

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, ANO, 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 PRELIMINAR DESCRITORES POR MEIO DE SISTEMAS "KWIC" ETC. EM ALGUNS PEQUENOS SISTEMAS EXISTE A POSSIBILIDADE DE FAZER UMA BUSCA TAMBÉM NO TEXTO MESMO (P.EX. STAIRS/ILM). MESMO QUANDO SE ELIMINAM AS PALAVRAS "DE FREQUÊNCIA" (ARTIGOS, PRONOMES ETC.) E MEMBRAS DE UMA LISTA DE PALAVRAS "VAZIAS" QUE VARIAM SEGUNDO O ASSUNTO TRATADO, HESTAM VÁRIOS PROBLEMAS SEM SOLUÇÃO.

- AINDA RESULTA MUITO "LIXO" NA PROCURA DA INFORMAÇÃO, A PRECISÃO É BAIXA
- NEM TODAS AS FORMAS DA PALAVRA PESQUISADA SÃO ENCONTRADAS (VERBOS, SUFFIXOS, PLURAIS IRREGULARES ETC.)
- HOMOGRAFIAS SINTÁTICAS E SEMÂNTICAS CAUSAM PROBLEMAS ADICIONAIS ("SEGUNDO" = NOME 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 NUM MEIO ELETRÔNICO (O QUE VAI SER O CASO NUM FUTURO MUITO PRÓXIMO).

ESTA ANÁLISE DA LÍNGUA É UM PROCESSO DISPENDIOSO PORQUE PRECISA DE MUITOS DADOS (PENSA-SE NUM DICCIONÁRIO) E ALGORITMOS MUITO COMPLEXOS. A LÍNGUA HUMANA É ALGO COMPLICADO. POR ISSO, PODE-SE COMEÇAR COM A APLICAÇÃO SOMENTE DOS PRIMEIROS PASSOS EM SEGUNDA DESCRIÇÃO DOS 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:

- CRIAÇÃO DE FICHAS DE DACTILOGRAFIA
- INDEXAÇÃO AUTOMÁTICA
- PRE-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 ALGUNS DOS DEPARTAMENTOS DE LINGÜÍSTICA E DE BIBLIOTECOMIA, SOB A COORDENAÇÃO DO AUTORA, NO COMPUTADOR BURROUGHS 6700 DO CPD DA UNB.

1. ANÁLISE LINGÜÍSTICA DE UM TEXTO EM LINGÜAGEM NATURAL

1.1. LEITURA DE TEXTOS

AINDA HOJE EM DIA, UM DOS MAIORES PROBLEMAS DA CHAMADA "LINGÜÍSTICA COMPUTACIONAL" QUE SE OCORRE NA ANÁLISE AUTOMÁTICA DE TEXTOS É O FATO QUE OS TEXTOS A SEREM TRATADOS NÃO EXISTEM ARMAZENADOS EM UNIDADES DA PERIFERIA DE UM COMPUTADOR. QUANDO ISSO OCORRE, SIGNIFICA QUE O TEXTO TEM QUE SER DACTILOGRAFADO DE NOVO OU ENTÃO LIDO POR UMA MÁQUINA CHAMADA "LEITORA ÓPTICA". COMO O SEGUNDO PROCESSO AINDA NÃO ESTÁ DESENVOLVIDO EM GRANDE ESCALA, NOS DOIS CASOS SURTEM MUITOS ERROS DE DACTILOGRAFIA OU DE LEITURA O QUE ENCAPECE RASANTE O PROCESSO DA ANÁLISE DO TEXTO, ALÉM DE ANULAR A VANTAGEM DO TEMPO. MAS, COMO ESTÁ SENDO INTRODUZIDA MAIS E MAIS A MÁQUINA DE ESCRITA ELETRÔNICA OU O MINI-COMPUTADOR PARA A PRODUÇÃO DE TEXTOS E MESMO PARA A TRANSCRIÇÃO DELES, ESTE PROBLEMA TENDÊ A DESAPARECER NUM FUTURO BASTANTE PRÓXIMO. RESTA NESTE CONTEXTO NÃO PROPRIAMENTE LINGÜÍSTICO APENAS A TAREFA DE DETERMINAR QUAL É A INFORMAÇÃO ESTRUTURADA DE UM DOCUMENTO

(TABELAS, FÓRMULAS ETC.) E QUAL A INFORMAÇÃO NÃO-ESTRUTURADA (FRASES EM LINGÜAGEM NATURAL) QUE VAI SER SUBMETIDA À ANÁLISE 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") OU UM PONTO DE PARÁGRAFO ("1.") OU FAZ PARTE DE UMA ELÍPSE ("..."). ESTO É NECESSÁRIO PARA ESTABELEÇER AS UNIDADES A SEREM TRATADAS SEQUENCIALMENTE PELOS ALGORITMOS DA ANÁLISE LINGÜÍSTICA PROPRIAMENTE DITA. SO EM POUCOS CASOS VAI SER NECESSÁRIO RECORRER ÀS INFORMAÇÕES DE OUTRAS FRASES, POR EXEMPLO NA ANÁLISE DE PRONOMES

1.2. DICCIONÁRIOS

1.2.1. DICCIONÁRIO DE "FREQUÊNCIA"

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 PARA UM SISTEMA DE INFORMAÇÃO. A SEGUIR, UMA AMOSTRA DO DICCIONÁRIO DE FREQUÊNCIA DE PORTUGUÊS.

TABELA 11: AMOSTRA DO DICCIONÁRIO DE FREQUÊNCIA

FORMA	FAC	CNP	TIG	CNP	TIG	C
A	3095	F295	F18			
A4	1185					
AQUELA	21153	F1253	F			
AQUELAS	211P3	F12P3	F			
AQUELE	21153	M1253	M			
AQUELES	211P3	M12P3	M			
AQUILO	21153	N1253	N			
A45	118P					
ACASO	2015	M26	M			
ADIANTE	126	L				
AGORA	126	T				
AÍ	126	L				
AÍNDÁ	126	T				
ALGO	21253	N				
ALGUEM	21153	N1253	N			
ALGUM	21153	M1253	M			
ALGUMA	21153	F1253	F			
ALGUMAS	211P3	F12P3	F			
ANTE	118					
ANTES	218	T26	T			

LEGENDA:

- FAC = FATOR DE FUNÇÕES (1 = SO UMA FUNÇÃO SINTÁTICA, 2 = 2 FUNÇÕES ETC.)
- C = CATEGORIA SINTÁTICA (POR EX. 09 = DETERMINADOR, 26 = ADVERBIO ...)
- N = NÚMERO (SINGULAR, PLURAL)
- P = GÊNERO (1, 2, 3)
- T1 = TEMPO (PRESENTE, PRETERITO, FUTURO)
- G = GÊNERO NAS FORMAS NOMINAIS (MASC., FEM., NEUTRO)
- M = MODO NAS FORMAS VERBAIS (INDICATIVO, SUBJUNTIVO, CONDICIONAL)

NA VERSÃO ATUAL DO DICIONÁRIO DE FREQUÊNCIA DO PORTUGUÊS, AINDA TEM POUCAS INFORMAÇÕES SEMÂNTICAS, SO PARA OS PRONOMES, QUANTORES (POSITIVO, NEGATIVO, TODO, 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 ELAS SÃO NECESSÁRIOS, POR EXEMPLO, PARA UM SISTEMA DE TRADUÇÃO. PORÉM, SE EXISTE UM TESAUROUS PRÉ-ESTABELECIDO PARA A INDEXAÇÃO DE TEXTOS DE UM CERTO ASSUNTO, PODEREM SER AGREGADAS INFORMAÇÕES MORFOLÓGICAS, SINTÁTICAS E SEMÂNTICAS A ESTAS PALAVRAS PARA FACILITAR O TRABALHO DA ANÁLISE LINGÜÍSTICA. O QUE É PRECISO DE TODOS MODOOS, SÃO LISTAS DAS FORMAS IRREGULARES DE PALAVRAS E RAÍZES PARA O RECONHECIMENTO E A CORRESPONDÊNCIA "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 TOMA 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 MODOOS 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. ESTA CLARO QUE ESTAS EXPRESSÕES MUITO POUCAS VEZES VÃO ENTRAR NA INDEXAÇÃO DE UM TEXTO.

1.2.4. EQUIVALENCIAS 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 ARMAZENADAS NUM SO 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, PRECISARA DAS POSSÍVEIS TERMINAÇÕES DE VERBOS, ADJETIVOS, SUBSTANTIVOS E ADVERBOS, E TAMBÉM INFORMAÇÕES SOBRE A COMBINAÇÃO COM AS FORMAS BÁSICAS ("LEXEMAS") DOS DICIONÁRIOS. PODEREM 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 FEITO 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 PORTUGUESA.

TABELA 21 PRIMEIRA PARTE DA TABELA DE MORFOLOGIA VERBAL

LETTER T M P U XX P-L P-S N-S COMMENT

A		3		V	11	45	
E				S	15	46	
I		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				K	43	44	-IA(M,S,MOS)
-		PRS	? 0 ?R		99	99	-A(M)
S				S	62	63	-SE(M)
-		PRS	? 7 0 ?R		99	99	-F(M,S,MOS)
E					75	76	-I
P		FUT	I 0		99	22	-REI, REMOS
U		PRT	I		52	21	-UEI
-		PRT	0 AR		99	99	-EI
A				R	24	25	-AM
E				S	15	16	-EM

LEGENDA:

LETTER	=	ÚLTIMA LETRA DA PALAVRA (ATE *****)
	=	LETRA PRECEDENTE (RESTO)
T	=	TEMPO
M	=	MODO
P	=	PESSOA
C	=	CUT (QUANTAS LETRAS CORTAR)
XX	=	TERMINAÇÃO DO INFINITIVO ("?R" NOS CASOS DE INDECISÃO: -U, -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. SO ASSIM É POSSÍVEL TAMBÉM NO CASO DE VÁRIAS PALAVRAS DESCONHECIDAS PELO DICIONÁRIO ESTABELECEM 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

ENFURA SEJA POSSÍVEL FAZER ESTA ANÁLISE EM UM PASSO SO, E SE PREFERIR

FAZ-LA EM DUAS ETAPAS: A DESAMBIGUACAO DAS HOMOGRAFIAS SINTATICAS E A CONSTRUCAO DA ARVORE ESTRUTURAL. OS SISTEMAS EXISTENTES QUE PREFEREM FAZER DESDE JA ANALISES ESTRUTURAIS DAS VARIAS COMBINACOES DE HOMOGRAFOS POSSIVEIS SAO TODOS RESTRITOS A RESPEITO DO VUCABULARIO E DAS ESTRUTURAS TRATADAS; NO CASO DA ENTRADA DE UM TEXTO QUALQUER, OS PROBLEMAS DO ALTO NUMERO DE POSSIBILIDADES A ANALIZAR E DOS RETORNOS DOS CAMINHOS ERRAOS NAU FORAM AINDA RESOLVIDOS. TODOS OS SISTEMAS OPERACIONAIS NO AMBITO DA INDUSTRIA OU DE ORGAOS DE GOVERNO TRABALHAM COM UMA ANALISE SINTATICA EM DUAS PASSOS SEPARADOS.

1.4.1. RESOLUCAO DAS HOMOGRAFIAS SINTATICAS

PARA DECIDIR QUAL FUNCAO SINTATICA DE UM HOMOGRAFO ESTA REALIZADA NUMA DETERMINADA FRASE, SE UTILIZA UM PROCESSO DE APROXIMACAO PARTINDO DO CONTEXTO DA PALAVRA. NESTE PROCESSO, SE FAZEM VARIAS PASSAGENS PELA FRASE, COMECANDO COM A PRIMEIRA OU A ULTIMA PALAVRA (DEPENDENDO DA LINGUA ANALIZADA). SE RESOLVEM PRIMEIRO OS CASOS MAIS FACIS E SEGUROS. GERALMENTE, TAMBEM E MELHOR NAU DETERMINAR EM SEGUICA A FUNCAO ATUAL DA PALAVRA SENAO COMECAR POR EXCLUIR O QUE ELA NAU PODE SER NO CONTEXTO ATUAL. UM EXEMPLO SERIA O SEGUINTE: SE DENTRO DE UMA FRASE SEM NENHUMA VIRGULA NEM CONJUNCAO, SE ENCONTRA JA UMA FORMA VERBAL QUE SOIENTE TEM ESTA FUNCAO POSSIVEL, ESTA POSSIBILIDADE DE SER VERBO VAI SER EXCLUIDA COM CERTEZA NOS OUTROS HOMOGRAFOS. NO FIM DE CADA PASSAGEM E ANOTADO SE ALGUMA COISA PODIA SER FEITO E SE AINDA EXISTEM HOMOGRAFIAS. SE, DEPUIS DE UMA PASSAGEM NAU FOI RESOLVIDO MAIS NADA, EXISTINDO AINDA HOMOGRAFIAS, SE PROCEDE AO SEGUNDO GRUPO DE REGRAS QUE SE BASEIAM NA PROBABILIDADE. POR EXEMPLO, O VERBO PORTUGUES GERALMENTE VEM DEPUIS DE UM SUBSTANTIVO OU ADJETIVO (QUE FAZEM PARTE DO GRUPO NOMINAL SUJEITO) E PODE SER DECIDIDO NESTE SENTIDO. UMA OUTRA POSSIBILIDADE E PORTANTO DEIXAR AS POUCAS ESTRUTURAS AINDA POSSIVEIS A ANALISE ESTRUTURAL PARA DEFINIR QUAL E A ESTRUTURA CORRETA; ESTA POSSIBILIDADE DA CONTA TAMBEM DE FRASES COM DUAS OU VARIAS ESTRUTURAS POSSIVEIS COMO ELAS APARECEM FREQUENTEMENTE EM TRABALHOS DA LINGUISTICA TECNICA.

1.4.2. CONSTRUCAO DA ARVORE ESTRUTURAL (DE DEPENDENCIA)

PARA ESTABELECEER AS DEPENDENCIAS DAS PALAVRAS DENTRO DA FRASE, E PRECISO PRIMEIRO ESTABELECEER AS FRONTEIRAS DAS PARTES DA FRASE (PRINCIPAL, SUBORDENADAS, GRUPOS ADVERBIAIS ETC.). ISTO E FEITO COM A DETERMINACAO DE VIRGULAS, CONJUNCOES DE SUBORDINACAO E COORDENACAO ETC. O PROXIMO PASSO CONSISTE EM ESTABELECEER OS GRUPOS NOMINAIS, VERBAIS, PREPOSICIONAIS E ADVERBIAIS DENTRO DESTAS PARTES. AQUI SE USA COMO AJUDA A INFORMACAO MORFOLOGICA DAS PALAVRAS (GNERO, NUMERO, TEMPO, MODO ETC.). NO ULTIMO PASSO SE DETERMINAM ENTAO AS FUNCOES SINTATICAS DESTES GRUPOS: SUJEITO, OBJETO (DIRETO E INDIRETO), COMPLEMENTOS NOMINAIS, PREDICADOS (GRUPOS VERBAIS) E A DEPENDENCIA 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 INFORMACOES MORFOLOGICAS E SINTATICAS. USA-SE O MODELO DA GRAMATICA DE DEPENDENCIA QUE TOMA COMO NUCLEO DA FRASE O COMPLEXO VERBAL. O RESULTADO DESTA ANALISE SINTATICA E APRESENTADO EM FORMA DE UMA ARVORE ESTRUTURAL.

----- DESENHO ARVORE -----

1.5. ANALISE DO TEXTO

EM ALGUNS CASOS, SERA NECESSARIO TER ACESSO A RESULTADOS DE OUTRAS FRASES; POR EXEMPLO, QUANDO, NA TRADUCAO, O PRONOME INGLES "THEY" SE PODE REFERIR A UM SUBSTANTIVO NA FRASE ANTERIOR. PARA DIZER SE, EM PORTUGUES, E "ELES" OU "ELAS", TEMOS QUE SABER A EQUIVALENCIA PORTUGUESA DO SUBSTANTIVO REFERIDO NA FRASE ANTERIOR.

2. APLICACAO NA INDEXACAO

2.1. DESCRITORES SIMPLES

DEPUIS DA RESOLUCAO DAS HOMOGRAFIAS SINTATICAS (PASSO 1.4.1. DA ANALISE LINGUISTICA) E POSSIVEL FORNECER UMA LISTA DE SUBSTANTIVOS E ADJETIVOS (EVENTUALMENTE TAMBEM VERBOS E ADVERBIOS) NAS SUAS FORMAS LEXICALIZADAS, JUNTO COM A FREQUENCIA ABSOLUTA NO TEXTO TRATADO, NO GRUPO DE TEXTOS TRATADOS E A FREQUENCIA 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 TESAURO NUMA AREA NOVA. SE JA EXISTE UM TESAURO PARA A AREA CORRESPONDENTE, OS ELEMENTOS DA LISTA PODEM SER COMPARADOS COM OS ELEMENTOS DO TESAURO E ASSIM ESTABELECIDOS DESCRITORES PARA CADA DOCUMENTO. COM UM PROCESSAMENTO ESTATISTICO MAIS REFINADO PODE-SE TAMBEM PENSAR NA AUTOMATIZACAO COMPLETA DA ELABORACAO DE UM TESAURO E DA INDEXACAO DE UM NOVO DOCUMENTO ONDE SO SERIA NECESSARIO UM CONTROLE DOS RESULTADOS PELO INDEXADOR HUMANO.

2.2. ATRIBUICAO DE "PESO" AOS DESCRITORES

COMO A ANALISE LINGUISTICA NAU SO FORNECE A FORMA BASE, SENAO COM O PASSO DA ANALISE SINTATICA TAMBEM A FUNCAO DA PALAVRA DENTRO DA FRASE, PODE-SE ATRIBUIR UM PESO "LINGUISTICO" A CADA SUBSTANTIVO OU ADJETIVO, DEPENDENDO POR EXEMPLO SE ELE FUNCIONA COMO SUJEITO, COMO OBJETO, COMO PREDICADO, SE ELE SE ENCONTRA DENTRO DA FRASE PRINCIPAL, DA SUBORDENADA, FRASES NEGATIVAS ETC. ESTAS INFORMACOES PODEM CORRIGIR A MERA ESTATISTICA NO PROCESSO DA CONSTRUCAO AUTOMATICA DO TESAURO OU NA INDEXACAO.

2.3. DESCRITORES COMPLEXOS

A ANALISE SINTATICA (1.4.2.) E TAMBEM CAPAZ DE GERAR UMA LISTA DE DESCRITORES COMPLEXOS, POR EXEMPLO TODOS OS SEGUINTE GRUPOS:

- SUBSTANTIVO-ADJETIVO
- SUBSTANTIVO-NEGACAO-ADJETIVO
- SUBSTANTIVO-GRUPO PREPOSICIONAL

ETC., QUE GERALMENTE SAO EXTRAIDOS DE UM TEXTO NO PROCESSO DA INDEXACAO.

PODEM SER DESENVOLVIDOS MECANISMOS QUE ANALISAM DIVERSAS FORMAS LINGÜÍSTICAS QUE GERALMENTE TÊM 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 ?) NA FREQUÊNCIA DESTES SUBSTANTIVO O QUE DÁ MAIS RELEVÂNCIA ÀS FREQUÊNCIAS ABSOLUTAS E RELATIVAS.

3. POSSIBILIDADES FUTURAS: REDE SEMÂNTICA

A ANÁLISE LINGÜÍSTICA DESCRITA ESTÁ SENDO APLICADA À VÁRIAS BASES DE DADOS. DEPOIS DE ATINGIR UM CERTO GRAU DE ESTABILIDADE, SEFA DESENVOLVER O PRÓXIMO PASSO: A CONSTRUÇÃO DE UMA REDE SEMÂNTICA DE UM TEXTO. NESTA REDE, OS SUBSTANTIVOS (OU UM SUBGRUPO DELES, P. EX. OS ELEMENTOS DE UM TESAURO), FORMAM OS NÓS, E OS VERBOS SERÃO OS ARCOS. ALGUNS VERBOS PODEM FUNCIONAR DENTRO DE UMA SUA 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 À PEQUENA REDE FORMADA PELA ANÁLISE LINGÜÍSTICA APLICADA À FRASE-PERGUNTA. SERÁ ATRIBUÍDO O PESO AO DOCUMENTO À 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 À 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 LINGÜISTIC 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 LINGÜISTIC ANALYSIS IN THE PRESENT MODEL ARE:

- SEGMENTATION (PHRASES, WORDS)
- DICTIONARY SEARCH AND MORPHOLOGICAL ANALYSIS (PORTUGUESE)
- DISAMBIGUATION OF SYNTACTIC HOMOGRAPIHS
- CONSTRUCTION OF A DEPENDENCY TREE
- PRONOUN ANALYSIS.

WITH THESE FEATURES, VARIOUS ITEMS 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 SERIES OF TESTS WITH VARIOUS DATA BASES; FURTHER DEVELOPMENT IS DIRECTED TO THE CONSTRUCTION OF A SEMANTIC NETWORK.

REFERÊNCIAS

- RELY, N. / ROMILLO, A. / VARELA, J. / SIOT-DECAUVILLE, N.: PROCEDURES D'ANALYSE SEMANTIQUE APPLIQUEES A LA DOCUMENTATION SCIENTIFIQUE CNRS - PARIS 1970
- RIDEMAN, P.T.: LINGÜÍSTICA COMPUTACIONAL, UM DESAFIO DE TRINTA ANOS EM: DADOS E IDEIAS, JUNHO/JULHO 1977, PAG. 29-39
- FISCHER, H.G.: C O N D O R , AN INTEGRATED DATA-BASE INFORMATION AND RETRIEVAL SYSTEM FOR STRUCTURED AND UNSTRUCTURED DATA SIEMENS FORSCH. UND ENTW.-BERICHTE BERLIN/NEW YORK 1981
- HUTCHINS, M.J.: PROGRESS IN DOCUMENTATION - MACHINE TRANSLATION AND MACHINE-AIDED TRANSLATION EM: JOURNAL OF DOCUMENTATION, VOL. 34, NO. 2 JUNE 197
- SALTON, G.: AUTOMATIC INFORMATION ORGANISATION AND RETRIEVAL NEW YORK 1968
- SCHANK, R.C./COLEY, 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