CDU: 025.4:681.3

ANÁLISE AUTOMÁTICA DE TEXTOS EM SISTEMAS DE INFORMAÇÃO

JOHANN HALLER
Departamento de Letras e lingüística, Universidade de Brasília, Brasília, D.F

Uma análise lingüística completa é um instrumento poderoso para melhorar a capacidade de um sistema automatizado de armazenamento e recuperação de informações. O presente artigo descreve um programa experimental que está sendo desenvolvido na Universidade de Brasília. Os passos da análise lingüística são, no presente modelo: segmentação (de um texto em frases e palavras); procura no dicionário e análise morfológica de palavras portuguesas; clareza de homografias sintáticas; construção de uma árvore de dependência; análise de pronomes. Com estes algoritmos, várias contribuições podem ser produzidas para ajudar o indexador humano 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; correção estatística com as referências dos pronomes. O programa experimental está sendo testado com várias bases de dados. O desenvolvimento futuro vai se dirigir para a construção de redes semânticas a partir de textos em linguagem natural.

0. INTRODUÇÃO

Na maioria dos sistemas de armazenamento e recuperação de informação existentes, são armazenados apenas os dados “estruturados” de um documento (AUTOR, AND, LUGAR, etc.), o título e eventualmente um pequeno resumo. Além disso, faz-se uma indexação manual ou semi-automática onde o computador às vezes ajuda propondo descritores por meio de sistemas KWIC etc.

Em alguns poucos sistemas existe a possibilidade de fazer uma busca também no texto mesmo (p. ex. STAIRS/ILM).

Mesmo quando se eliminam as palavras “de freqüência” (artigos, pronomes, etc.) e membros de uma lista de palavras “vazias” que variam segundo o assunto tratado, restam vários problemas sem solução.

- Ainda resulta muito “lixo” na procura da informação,
  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
- homografias sintáticas e semânticas causam problemas adicionais (SEGUNDO - NON OU PRP?)
- Não é possível trabalhar com grupos de descritores lingüísticamente coerentes.

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

Esta análise da língua é um processo dispendioso porque precisa de muitos dados (pensa-se num dicionário) e algoritmos muito complexos. A linguagem humana é algo complicado. Por isso, pode-se começar com a aplicação somente dos primeiros passos em seguida descritos aos textos destinados a indexação. A análise tem que ser considerada dentro de um contexto mais amplo da produção, armazenagem, busca e tradução de textos, onde ela abre um grande gama de possibilidades para os usuários de um documento.

- Correção de erros de datilografia
- Hifenação automática
- Pré-tradução automática de textos técnicos
- Análise da pergunta do usuário contribuindo para uma melhor comunicação do usuário com o sistema
- Etc.

A seguir, serão descritos os passos de uma análise lingüística automática e sua importância para a indexação de um texto ou de uma frase-pergunta.

Um programa experimental está sendo desenvolvido no âmbito da pós-graduação na Universidade de Brasília, com a colaboração de alunos dos departamentos de lingüística e de biblioteconomia, sob a coordenação do autor, no computador Burroughs 6700 do CPD da UnB.

1. Análise lingüística de um texto em linguagem natural

1.1 leitura de textos.

Ainda hoje em dia, um dos maiores problemas da chamada lingüística computacional que se ocupa da análise automática de textos e o fato que os textos a serem tratados não existem armazenados em entidades da periferia de um computador. Quando isso ocorre, significa que o texto tem que ser datilografado de novo ou então lido por uma máquina chamada “leitura ótica”. Como o segundo processo ainda não está desenvolvido em grande escala, nos dois casos sofrem muitos erros de datilografia ou de leitura, o que encarece bastante o processo da análise do texto, , além de anular a vantagem do tempo mas, como está sendo introduzido mais e mais a máquina de escrever eletrônica ou o mini-computador para a produção de textos e mesmo para a transmissão deles, este problema tende a desaparecer num futuro bastante próximo.
Resta neste complexo não propriamente linguístico apenas a tarefa de determinar qual : a informação estruturada de um documento

(tabelas, fórmulas, etc) e qual a informação não-estruturada (frases em linguagem natural) que vai ser submetida a analise lingüística.
Este algoritmo tem por exemplo que decidir quando um ponto realmente significa o fim de uma frase e quando ele é apenas um ponto decimal (“1.000”), um ponto de parágrafo (“1.) ou faz parte de uma reticência (“...”).
Esto é necessário para estabelecer as unidades a serem tratadas sequencialmente pelos algoritmos da análise lingüística propriamente dita. Só em poucos casos vai ser necessário recorrer as informações de outras frases, por exemplo na análise de pronomes.

1.2 Dicionários

1.2.1 Dicionário de “freqüência”.

Neste arquivo são armazenadas as palavras mais frequentes de uma língua que constituem mais ou menos 50% de um texto normal. Uma procura econômica e assim garantida.
Estas palavras são, ao mesmo tempo, as mais importantes da língua para determinar a estrutura 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 tem “sentido próprio” e não podem entrar por exemplo, como descritores pra um sistema de informação.
A seguir, uma amostra do dicionário de freqüência de portugues.

Tabela 1: Amostra do dicionário de freqüência

FWORD	FAC CNP TIG	CNP TIG	C
A	3095	F29S	F18
AQ	118S		
AQUELA	211S3	F12S3	F
AQUELAS	211P3	F12P3	F
AQUELE	211S3	M12S3	M
AQUELES	211P3	M12P3	M
AQUILO	211S3	N12S3	N
AS	118P		
ACASO	2015	M26	M
ADIANTE	126	L	
AGORA	126	T	
AI	126	L	
AINDA	126	T	
ALGO	212S3	N	
ALGUEIM	211S3	N12S3	N
ALGUM	211S3	M12S3	M
ALGUMA	211S3	F12S3	F
ALGUMAS	211P3	F12P3	F
ANTE	118		
ANTES	218	T26	1

LEGENDA:

FAC = FATOR DE FUNCOES (1 = SO UMA FUNCAO SINTATICA, 2 = 2 FUNCOES ETC.)
C = CATEGORIA SINTATICA (POR EX. 09 = DETERMINADOR, 26 = ADVERBIO ..)
N = NUMERO (SINGULAR, PLURAL)
P = PESSO (1,2,3)
TI = TEMPO (PRESENTE, PRETERITO, FUTURO)
G = GENERO NAS FORMAS NOMINAIS (MASC., FEM., NEUTRO)
= MODO NAS FORMAS VERBAIS (INDICATIVO, SUBJUNTIVO, CONDICIONAL)

Na versão atual do dicionário de freqüência do portugues, ainda tem poucas informações

semânticas, só para os pronomes, clantures [1](positivo, negativo, 1000, poucos, nada etc.) e algumas preposições. Este dicionário é geralmente uma parte das palavras "vazias" dentro de um sistema de indexação.

1.2.2. Outros dicionários.

Um sistema de indexação não precisa absolutamente de mais dicionários como eles são necessários, por exemplo, para um sistema de tradução, porém, se existe um tesaurus pré-estabelecido para a indexação de textos de um certo assunto, podem ser acrescentadas informações morfológicas, sintáticas e semânticas a estas palavras para facilitar o trabalho da análise lingüística.
O que é preciso de todos modos, 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 masculino do adjetivo etc.).

1.2.3. Dicionário de expressões fixas

Muitas vezes acontece que dentro de uma língua, um grupo de palavras tome um novo sentido e até uma função sintática diferente quando ocorre junto numa frase.
Estes casos tem que ser armazenados também num arquivo, classificados de preferência pela palavra chave da expressão. Um exemplo em português pode ser "bater um papo" o qual tem que ser reconhecido como grupo especial - com o verbo "bater" podendo aparecer em todos os tempos e todos os modos possíveis - para ser traduzido por exemplo corretamente para outra língua.
Em inglês, "in the process of" toma como grupo a nova função de preposição que introduz um grupo preposicional.
Está claro que estas expressões muito poucas vezes vão entrar na indexação de um texto.

1.2.4. Equivalências em línguas estrangeiras

Este arquivo contem as formas bases de palavras da língua estrangeira com informações sobre declinação e flexão das palavras para servir de referência num sistema de informação multilingual ou num sistema de tradução.
Nos sistemas de tradução mais antigos, as partes 1.2.2 e 1.2.4 estão armazenados num só arquivo isto é uma desvantagem quando se quer trabalhar com várias línguas.

1.3. Análise morfológica e procura no dicionário

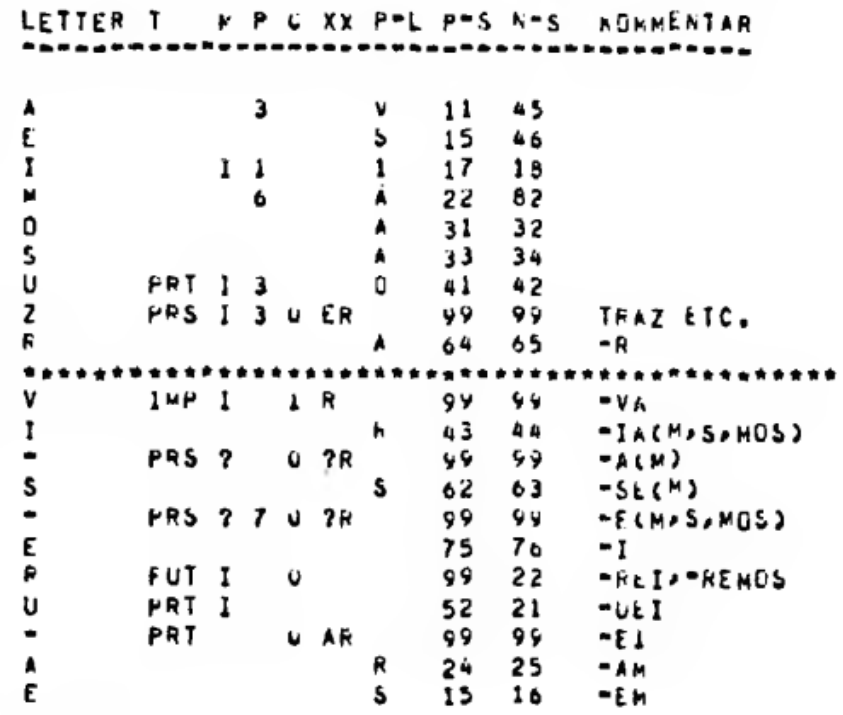
Para se fazer uma análise morfológica, precisará das possíveis terminações de verbos, adjetivos, substantivos e advérbios, e também informações sobre a combinatória com as formas básicas ("lexemas") dos dicionários.
Podem acontecer homografias entre estas classes e também entre elas e o dicionário de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

Exemplos: "Para" preposição, forma verbal
"Nada" pronome, forma verbal
"Casa" substantivo, forma verbal, etc.

Por esta razão, a procura no dicionário tem que ser feita com todas as possibilidades que a terminação permite. A seguir, como exemplo, a primeira parte da tabela para a análise dos verbos regulares da língua portugesa.

[1] não foi possível identificar esta palavra no artigo original (nota da Biblioteca Central da UnB em 10 de janeiro de 2023)

Tabela 2: Primeira parte da tabela de morfologia verbal



LETTER	T	M	P	C	XX	P-L	P-S	N-S	KOMMENTAR
A			3			V	11	45	
E						S	15	46	
I			1	1		1	17	18	
M			6			A	22	82	
O						A	31	32	
S						A	33	34	
U	PRT	I	3			O	41	42	
Z	PRS	I	3	U	ER		99	99	TRAZ ETC.
R						A	64	65	-R
V	IMP	I		1	R		99	99	-VA
I						H	43	44	-IA(M,S,MOS)
-	PRS	?		0	?R		99	99	-A(M)
S						S	62	63	-SE(M)
-	PRS	?	7	U	?R		99	99	-E(M,S,MOS)
E							75	76	-I
R	FUT	I		0			99	22	-REI,-REMOS
U	PRT	I					52	21	-UEI
-	PRT			U	AR		99	99	-EI
A						R	24	25	-AM
E						S	15	16	-EM

Legenda:
Letter = ultima letra da palavra (ate********)
Letra precedente (resto)
T = Tempo
M = mudo
P = Pessoa
C = Cul (Quantas letras curtar)
XX = terminação do infinitivo ("?R" nos casos de indecisão : -o, -a, etc., que são resolvidos na procura no dicionário)
P-L = Letra precedente
P-S = Próxima linha para continuar
Se letra precedente acontece na palavra
N-S = Linha para continuar no caso contrário
Comment = Comentário

Enquanto os dicionários ainda são pequenos - e mesmo depois, porque nunca se pode contar com todas as palavras de uma língua dentro do dicionário (neologismos etc.) - a análise morfológica tem que ser aplicada a todas as palavras técnicas para dar as classes sintáticas possíveis e o máximo de informações que se podem derivar da terminação. Só assim é possível também no caso de várias palavras desconhecidas pelo dicionário estabelecer a estrutura correta da frase.
Se por exemplo, a palavra "mundo" ainda não existe no dicionário, ela seria considerada como um possível substantivo (o que ela é de verdade), um possível adjetivo ("mundo, -a") e um possível verbo ("mundar" ou "munder" etc.).
A maioria destes casos duvidosos vão ser resolvidos na seguinte etapa da análise da homografia e da estrutura sintática

1.4. Análise sintática

Embora seja possível fazer esta análise em um passo só e de preferir
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 a respeito do vocabulário e das estruturas tratadas no caso da entrada de um texto qualquer, os problemas do alto número de possibilidades a analizar 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.

1.4.1. Resolução das homografias sintáticas

Para decidir qual função sintática de um homógrafo está realizada numa determinada frase, se utiliza um processo de aproximação partindo do contexto da palavra.
Neste processo, se fazem várias passagens pela frase, começando com a primeira ou a última palavra (dependendo da língua analizada). Se resolvem 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 esta função possível, esta possibilidade de ser verbo vai ser excluída com certeza nos outros homógrafos.
No fim de cada passagem é anotado se alguma coisa podia ser feita e se ainda existem homografias.
Se, depois de uma passagem, não foi resolvido mais nada, existindo ainda homografias, se procede ao segundo grupo de regras que se baseiam na probabilidade. Por exemplo, o verbo português geralmente vem depois de um substantivo ou adjetivo (que fazem parte do grupo nominal sujeito) e pode ser decidido neste sentido.
Uma outra possibilidade é portanto deixar as poucas estruturas ainda possíveis a analise estrutural para definir qual é a estrutura corretas esta possibilidade da conta também de frases com duas ou várias estruturas possíveis como elas aparecem frequentemente em trabalhos da lingüística técnica.

1.4.2. 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 destas 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 destes grupos: sujeito, objetos (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 tem que ter um conjunto comum destas informações morfológicas e sintáticas.
Usa-se o modelo da gramática de dependência que toma como núcleo da frase o complexo verbal.
O resultado desta análise sintática é apresentado em forma de uma árvore estrutural.

—---------- Desenho árvore —----------

1.5. Análise no texto

Em alguns casos, será necessário ter acesso a resultados de outras frases: por exemplo, quando, na tradução, o pronome inglês “They” se pode referir a um substantivo na frase anterior, para dizer se, em português, e “eles” ou “elas”, temos que saber a equivalência portuguesa do substantivo referido na frase anterior.

2. Aplicação na indexação

2.1. Descritores simples

Depois da resolução das homografias sintáticas (passo 1.4.1 da análise lingüística) é possível fornecer uma lista de substantivos e adjetivos (eventualmente também verbos e advérbios) nas suas formas lexicalizadas, junto com a frequencia absoluta no texto tratado, no grupo de textos tratados e a frequência relativa do texto dentro do grupo.
Esta lista pode servir como proposta de descritores para o indexador humano o que lhe facilita muito o trabalho, por exemplo, de construir um tesaurus numa área nova.
Se já existe um tesaurus para a área correspondente, os elementos da lista podem ser comparados com os elementos do tesauros e assim estabelecidos descritores para cada documento.
Com um processamento estatístico mais refinado pode-se também pensar na automatização completa da elaboração de um tesaurus e da indexação de um novo documento onde só seria necessário um controle dos resultados pelo indexador humano.

2.2. Atribuição de “peso” aos descritores

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 “linguístico” a cada substantivo do adjetivo, dependendo por exemplo se ele funciona como subjeito, como objeto, como predicado, se ele se encaixa dentro da frase principal, da subordinada, frases negativas etc.
Estas informações podem corrigir a mera estatística no processo da construção automática do tesaurus ou na indexação.

2.3. Descritores complexos

A análise sinática (1.4.2.) é também capaz de gerar uma lista de descritores complexos, por exemplo, todos os seguintes grupos:

- Substantivo-adjetivo
- Substantivo-negação-adjetivo
- Substantivo-verbo-preposicional

Etc., que geralmente são extraídos de um texto no processo da indexação.

Podem ser desenvolvidos mecanismos que analisam diversas formas lingüísticas que geralmente tem o mesmo conteúdo, por exemplo Frase Relativa - Grupo com particípio etc.

2.4 Correção mediante pronomes

Geralmente, os pronomes não são tomados em conta na estatística de palavras. Como a análise lingüística identifica o substantivo correspondente, pode ser acrescentado um ponto (ou meio ponto 9) na freqüência deste substantivo o que dá mais relevância as freqüências absolutas e relativas.

3. Possibilidades futuras: rede semântica

A análise lingüística descrita está sendo aplicada a várias bases de dados. Depois de atingir um certo grau de estabilidade, será desenvolvido o próximo passo: a construção de uma rede semântica de um texto.
Nesta rede, os sustantivos (ou um subgrupo deles, p. ex. os elementos de um tesauros) foram os nós e os verbos serão os arcos.
Alguns verbos podem funcionar dentro de uma só relação, simplificando assim consideravelmente a rede.
Uma procura de informação poderá tomar uma palavra como ponto de partida e tentar achar na rede um “caminho” semelhante a pequena rede formada pela análise lingüística aplicada a frase-pergunta.
Será atribuído o peso ao documento, a base das semelhanças encontradas entre a pergunta e a rede do texto, podendo-se apresentar estas frases do texto ao usuário quando este quer verificar se o documento corresponde a necessidade dele.
Com estes procedimentos, será dado um grande passo para uma melhor comunicação entre o homem e o computador e para o melhor uso de sistemas de informação.

Abstract
Automatic analysis of texts in information systems

One of the most powerful instruments to improve the capacity of automatic information storage and retrieval systems would be a complete linguistic analysis of the full text of a document. The article describes an experimental approach being developed at the University of Brasília, Brazil.
The passes of the linguistic analysis in the present model are:

- Segmentation (phrases, words)
- Dictionary search and morphological analysis (Portuguese)
- Disambiguation of syntactic homographies
- Construction of a dependency tree
- Pronoun Analysis

With these features, various itens can be provided to help the human indexer or to integrate, in conjunction with additional statistic procedures, an automatic indexing process:

- Single descriptors in lexicalized form
- Compound descriptors (noun groups etc)
- Weighted descriptors (weight based on the syntactic functions)
- Statistic correction with the references of pronouns.

The experimental program will undergo now a serie of testes with various data bases further development is directed to the construction of a semantic network.

REFERÊNCIAS

BÉLY, N./ BORILLO, A/ VIRBEL, J./ SIOT-DECAUVILLE, N. Procédures d'analyse sémantique appliquées à la documentation scientifique. CNRS - Paris 1970.

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

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.

HUTCHINS, W. J. Progress in documentation - machine translation and machine-aided translation. EM Journal of documentation, vol. 34, no. 2 June 1978

SALTON, G. I. Automatic information organisation and retrieval. New York 1968.

SCHANK, R.C. / COLBY, K. M. Computer models of thought and language. San Francisco 1973.

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