}_{ji}) \right),$$

onde:

$|\mathbf{q}|$ é o número de termos na consulta;

$|\mathbf{D}|$ é o número de documentos no *corpus*; e

$\mathbf{p}_{ji}$ é o peso do termo $\mathbf{t}_j$ no documento $\mathbf{d}_i$.

Considerando a consulta "orientação a objetos", o peso de cada documento recuperado é obtido pelo cálculo de similaridade, resultado da soma, neste caso, dos pesos dos termos "orient" e "objet", contidos nos documentos.

6.3. Busca e classificação com expansão de consulta

O Yahinho, na versão 1.0, utiliza praticamente o mesmo procedimento da versão 0.0. para selecionar e classificar os documentos relevantes, após a consulta ser expandida [Gonzalez2001c, Gonzalez2001d]. A expansão de consulta é realizada com a utilização de um thesaurus (T-Lex), construído parcialmente contendo termos do contexto da aplicação. Ou seja, foram inseridos alguns termos da área de sistemas de informação, a mesma do *corpus* utilizado.

O thesaurus T-Lex possui uma estruturação semântica projetada para implementar relacionamentos lexicais, levando em conta aspectos da teoria do Léxico Gerativo. Cada item lexical α, no T-Lex, pode apresentar alguns dos seguintes relacionamentos (ou papéis):

- formal, indicando hiponímia de α;
- constitutivo, indicando meronímia de α;
- agentivo, indicando o responsável pela existência de α, ou a causa ou condição da criação ou da inicialização de α, ou o que é necessário para α ocorrer; e
- télico, indicando a função ou o próposito de α, ou o que é conseqüência da ocorrência de α.

Um item lexical α, então, está relacionado a outros itens lexicais (descritores de α) através de um destes relacionamentos. Os termos obtidos na expansão de um termo α, de uma consulta, são selecionados entre os descritores de α contidos no T-Lex.

Na tela de entrada do Yahinho, versão 1.0, é possivel selecionar quais os relacionamentos do thesaurus que serão utilizados na expansão, qual o nível máximo de expansão e se os verbos serão utilizados como termos expandidos.

Para obter a classificação dos documentos recuperados, no cálculo de similaridade destes documentos com a consulta, cada termo desta tem peso 1 na versão 0.0. Na versão 1.0, o peso de cada termo de uma consulta $\mathbf{q}$, é calculado da seguinte forma:

- termo original: peso = 1
- termo expandido: peso = **v**/**e**/**n**, sendo:
  v = número de termos de onde o termo expandido se originou,
  e = nível de expansão (1 para a expansão direta de um termo original, 2 para a expansão de um termo expandido de um original, e assim por diante), e
  n = número de termos originais de $\mathbf{q}$.

6.4. Avaliação

Na avaliação comparativa das versões do Yahinho, a versão 1.0, em geral, apresentou resultados superiores. A Figura 6 mostra o gráfico Precisão/Resposta correspondente duas versões.

As curvas Precisão/Resposta do gráfico da Figura 6 são das seguintes configurações:

- ZERO: versão 0.0, sem expansão;
- CT1: versão 1.0, com expansão utilizando os papéis constitutivo e télico, até o nível 1;
- CAT1: versão 1.0, com expansão utilizando os papéis constitutivo, agentivo e télico, até o nível 1; e
- CAT3: idem CAT1 até o nível 3.

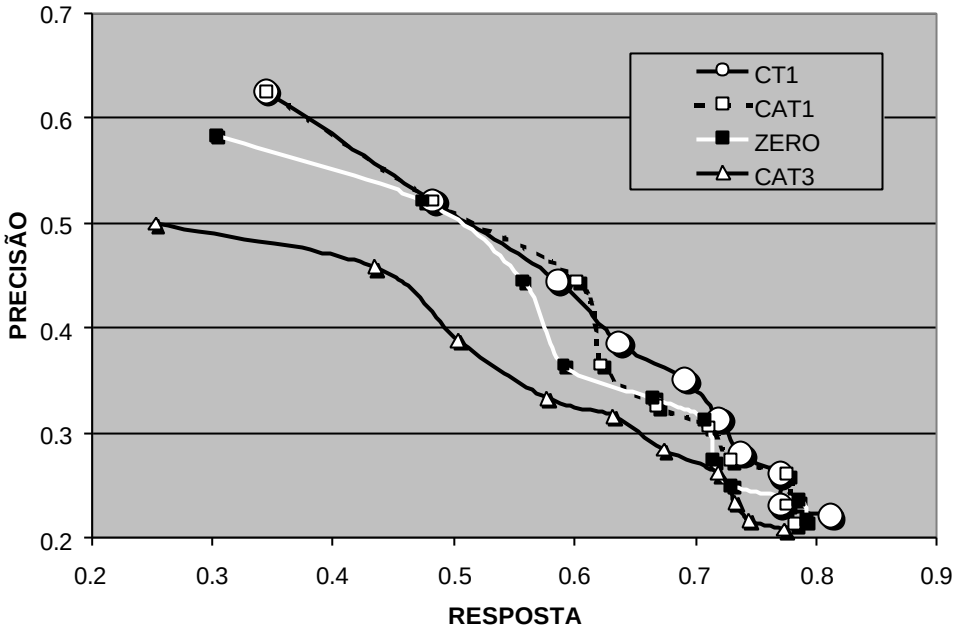

Figura 6. Gráfico Precisão / Resposta

A curva de melhor comportamento é a da configuração CT1. A pior é a da configuração CAT3, provavelmente em razão da expansão até o nível 3. As duas configurações da versão 1.0, com expansão até o nível 1, foram superiores à versão 0.0.

7. Considerações finais

Os dois grandes recursos de que dispõe o PLN são o conhecimento lingüístico e a estatística. O ideal é que ambos estejam presentes. Justamente, uma das dificuldades enfrentadas por estratégias baseadas apenas em métodos estatísticos, em RI, é a complexidade inerente ao processamento lingüístico, o que causa, por exemplo, a falta de esquemas adequados para estabelecimento de pesos para termos compostos [Voorhees1999]:

Há perdas para a RI, quando as abordagens são puramente estatísticas [Allen2000], pois nenhum esforço é feito em nível de processamento semântico. Também há a possibilidade de falha nas comparações simples entre itens lexicais devido a problemas de ambigüidade e de composicionalidade dos termos.

Por outro lado, abordagens puramente lingüísticas também encontram dificuldades [Voorhees1999, Allen2000]. Isto ocorre porque: (a) o conhecimento lingüístico é muitas vezes aplicável em domínios específicos, não sendo portável; (b) o sucesso da RI depende de propriedades da consulta que dispensam algumas técnicas lingüísticas, essencialmente por serem, em geral, constituídas por poucos termos ou por termos independentes sem estrutura sintática; (c) os erros, quando ocorrem, trazem prejuízos que não são compensados pelos benefícios decorrentes da sua aplicação; e (d) as técnicas lingüísticas (como etiquetagem de categorias gramaticais, resolução de ambigüidade, análise sintática, etc.) necessitam ter alto grau de precisão para trazer benefícios.

Muitas vezes, as técnicas não lingüísticas já exploram implicitamente o conhecimento lingüístico que, ao ser agregado, na verdade, relativamente pouca contribuição terá [Voorhees1999].

As variações lingüísticas também explicam as limitações impostas à aplicação do conhecimento lingüístico à RI, em nível semântico [Smeaton1997]: (a) diferentes palavras podem assumir o mesmo significado, como "sapato" e "calçado"; (b) frases com as mesmas palavras em ordens diferentes podem possuir significados diferentes, como "vítima juvenil de crime" e "vítima de crime juvenil"; e (c) o mesmo item lexical pode assumir diferentes significados em contextos diferentes, como "agudo" na medicina e na geometria.

Estes problemas também afetam as abordagens estatísticas, mas com menor repercussão porque é ao conhecimento lingüístico que é atribuída a tarefa de resolvê-los. O PLN aplicado à RI deve procurar aproximar da prática a teoria lingüística, e esta aproximação deve ser feita da forma mais perfeita possível, para trazer vantagens [Voorhees1999].

Sem abandonar os métodos estatísticos, há concretamente dois motivos para que os pesquisadores busquem aplicar PLN com conhecimento lingüístico à RI: (a) as abordagens estatísticas ainda não resolveram satisfatoriamente o problema da RI; e (b) a RI trata inerentemente de processamento de textos, e esta característica faz com que o conhecimento lingüístico não possa ser descartado.

Mesmo que a RI ainda deva esperar muito do PLN motivado por conhecimento lingüístico, a lingüística computacional, em contrapartida, tem encontrado na RI um campo experimental que tem possibilitado evoluções significativas. Isto se deve à exigência de grande precisão nos procedimentos aplicados, incentivando aprimoramentos.

O futuro aponta para a necessidade de maior interação entre as abordagens estatísticas e lingüísticas. Para citar exemplos, de um lado, o cálculo da freqüência de co-ocorrência de itens lexicais pode capturar informações importantes sobre composição de termos; de outro, a captura de padrões léxico-sintáticos levando em conta a identificação de categorias gramaticais pode dar qualidade tanto à representação do textos na indexação, quanto ao cálculo de similaridade na busca.

A questão da RC e, embutida nela, o processamento do significado são pontos cruciais que afetam a performance da RI. Se, com a intenção de atender bem as NIs de seus usuários, um sistema pretender: (a) representar adequadamente documentos e consultas, através de uma forma lógica coerente, e (b) realizar busca, utilizando associações de termos e cálculos de similaridades semanticamente qualificados,

então não poderá dispensar o conhecimento lingüístico por uma simples razão: trata com textos e, portanto, com linguagem natural.

Referências bibliográficas

[Allen1995] Allen, J. *Natural Language Understanding*. Redwood City, CA: The Benjamin/Cummings Pub. Co., 1995. 654 p.

[Allen2000] Allen, J. Natural Language Processing for Information Retrieval. Seattle, Washington: tutorial apresentado em NAACL/ANLP Language Technology Joint Conference, 2000.

[Arampatzis1997] Arampatzis, Avi; Tsoris, C.H.; Koster, C.H.A. IRENA: Information Retrieval Engine Based on Natural Language Analysis. RIAO'97 – Computer-Assisted Information Searching on Internet, McGill University, Montreal, Canadá, 1997. p.159-175.

[Arampatzis2000] Arampatzis, A.T.; Van Der Weide, T. P.; Koster, C.H.A.; Van Bommel, P. Linguistically-motivated Information Retrieval. Encyclopedia of Library and Information Science, V.69, 2000. p.201-222.

[Baeza-Yates1999] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. New York: ACM Press, 1999. 513 p.

[Beardon1991] BEARDON, C.; LUMSDEN, D.; HOLMES, G. *Natural Language and Computational Linguistics*. Melksham-Wiltshire, England: Ellis Horwood Ltd., 1991.

[Berland1999] Berland, M.; Charniak, E. Finding Parts in Very Large *Corpora*. ACL'99, Maryland, 1999. p57-64.

[Bick1998] Bick, Eckhard. Structural Lexical Heuristics in the Automatic Analysis of Portuguese. 11th Nordic Conference on Computational Linguistics, Copenhagen, 1998. p.44-56.

[Bod1995] Bod, Rens. Enriching Linguistics with Statistics: Performance Models of Natural Language. Tese de doutorado. Institute for Logic, Language and Computation (ILLC), Universidade de Amsterdã, 1995.

[Bouillon1998] Bouillon, Pierrette. *Traitement Automatique des Langues Naturelles*. Bruxelas, Paris: Aupelf-Uref, Editions Duculot, 1998. 245 p.

[Bouillon2000] Bouillon, P.; Fabre, C.; Sébillot, P.; Jacquemin, L. Apprentissage de Ressources Lexicales pour l'Extension de Requêtes. In: Jacquemin, Christian (editor). *Traitement Automatique des Langues pour les Recherche d'Information*. Hermès Science Publications, Paris, 2000. p.367-393.

[Buitelaar1998] Buitelaar, Paul. CoreLex: An Ontology of Systematic Polysemous Classes. FOIS'98 – International Conf. on Formal Ontology in Inf. Systems, Trento, Itália, 1998.

[Brown1992] Brown, Peter F.; Della Pietra, Vicent J. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, v.18, n.4, dezembro 1992. p.467-479.

[Byrd1999] Byrd, Roy J.; Ravin, Yael. Identifying and Extracting Relations in Text. NLDB'99 – 4th Int. Conf. on Applications of Natural Language to Information Systems. Fliedl, G.; Mayr, H.C. (editores) OCG Schriftenreihe (Lectures notes), V.129, 1999. p.149-154.

[Clark2000] Clark, P.; Thompson, J.; Holmback, H.; Duncan, L. Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search. 12th Conf. on Innovative Applications of AI (AAAI / IAAI'2000), 2000. p.988-995.

[Contreras2001] Contreras Z., H.Y.; Dávila Q., J.A. Procesamiento del Lenguaje Natural basado en una "Gramática de Estilos" para el Idioma Español. CLEI'2001, Mérida, 2001. CD-ROM.

[Corston-Oliver1999] Corston-Oliver, S.H.; Dolan, W. B. Eliminating Index Terms from Subordinate Clauses. 37th Annual Meeting of the ACL. Maryland, Univ. of Maryland, 1999. p.349-356.

[Croft2000] Croft, W. B. (Editor) Advances in Information Retrieval. London: Kluwer Academic Publishers, 2000

[Date1991] Date, C.J. *Introdução a Sistemas de Banco de Dados*. Rio de Janeiro: Editora Campus, 1991. 674 p.

[Davis1993] Davis, R.; Shrobe, H.; Szolovits, P. What is a Knowledge Representation? *AI Magazine*, V.14, N.1, 1993. p.17-33.

[Dias2000] Dias, Gaël; Guillore, Sylvie; Bassano, Jean-Claude; Lopes, José Gabriel P. Extraction Automatique d'Unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire. In: Jacquemin, Christian (editor). *Traitement Automatique des Langues pour les Recherche d'Information*. Hermès Science Publications, Paris, 2000. p.447-493.

[Evens1992] Evens, M. W. (Editora) *Relational Model of the Lexicon: Representing Knowledge in semantic networks*, New York: Cambridge University Press, 1992. p.41-74.

[Fellbaum1998] Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998. 423 p.

[Franconi2001] Franconi, Eurico. Description Logics for Natural Language Processing. In: Baader, F.; McGuinness, D. L.; Nardi, D.; Patel-Schneider, P. P. (editores) *Description Logics Handbook*. Cambridge: Cambridge University Press, 2001. Cap.18.

[Frakes1992] Frakes, W. B.; Baeza-Yates, R. (editores). *Information Retrieval – Data Structures & Algorithms.* New Jersey: Prentice-Hall, 1992.

[Frantz1997] Frantz, V.; Shapiro, J.; Voiskunskii, V. *Automated Information Retrieval: Theory and Methods*. San Diego, CA: Academic Press, 1997. 365 p.

[Friburger2002] Friburger, N.; Maruel, D. Textual Similarity based on Proper Names. ACM SIGIR 2002 Conference – Mathematical / Formal Methods in Information Retrieval, Finlândia, 2002.

[Gamallo2002] Gamallo, Pablo; Gonzalez, M.; Agustini, A.; Lopes, G; Lima, Vera L. S. Mapping Syntactic Dependencies onto Semantic Relations. ECAI'02, Workshop on Natural Language Processing and Machine Learning for Ontology Engineering, Lyon, France, 2002. p15-22.

[Gauch1994] Gauch, Susan; Futrele, Robert. Experiments in Automatic Word Class and Word Sense Identification for Information Retrieval. 3th Annual Symposium on Document Analysis and Information Retrieval, 1994. p.425-434.

[Genesereth1988] Genesereth, M.; Nilsson, N. *Logical Foundations of Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann Ed., 1988.

[Gonzalez2000] 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.

[Gonzalez2000a] Gonzalez, M. Representação Semântica de Sentenças em Linguagem Natural e sua aplicação na Recuperação de Informação. Trabalho Individual II, PPGCC, Faculdade de Informática, PUCRS, setembro 2000. 71 p.

[Gonzalez2001] Gonzalez, M. Thesauri. Trabalho Individual III, PPGCC, Faculdade de Informática, PUCRS, maio 2001. 98 p.

[Gonzalez2001a] Gonzalez, M.; Lima, Vera L. S. Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação. ENIA'2001, Fortaleza, Brasil, 2001. CD-ROM, ISBN 85-88442-04-3.

[Gonzalez2001b] Gonzalez, Marco; Lima, Vera L. S. de. Recuperação de Informação e Expansão Automática de Consulta com Thesaurus: uma avaliação. XXVII Conferencia Latinoamericana de Informatica (CLEI'2001). Ciudad de Mérida, Venezuela, 2001. CD-ROM, ISBN 980-110527-5.

[Gonzalez2001c] Gonzalez, Marco; Lima, Vera L. S. de. Thesaurus com Estruturação Semântica e Operações Gerativas. XXVII Conferencia Latinoamericana de Informatica (CLEI'2001). Ciudad de Mérida, Venezuela, 2001. CD-ROM, ISBN 980-110527-5.

[Gonzalez2001d] Gonzalez, Marco; Lima, Vera L. S. de. Semantic Thesaurus for Automatic Expanded Query in Information Retrieval. 8th Symposium on String Processing and Information Retrieval (SPIRE'01), Chile, 2001. IEEE Computer Society Publications, ISBN 0-7695-1192-9. Proceedings, p.68-75.

[Gonzalo1998] Gonzalo, J.; Verdejo, F.; Chugur, I.; Cigarran, J. Indexing with WordNet Synsets can Improve Text Retrieval. COLING/ACL'1998 – Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998. p.1-7.

[Guthrie1996] Guthrie, L.; Pustejovsky, J.; Wilks, Y.; Slator, B. M. The Role of Lexicons in Natural Language Processing. *Communications of the ACM*, V.39, N.1, janeiro 1996. p.63-72.

[Haller1986] Haller, Johann. Análise Lingüística e Indexação Automática de Textos. *VERITAS*, V.31, N.123, setembro 1986. p.393-405.

[Hearst1992] Hearst, M. A. Automatic Acquisition of Hyponyms from Large Text *Corpora*. Fourteenth International Conference on Computational Linguistics, 1992.

[Jacquemin1997] Jacquemin, Christian; Klavans, Judith L.; Tzoukermann, Evelyne. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology

and Syntax. 35th Annual Meeting of ACL – 8th Conf. of the European Chapter of the ACL, Madri, 1997. p.24-31.

[Jacquemin2000] Jacquemin, Christian. Traitement Automatique des Langues pour la Recherche d'Information. Paris: Hermès Science Publications, 2000. 591 p.

[Järvelin1996] Järvelin, K.; Kristensen, J.; Niemi, T.; Sormunen, E.; Keskustalo, H. A Deductive Data Model for Query Expansion. In: Frei, H.P. et al. (Editores.), Proceedings do 19th Annual International ACM SIGIR - Conference on Research and Development in Information Retrieval (ACM SIGIR'96), Zurique, 1996. p. 235-243.

[Jing1994] Jing, Y.; Croft, W.B. An Association Thesaurus for Information Retrieval. RIAO, 1994. p.146-160.

[Jurafsky2000] Jurafsky, D.; Martin, J. Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey, USA: Prentice-Hall, 2000. 934 p.

[Kahane2001] Kahane, Sylvain; Polguere, Alain. Formal Foundation of Lexical Functions. ACL'2001 – Workshop on Collocation, Toulouse, 2001.

[Kaji2000] Kaji, H.; Morimoto, Y.; Aizono, T.; Yamasaki, N. Corpus-dependent Association Thesauri for Information Retrieval. 18th International Conference of Computational Linguistics – Coling 2000, Nancy, 2000. p.1-7

[Kim2000] Kin, M.; Lu, F.; Raghavan, V. V. Automatic Construction of Rule-based Trees for Conceptual Retrieval. 7th International Symposium on String Processing and Information Retrieval (SPIRE), 2000. p.153-161.

[Kowalski1997] Kowalski, G. *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers, 1997. 282 p.

[Krenn1997] Krenn, Brigitte; Samuelsson, Christer. The linguist's guide to statistics. Material de apoio para o curso de Statistical Approaches in Computational Linguistics, da Universidade de Saarland. Versão de dezembro de 1997. 172 p.

[Krovetz1997] Krovetz, Robert. Homonymy and Polysemy in Information Retrieval. 35th Annual Meeting of the ACL and 8th Conf. of the European Chapter of the ACL. UNED, Madrid, 1997. p.7-12.

[Lewis1996] Lewis, David D.; Jones, Karen S. Natural Language Processing for Information Retrieval. *Communications of the ACM*, V.39, N. 1, 1996. p.92-101.

[Lobato1986] Lobato, L. M. P. *Sintaxe Gerativa do Português: da Teoria Padrão à Teoria da Regência e Ligação*. Rio de janeiro: Ed. Vigília, 1986. 558 p.

[Loukachevitch1999] Loukachevitch, N. V.; Salli, A. D.; Dobrov, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. NLDB'99 – 4ª Int. Conf. on Applications of Natural Language to Information Systems. OCG Schriftenreihe, Lecture Notes, V.129, 1999. p.203-208.

[Lyons1977] Lyons, J. *Semantics*. Cambridge: Cambridge University Press, 1977. V. 1 e 2.

[Mandala1999] Mandala, R.; Tokunaga, T.; Tanaka, H. Complementing WordNet with Roget's and *Corpus*-based Thesauri for Information Retrieval. EACL'99. Nineth Conf. of the European Chapter of the ACL. Noruega, 1999. p.94-101.

[Meadow2000] Meadow, C.T.; Boyce, B.R.; Kraft, D.H. *Text Information Retrieval Systems.* San Diego: Academic Press, 2000. 364 p.

[Mel'cuk1992] Mel'Cuk, Igor, Zholkovsky, Alexander. The Explanatory Combinatorial Dictionary. In: EVENS, Martha W. (Editora) *Relational Model of the Lexicon: Representing Knowledge in semantic networks.* New York: Cambridge University Press, 1992, p.41-74.

[Moens2000] Moens, Marie-Francine. Automatic Indexing and Abstracting of Document Texts. Boston: Kluwer Academic Publishers, 2000. 265 p.

[Moldovan2000] Moldovan, D.I.; Mihalcea, R. Using WordNet and Lexical Operators to Improve Internet Searches. IEEE Internet Computing, V.4, N.1, janeiro-fevereiro 2000. p.34-43.

[Morin1999] Morin, E. Jacquemin, C. Projecting *Corpus*-Based Semantic Links on a Thesaurus. 37th Annual Meeting of the Ass. for Comp. Linguistics, Maryland, 1999. p.20-26.

[Nunes1999] Nunes, M. das Graças Volpe; Silva, Bento C. D. da; Rino, Lúcia H. M.; Oliveira Jr., Osvaldo N. de; Martins, Ronaldo T; Montilha, Gisele. Introdução ao Processamento das Línguas Naturais. Notas Didáticas do ICMC (Instituto de Ciências Matemáticas e de Computação), São Carlos, 1999.

[Orengo2001] Orengo, V. M.; Huyck, C. A Stemming Algorithm for the Portuguese Language. Eighth Symposium on String Processing and Information Retrieval (SPIRE 2001), Chile, 2001. p.186-193.

[Pearce2001] Pearce, D. Synonymy in Collocation Extraction. NAACL 2001 Workshop – WordNet and Other Lexical Resources, Pittsburgh, USA, 2001. p.1-6.

[Pustejovsky1995] Pustejovsky, J. *The Generative Lexicon*. Cambridge: The MIT Press, 1995, 298 p.

[Pustejovsky1997] Pustejovsky, J.; Boguraev, B.; Verhagen, M.E.; Buitelaar, P.; Johnston, M. Semantic Indexing and Typed Hyperlinking .Working Notes of AAAI'97 – Spring Symposium on Natural Language Processing for the World Wide Web, 1997.

[Rijsbergen1979] van Rijsbergen, C.J. *Information Retrieval*. London: Bitterworths, 1979.

[Roget1958] Roget, Peter M.; Roget, John L.; Roget, Samuel R. *Thesaurus of English Words and Phrases*. London: Longmans, Green and Co., 1958.

[Russel1995] Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[Sacconi1999] Sacconi, Luiz Antonio. *Nossa Gramática: Teoria e Prática*. São Paulo: Atual Editora, 1999. 576 p.

[Saint-Dizier1999] Saint-Dizier, P. On the Polymorphic Behavior of Word-senses. In: *Lingüística Computacional: Investigação Fundamental e Aplicações.* Lisboa: Edições Colibri, 1999. p.29-56.

[Sanderson2000] Sanderson, M. Retrieving with Good Senses. *Information Retrieval*, V.2, N.1, 2000. p.49-69.

[Sanderson2000a] Sanderson, M.; Dawn, L. Building, Testing, and Applying Concept Hierarchies. In: Croft, W. B. (Editor) Advances in Information Retrieval. London: Kluwer Academic Publishers, 2000. p.235-266.

[SantaMaria1999] Santa Maria Fº., I.; Chishman, R. L. de O.; Lima, Vera L. S. de. Indexação Semântica de Documentos: uma aplicação da Teoria do Léxico Gerativo. IV Encontro para o Processamento da Língua Portuguesa Escrita e Falada (PROPOR'99), Évora, Portugal, 1999.

[Santos1996] Santos, Diana. Português Computacional. In: Duarte, I.; Leiria, I. (editores). *Actas do Congresso Internacional sobre o Português*. Lisboa: Edições Colibri, 1996. p.167-184.

[Santos2000] Santos, Diana. O Projecto Processamento Computacional do Português: Balanço e Perspectivas. V Encontro para o Processamento da Língua Portuguesa Escrita e Falada (PROPOR'2000). Atibaia, São Paulo, Brasil, 2000. p.105-113.

[Santos2001] Santos, Diana. Introdução ao Processamento de Linguagem Natural através das Aplicações. In: Ranchhod, E. (editora). *Tratamento das Línguas por Computador – Uma Introdução à Lingüística Computacional e suas Aplicações*. Lisboa: Caminho, 2001. p.229-259.

[Saracevic1975] Saracevic, T. Relevance: a review of and a framework for thinking on the notion in information science. Journal of the American Society for Information Science, N.26, 1975. p. 321-343. In: Sparck Jones, K; Willet, P. (editores). *Readings in Information Retrieval*. California: Morgan Kaufmann Publishers, Inc., 1997. p. 15-20.

[Scapini1995] Scapini, I.K.; Relações entre Itens Lexicais. In: Poersch, J. M.; Wertheimer, A.M.C.; Ouro, M.E.P.; Ludwig, E.M.; Scapini, I.K.; Becker, B.F. Fundamentos de um Dicionário Remissivo. 1º Encontro do CELSUL, Florianópolis, novembro 1995. Anais, V. 1, p.393-429.

[Schatz1997] Schatz, B. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, V.275, 1997. p.327-334.

[Shatkay2000] Shatkay, H.; Wilbur, J. Finding Themes in Medline Documents. Probabilistic Similarity Search, IEEE, Advances in Digital Libraries, 2000. p.1-10.

[Smeaton1997] Smeaton, A.F. Information Retrieval: Still Butting Heads with Natural Language Processing. In: Pazienza, M.T. (editor). *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Springer-Verlag Lecture Notes in Computer Science, N.1299, 1997. p.115-138.

[Soergel1997] Soergel, D. Multilingual Thesauri in Cross-language Text and Speech Retrieval. AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997. p.1-8.

[Sparck-Jones1986] Sparck-Jones, K. *Synonymy and Semantic Classification*. Edinburgh: Edinburgh University Press, 1986. 285 p.

[Sparck-Jones1997] Sparck-Jones, K.; Willet, P. (editores). *Readings in Information Retrieval*. California: Morgan Kaufmann Publishers, Inc., 1997.

[Sparck-Jones1999] Sparck-Jones, K. Information Retrieval and Artificial Intelligence. *Artificial Intelligence*, N. 114, 1999. p.257-281.

[Sparck-Jones1999a] Sparck-Jones, K. What is the Role of NLP in Text Retrieval? In: Strzalkowski, Tomek (editor). *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999. p.1-24

[Storb1998] Storb, B. H.; Wazlawick, R. S. Um Modelo de Recuperação de Documentos para a Língua Portuguesa utilizando Stemming Difuso. III Encontro para o Processamento da Língua Portuguesa Escrita e Falada (PROPOR'1998), Porto Alegre, Brasil, 1998. p79-87.

[Strzalkowski1999] Strzalkowski, Tomek (editor). *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999. 385 p.

[Strzalkowski1999a] Strzalkowski, Tomek; Perez-Carballo, José; Karlgren, Jussi; Hulth, Anette; Tapanainen, Pasi; Lahtinen, Timo. Natural Language Information Retrieval: TREC-8 Report. 8th Text REtrieval Conference. NIST Special Publication, 1999.

[Swanson1988] Swanson, D. R. Historical note: information retrieval and the future of an illusion. Journal of the American Society for Information Science, V.39, p.92-98. In: Sparck-Jones, K.; Willet, P. (editores). *Readings in Information Retrieval*. California: Morgan Kaufmann Publishers, Inc., 1997.

[Vieira2000] Vieira, Renata. Textual Co-reference annotation: a Study on Definite Descriptions. VIII Congresso da Sociedade Argentina de Lingüística, Mar del Plata, Argentina, 2000.

[Vieira2001] Vieira, Renata; Lima, Vera L. S. de. Lingüística Computacional: Princípios e Aplicações. JAIA, SBC, Fortaleza, Brasil, 2001.

[Voorhees1999] Voorhees, E. M. Natural Language Processing and Information Retrieval. In: Information Extraction: towards scalable, adaptable systems. Lecture notes in Artificial Intelligence, N.1714, 1999. p.32-48.

[Voorhees2000] Voorhees, E. H.; Harman, D. Overview of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication, 2000.

[Yule1998] Yule, George. *The Study of Language*. Cambridge: University Press, 1998. 294 p.

[Wertheimer1995] Wertheimer, A. M. C. O Dicionário Remissivo Comparado aos Outros Dicionários Existentes. In: Poersch, J. M.; Wertheimer, A. M. C.; Ouro, M. E. P.; Ludwig, E. M.; Scapini, I. K.; Becker, B. F. Fundamentos de um Dicionário Remissivo. 1º Encontro do CELSUL, Florianópolis, novembro 1995. Anais, V. 1, p.393-429.

[Wilks1996] Wilks, Y. A.; Slator, B. M.; Guthrie, L. M. *Electric Words: Dictionaries, Computers and Meanings*. Cambridge: The MIT Press, 1996. 289 p.