INDEXAÇÃO AUTOMÁTICA DE TEXTOS

JOHAN HALLER
Departamento de Letras e Lingüística
Universidade de Brasília
70910 Brasília, DF

Descreve um programa experimental de análise automática de textos desenvolvido na Universidade de Brasília. Os passos da análise lingüística seguidos no presente modelo são:
- segmentação (de um texto em frases e palavras);
- procura no dicionário e análise morfológica de palavras portuguesas;
- desembiguação de homografias sintáticas;
- construção de uma árvore de dependência.

Com estes algoritmos, várias contribuições podem ser produzidas para ajudar o indexador ou para integrar, junto com processos estatísticos adicionais, um sistema de indexação automática:
- descritores simples em forma-base;
- descritores compostos (grupos nominais);
- descritores com peso, baseado nas funções sintáticas dentro da frase e em traços semânticos, com listas e/ou tesauros.

O programa experimental está sendo testado com várias bases de dados

1. INTRODUÇÃO

Nas últimas duas décadas foram criados vários programas e sistemas para o armazenamento e a recuperação de informações, os quais estão sendo modificados e adaptados, como no caso das repartições públicas no Brasil (PRODASEN, EMBRAPA, etc.). Estes sistemas apresentam vários problemas:

- ainda resulta muito *lixo* na procura da informação, e a precisão é pequena;
- nem todas as formas da palavra pesquisada são encontradas: (verbos, sufixos, plurais irregulares, etc.); o *recall* é pequeno;
- homografias sintáticas e semânticas causam problemas adicionais (SEGUNDO – NON OU PRP?);

— não é possível trabalhar com grupos de descritores lingüisticamente coerentes.

O tratamento da maioria destes problemas pode ser melhorado com a aplicação de uma análise lingüística completa do texto, se ele existe num meio eletrônico (o que vai ser o caso num futuro muito próximo).

Aqui serão descritos primeiro brevemente os passos da análise lingüística necessária (Haller 1982), e depois os algoritmos desenvolvidos para obter uma lista de candidatos a descritores simples e compostos.

Um programa experimental está sendo desenvolvido no âmbito da pós-graduação na Universidade de Brasília, com a coloboração dos alunos dos Departamentos de Letras e Lingüística e de Biblioteconomia, sob a coordenação do autor, no computador Burroughs 6700 do CPD da UnB.

2. ANÁLISE LINGÜÍSTICA

2.1 Segmentação

A segmentação estabelece as unidades a serem tratadas seqüencialmente pelos algoritmos da análise lingüística propriamente dita. Em geral, a lingüística trata de **frases** como unidade de análise; só em casos especiais será necessário recorrer a informações da frase anterior, como, por exemplo, no caso dos pronomes, se estes (e as referências) são usados para fins de indexação estatística.

2.2. Dicionários

2.2.1 Dicionário de Freqüência

Nesse arquivo são armazenadas as palavras mais freqüentes de uma língua, que constituem mais ou menos 50% de um texto normal. Uma procura econômica é assim garantida. Essas palavras são, ao mesmo tempo, as mais importantes na língua para determinar a estrura e até a semântica (lógica) de uma frase, e por isso elas precisam de um máximo de informações morfológicas, sintáticas e semânticas. Em compensação, elas não têm **sentido próprio** e não podem entrar, por exemplo, como descritores para um sistema de informação.

2.2.2 Outros Dicionários

Um sistema de indexação não precisa absolutamente de mais dicionários do que eles são necessários, por exemplo, para um sistema de tradução. Porém, se existe um *thesaurus* preestabelecido para a indexação de textos de um certo assunto, podem ser acrescentadas informações morfológicas, sintáticas e semânticas a essas palavras, para facilitar o trabalho de análise lingüística. O que é preciso, de todo modo, são listas das formas irregulares de palavras e raízes, para o reconhecimento e a correspondente **normalização** da palavra, i.e, a produção da forma

básica ou lexema (infinitivo do verbo, forma singular do substantivo, forma singular masculina do adjetivo, etc.).

2.3 Análise Morfológica e Procura no Dicionário

Para se fazer uma análise morfológica precisar-se-á das possíveis terminações de verbos, adjetivos, substantivos e advérbios, e também de informações sobre a combinatória com as formas básicas (lexemas) dos dicionários. Podem acontecer homografias entre estas classes e também entre elas e o dicionário de freqüência.

EXEMPLOS: PARA — preposição, forma verbal
NADA — pronome, forma verbal
CASA — *substantitivo, forma verbal, etc.*

Por esta razão, a procura no dicionário tem que ser feita com todas a possibilidades que a terminação permite.

2.4 Análise Sintática

Embora seja possível fazer essa análise em um passo só, preferimos fazê-la em duas etapas: a desambiguação das homografias sintáticas e a construção da árvore estrutural. Os sistemas existentes, que preferem fazer desde já análises estruturais das várias combinações de homógrafos possíveis, são todos restritos com respeito ao vocabulário e às estruturas tratadas; no caso da entrada de um texto qualquer, os problemas do alto número de possibilidades a analisar e dos retornos dos caminhos errados não foram ainda resolvidos. Todos os sistemas operacionais no âmbito da indústria ou de órgãos de governos trabalham com uma análise sintática em dois passos separados.

2.4.1 Resolução das Homografias Sintáticas

Para decidir qual função sintática de um homógrafo está realizada numa determinada frase, utiliza-se um processo de aproximação partindo do contexto da palavra. Nesse processo, fazem-se várias passagens pela frase, começando com a primeira ou com a última palavra (dependendo da língua analisada). Resolvem-se primeiro os casos mais fáceis e seguros. Geralmente também é melhor não determinar em seguida a função atual da palavra, senão começar por excluir o que ela não pode ser no contexto atual.

Um exemplo seria o seguinte: se, dentro de uma frase sem nenhuma vírgula nem conjunção, se encontra já uma forma verbal que somente tem essa função possível, a possibilidade de ela ser verbo vai ser excluída com certeza nos outros homógrafos. *No fim de cada passagem é anotado se alguma coisa podia ser feito* e se ainda existem homografias. Se, depois de uma passagem, não foi resolvido mais nada, existindo ainda homografias, procede-se ao segundo grupo de regras, que se baseiam na probabilidade. Por exemplo, o verbo, em português, geralmente vem depois de um substantivo ou adjetivo (que fazem parte do grupo nominal sujeito) e pode ser decidido neste sentido.

2.4.2 Construção da Árvore Estrutural (de dependência)

Para estabelecer as dependências das palavras dentro da frase é preciso primeiro estabelecer as fronteiras das partes da frase (principal, subordinadas, grupos adverbiais, etc.). Isto é feito com a determinação de vírgulas, conjunções de subordinação e coordenação, etc. O próximo passo consiste em estabelecer os grupos nominais, verbais, preposicionais e adverbiais dentro dessas partes. Aqui se usa como ajuda a informação morfológica das palavras (gênero, número, tempo, modo, etc.). No último passo se determinam então as funções sintáticas desses grupos: sujeito, objeto (direto e indireto), complementos nominais, predicados (grupos verbais)e a dependência de um grupo adverbial (de um grupo nominal, do predicado ou da frase inteira); os membros de um grupo têm que ter um conjunto comum dessas informações morfológicas e sintáticas. O resultado dessa análise sintática é apresentado em forma de uma árvore estrutural.

3. APLICAÇÃO NA INDEXAÇÃO

3.1 Descritores Simples

Depois da resolução das homografias sintáticas é possível fornecer uma lista de substantivos e adjetivos (eventualmente também de verbos e advérbios) nas suas formas lexicalizadas, junto com a freqüência absoluta no texto tratado, no grupo de textos tratados e a freqüência relativa do texto dentro do grupo. Essa lista pode servir como proposta de descritores para o indexador humano, o que lhe facilita muito o trabalho, por exemplo, de construir um *thesaurus* numa área nova. Se já existe um *thesaurus* para a área correspondente, os elementos da lista podem ser comparados com os elementos do *thesaurus* e, assim, estabelecidos descritores para cada elemento. Com um processamento estatístico mais refinado pode-se também pensar na automatização completa da elaboração de um *thesaurus* e da indexação de um novo documento, onde só seria necessário um controle dos resultados pelo indexador humano.

3.2 Atribuição de Peso aos Descritores

3.2.1 Critérios Sintáticos

Como a análise lingüística não só fornece a forma-base, senão com o passo da análise sintática também a função da palavra dentro da frase, pode-se atribuir um peso lingüístico a cada substantito ou adjetivo, dependendo de, por exemplo, se ele funciona como sujeito, como objeto, como predicado, se ele se encontra dentro da frase principal, da subordinada, frases negativas, etc. Essas informações podem corrigir a mera estatística no processo da construção automática do *thesaurus* ou na indexação.

3.2.2. Critérios Semânticos

Além dos critérios sintáticos existem, obviamente, critérios semânticos, segundo os quais um indexador humano decide quais são as palavras-chave de um texto e quais não são. Por exemplo, há palavras que nunca vão ser descritores, como:

ÂMBITO
APOIO
ATO
FIM
FLEXIBILIDADE, etc.

De momento, estas palavras são coletadas e armazenadas numa lista; quando elas aparecem no texto, são eliminadas como candidatas a descritores. Uma análise sistemática dos traços semânticos precisa ser feita quando o volume dessa lista atinge um tamanho maior. Outras palavras são julgadas pelo indexador como impossíveis de serem candidatas a descritores simples, mas que indicam um descritor composto. Por exemplo:

AÇO
ACORDO
AMOSTRA
ESCALA, etc.

Valem as mesmas considerações feitas para o grupo anterior. Um terceiro grupo são palavras que seriam candidatas a descritores somente em textos muito especiais. Por exemplo:

CAUSA (Direito)
ÓRGÃO (Medicina)
PALAVRA (Lingüística)

Estas palavras devem entrar somente se a matéria tratada é conhecida de antemão; em compensação, elas podem até contribuir para fazer essa definição automaticamente, quando palavras de uma matéria são muito freqüentes dentro de um texto ou de um grupo de textos.

3.3 Descritores Complexos

A análise sintática (2.4.2.) é também capaz de gerar uma lista de descritores complexos, como, por exemplo, todos os seguintes grupos:

- substantivo-adjetivo
- substantitvo-negação-adjetivo
- substantivo-grupo preposicional, etc,

que geralmente são extraídos de um texto no processo da indexação. Geralmente será formado um descritor composto, se o substantivo principal (o head do grupo nominal) fizer parte da lista dos *indicadores de descritor complexo*, descrita em 3.2.2. Podem ser desenvolvidos mecanismos que analisam diversas formas lingüís-

ticas que geralmente têm o mesmo conteúdo, como por exemplo frase relativa — grupo com partícipio, etc.

3.4 Resultado Final

O resultado final dessas operações é agora uma lista (ou várias) de candidatos a descritores simples e compostos, cada um com um *peso*, que é a soma dos pesos em cada ocorrência dentro do texto ou do grupo de textos. Algumas experiências mostram que as palavras de maior peso correspondem, em geral, aquelas que o indexador humano extrairia do texto correspondente ou da pergunta do usuário. Torna-se agora necessária uma verificação com um volume maior de dados. Em todo caso, a lista de candidatos é uma ajuda preciosa para o indexador humano, especialmente quando se trata de uma área nova, na qual se pretende construir um vocabulário normativo ou um tesauro.

Abstract:

Automatic indexing of texts

Describes an experimental program of an automated analysis of texts developed in the University of Brasilia. The steps followed in the model are: segmentation of a text in phrases and words, words in a dictionary, and morphological analysis of Portufuese words.

REFERÊNCIAS

ANDREEWSKI. A./RUAS, V. Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informacão**, Vol. 12, Nº 1, 1983; 61-73.

BARANOW, U. G. Perspectivas na contribuição da lingüística e áreas afins à ciência da informação. In: **Ciência da informação**, vol. 12, Nº 1, 1983; 23-36.

BIDERMAN, M. T. Lingüística computacional, um desafio de trinta anos. In: **Dados e Idéias**, junho/julho 1977, pág. 29-39.

CINTRA, A. M. M. Elementos de lingüística para estudos de indexação. In: **Ciência da Informação**, vol. 12, Nº 1, 1983, 5-12.

DILLON, M./GRAY, A.: FASIT, A Fully Automatic Syntactically Based indexing System. In: Information Science, 34(2):099-108; 1983.

FISCHER, H. G. **CONDOR, an Integrated Data-Base Information and Retrieval System for Structured and Unstructured data**; Siemens Forsch. und Entw.-Berichte, Berlin/New York 1981.

HALLER, J. Processamento de textos em linguagem natural. In: **Anais do XV Congresso Nacional de Informática — SUCESU** Rio de Janeiro, out. 1982, p. 251-260.

HALLER, J. Análise automática de textOs em sistemas de informação. In: **Revista de Biblioteconomia de Brasília**, Vol. 11, Nº 1, Jan/Junho 1983, p. 105-114.

SALTON, G. **Automatic Information Organization and Retrieval**. New York, 1968.

WOODS, W.A.E.A. **The Lunar Sciences Natural Language Information System**, Final Report BBN 2378, Cambridge/Mass. 1972.

ZIMMERMANN, H. E. A. JUDO — Juristiche Dokumentanalyse Bericht 1977 — 1979, Universitaet Regensburg. Mai, 1980.