UM MÉTODO APROXIMATIVO PARA A SOLUÇÃO DE HOMOGRAFIAS EM TEXTOS DE LINGUAGEM NATURAL:

UMA CONTRIBUIÇÃO PARA UMA MELHOR COMUNICAÇÃO HOMEM-MÁQUINA

Johann Haller
Departamento de Lingüística
Universidade de Brasília

RESUMO

As linguagens naturais ainda não são aplicadas em grande escala na comunicação entre o homem e o computador. Uma das razões é o fato de que as linguagens naturais possuem muitos elementos com vários significados e também com diferentes funções sintáticas possíveis.

O homem resolve estas ambigüidades com a redundância da linguagem natural em todos os níveis lingüísticos e com os seus conhecimentos do mundo real e da situação concreta; o computador precisa, pelo menos enquanto não tem órgãos "sensitivos" desenvolvidos, de algoritmos formais atuando na forma de uma frase ou de um texto que lhe permita resolver estes problemas.

Como os conhecimentos do computador são restritos a regras e a elementos lingüísticos (dicionários, regras morfológicas e sintáticas, regras semânticas), ele tem que aplicar métodos diferentes para resolver ambigüidades como as seguintes da língua portuguesa:

— Nada: Pronome ou 3º pessoa sing. do verbo "nadar"?
— Pelo: Contração de preposição e artigo ou substantivo?
— Posto: Particípio do verbo "pôr" ou substantivo?

O presente trabalho mostra um algoritmo aproximativo (que trabalha em vários ciclos através da frase), sendo implantado no projeto de processamento de textos do autor (HALLER,

Letras de Hoje, 17 (2): 131 a 143

1982), que já exibe resultados bastante satisfatórios nas aplicações da indexação automática, do acesso em linguagem natural a bases de dados e na construção de glossários automáticos numa língua estrangeira.

INTRODUÇÃO: PROCESSAMENTO DE TEXTOS EM LINGUAGEM NATURAL

Como foi descrito pelo autor (HALLER, 1982), o processamento de textos em linguagem natural é uma das tarefas da lingüística computacional que começou com as primeiras tentativas de tradução automática nos anos cinqüenta. Como as esperanças eram exageradas, estas tentativas foram abandonadas e começaram as pesquisas mais básicas na lingüística, na informática e na inteligência artificial.

Ao mesmo tempo, surgiram outras aplicações possíveis do processamento de textos em linguagem natural, como a indexação automática, a correção automática de erros de datilografia, aplicação no ensino programado, na comunicação homem-máquina em geral, na automatização de escritórios, etc.

No Brasil, foram poucas, até hoje, as pesquisas no ramo da lingüística computacional. Nas primeiras tentativas são do ITA (pesquisa sobre a entropia da língua portuguesa), dos professores Maria Tereza Biderman e Paltônio Daun Fraga (BIDERMANN, 1977).

Recentemente, surgiu uma equipe na PUC/RJ, sob a coordenação do Professor Vitoriano Ruas e do Professor Andreewsky, da Universidade Paris-SUD, que querem realizar uma adaptação do sistema SPIRIT (em língua francesa) ao português (ANDREEWSKY, 1983).

Na Universidade de Brasília está sendo desenvolvido, desde 1980, um programa experimental de análise automática da língua portuguesa; existe uma versão preliminar em linguagem COBOL na Burroughs 6700 do CPD da UnB. Colaboram com o autor vários alunos de Pós-Graduação da Lingüística, da Biblioteconomia e do Processamento de Dados.

Um dos problemas maiores na análise da língua natural é o problema da Homografia e Homonímia, isto é, os elementos da língua humana não têm, na sua maioria, só uma função e um significado, como por exemplo nas linguagens de programação. Existem ambigüidades em todos os níveis, onde um texto pode ser analisado: na morfologia, na sintaxe e na semântica.

A mente humana usa duas classes de recursos para resolver este problema: intralinguais e extralinguais.

Os recursos intralinguais seriam a capacidade de falar e entender que a criança adquire com o crescimento quando ela imita as outras pessoas, junta os elementos adquiridos para novas frases e entende, assim, até frases que nunca ouviu. Mais tarde, na escola, se pretende aprender as regras explícitas da língua materna nos níveis morfológico, sintático e semântico, esta aprendizagem se torna absolutamente necessária quando alguém quer aprender uma língua estrangeira.

Os recursos extralinguais seriam o conhecimento do "mundo real" e a percepção da situação na qual se realiza a comunicação.

A lingüística computacional tenta formular as regras do primeiro grupo para poderem ser aplicadas a textos escritos para fins de indexação automática, tradução automática, etc.

As pesquisas nesta área alcançaram bons resultados na morfologia que trabalha com dicionários e tabelas das terminações possíveis para atribuir as funções sintáticas possíveis às palavras do texto.

Já para uma análise completa é necessário decidir qual das funções possíveis esta sendo realizada no contexto da frase.

Este problema tem sido atacado inicialmente junto com a análise das cadeias sintáticas se elas são corretas ou não. Como uma análise desta maneira é difícil de desenvolver (ela tem que voltar muitas vezes de caminhos errados) e também por razões econômicas, vários projetos de lingüística computacional optaram para um passo à parte, onde as homografias são desambiguadas.

Duas destas tentativas serão brevemente descritas no capítulo 2. Esta análise tem várias conseqüências também na língua portuguesa para os níveis seguintes de semântica e pragmática (que já seria um campo para os recursos da segunda classe).

Assim, as palavras

Estado	NOM/PAR	(Substantivo / Particípio)
Pelo	NUM/PRP	(Substantivo / Preposição)
Nada	VRB/PRN	(Verbo / Pronome)

ganham um sentido totalmente diferente dependendo da classificação sintática que um programa de análise automática atribuiria a elas: as traduções para uma língua estrangeira seriam diferentes, a importância delas num sistema de informação também.

Por esta razão, a desambiguação é um instrumento importante nesta análise automática; muito mais ainda se se toma em conta que os recursos da segunda classe (extralinguais) que poderiam ajudar na tradução ou indexação automática ainda só existem em pequenos sistemas de inteligência artificial ("mini-mundos").

A lingüística computacional tem que explorar ao máximo as regularidades formais da língua escrita se ela quer fornecer uma contribuição à melhor comunicação homem-máquina.

1. A HOMOGRAFIA NA LÍNGUA PORTUGUESA

1.1. Morfemas Multifuncionais

Na língua portuguesa ocorre muitas vezes que um substantivo deverbal coincida com uma forma do próprio verbo, sobretudo no caso das seguintes terminações:

— O	Abandono	
	Acordo	
	Apoio	etc.
— A(s)	Busca(s)	
	Ajuda(s)	etc.

Na maioria dos casos não resulta nenhuma diferença semântica, muito embora a mudança sintática possa adquirir importância na análise das relações entre os elementos da frase; às vezes, porém, são coisas bem diferentes que são designadas com as mesmas letras:

Caso	
Cobra(s)	
Faça	(subjuntivo de fazer)
Mata	etc.

Aqui é de suma importância resolver a ambigüidade sintática, tanto para um sistema de indexação quanto para uma tradução automática, dois casos típicos de processamento de textos em linguagem natural.

1.2. Palavras Multifuncionais

Além de morfemas multifuncionais, existem em português também palavras que podem ter várias funções. Estas palavras são homógrafas somente dentro da classe fechada das palavras "funcionais" que não têm valor semântico em si e não seriam considerados nunca descritos, como por exemplo

Tenho	(Verbo Auxiliar - Verbo Normal)
Estou	
Estamos	
Esta	(Verbo Auxiliar - Verbo - Verbo Normal - Pronome)
etc.	

Para uma análise sintática completa da frase como ela é necessária, por exemplo, para a tradução automática, também estas palavras têm que ser desambiguadas. No caso da indexação, teriam mais importância os seguintes casos onde existe uma homografia entre uma função sintática sem conteúdo semântico e outra proveniente de uma palavra que pode bem ser descritor:

Cedo	VRB/FAV	(Verbo - Advérbio)
Cerca	NOM/PRP	(Substantivo - Preposição)
Como	VRB/CON	(Verbo Conjunção)
Era	NOM/AUX	(Substantivo - Verbo Auxiliar)
Estado	NOM/PAR	(Substantivo - Particípio)
Nada	VRB/PRN	(Verbo - Pronome)
Para	NOM/VRB/PRP	(Substantivo - Verbo - Preposição)
etc.		

Nos sistemas de informação que trabalha com os textos inteiros (p. ex. STAIRS da IBM), estas palavras levam a uma precisão muito pequena porque causam a recuperação de muitos documentos que não pertencem à pergunta.

2. A DESAMBIGUAÇÃO DA HOMOGRAFIA

2.1. Parsing

O método que foi usado nas primeiras tentativas da análise automática de textos foi o "parsing" direto depois de identificar as palavras no dicionário e atribuídas todas as funções sintáticas possíveis, foram criadas todas as cadeias que resultaram das combinações de tais funções sintáticas. Cada ca-

dela foi analisada se era correta segundo uma gramática que existia no computador.

Outro procedimento trabalhava com uma predição: o que seria o mais provável num certo ponto da frase? Foi criada uma nova hipótese sobre a estrutura da frase a cada passo para frente ou para trás, se a análise começou com a último elemento como é preferível em várias línguas (KUNO, 1966).

O problema comum desses procedimentos é o alto número de possibilidades (de cadeias ou "caminhos") que deve ser analisado e que se deve, como foi dito no começo, ao fato de os elementos da linguagem natural tenderem a ser ambíguos (média matemática = 2,5 funções por palavras em porturguês).

Além de ser pouco econômico, este fato implica numa grande dificuldade de formulação do algoritmo e do programa, já que é muitas vezes necessário voltar atrás numa decisão e retomar a análise do último ponto que podia ser decidido com alguma segurança.

2.2. Dois métodos usados em sistemas de análise automática de textos

2.2.1. Satan (Saarbruecken)

Um projeto da Universidade de Saarbruecken (Alemanha Ocidental) para a análise de textos em língua alemã está descrito em Zimmermann (1980); "Satan" significa "Sistema de Análise Automática de Textos de Saarbruecken" — **S**aarbrue**c**kener automatische **T**extanalyse".

No passo da análise que reduz as homografias, são tratadas primeiro algumas palavras especiais que têm muitas funções. Elas correspondem às palavras portuguesas "como", "que", "o", "a", etc. Cada vez que se constata a impossibilidade de uma função num determinado contexto, tal função se exclui.

Este processo aplica-se passando pela frase da esquerda à direita e se repete até que não surja mais possibilidade de exclusão. As outras palavras são classificadas segundo um esquema muito detalhado com mais de 150 categorias sintáticas. Entre estas categorias existem também regras de exclusão; porém, a elaboração manual destas regras torna-se difícil. Em cada língua existem casos especiais que não acontecem com alta freqüência, mas que é necessário considerar. Se as regras são muitos amplas, sobram muitas homografias, se elas

ficam detalhadas, pode acontecer que se exclua uma seqüência permitida.

Para resolver esse problema atribui-se uma "probabilidade" a cada par de classes que se estabelece depois de análises estatísticas de textos, um método empírico que será aperfeiçoado no sistema SPIRIT (ver o seguinte capítulo).

Mesmo assim, sobram ainda bastantes cadeias de categorias para as quais se calculam as probabilidades de serem corretas, "geralmente", diz Maas (apud ZIMMERMANN 1980, p. 112), "a cadeia correta acha-se nas primeiras 12 cadeias apresentadas em ordem de probabilidades". O resto das decisões deixa-se para a análise sintática propriamente dita que segue um procedimento semelhante à análise descrita no capítulo 2.1., só que já se evita uma grande parte dos caminhos errados.

2.2.2. Spirit

O sistema SPIRIT (sistema para indexação e recuperação de informações textuais) foi desenvolvido pelo francês A. Andreewsky na Universidade Paris-SUD para a língua francesa. Ele está sendo testado por uma equipe da PUCRJ com respeito à sua aplicabilidade à língua portuguesa (ANDREEWSKY 1983).

O método com o qual ele tenta resolver o problema das ambigüidades sintáticas é chamado de "aprendizagem". Para este método é necessário criar um dicionário com todas as palavras dos textos que vão servir para esta "aprendizagem". No dicionário tem que constar todas as funções sintáticas possíveis de cada palavra. Em seguida, se atribui a cada palavra dentro do texto a função sintática atual e o texto entre no processo da "aprendizagem". Um programa extrai regras binárias, às vezes ternárias para a escolha da função certa para cada palavra. No começo, surgem mais regras binárias que com o crescimento do "corpus" aplicado tendem a se converter em ternárias porque na maioria dos casos não é suficiente considerar só um dos dois vizinhos da palavra.

Um exemplo seria a frase seguinte

Eu me **caso entre** a Páscoa e o fim de maio,
caso entre a curto prazo para essa firma.

Para desambiguar as palavras sublinhadas, até as regras ternárias ficam ambíguas, mas a decisão é feita combinando as regras ternárias que incluem a primeira palavra (caso) como último e a segunda palavra como primeiro elemento.

As conseqüências deste procedimento são as seguintes:

Da mesma maneira que em SATAN é necessário classificar muito detalhadamente as palavras da língua portuguesa. Juntando este fato à necessidade de haver todas as palavras dentro do dicionário, resulta um trabalho considerável na construção do dicionário. Como este trabalho tem que ser executado por várias pessoas, cresce a possibilidade de erros e inconsistências. Os recursos necessários para a adaptação deste método à língua portuguesa devem ser bastante elevados.

3. O MÉTODO DE APROXIMAÇÃO NO PROJETO "LINGA"

O método aplicado no projeto LINGA se baseia em trabalhos do autor numa pesquisa realizada na empresa Siemens, em Munique (Alemanha), descrita em FISCHER (1981) e HALLER (1981A e 1981B) que foi aplicada às línguas alemã e espanhola.

3.1. **Filosofia básica**

A filosofia básica da sub-rotina "DISAMB" no projeto LINGA está na aplicação conseqüente do princípio da aproximação e numa classificação mais simplec do que as acima descritas. LINGA trabalha somente com 32 categorias que podem ser aprendidas rapidamente, facilitando a correção dos resultados. Todos os passos descritos no próximo capítulo se repetem da primeira à última palavra até não mais ser possível nenhuma exclusão. As poucas homografias restantes são decididas à base de probabilidades de categorias soltas (não de pares), como descrito em 3.3.3.

3.2. **Descrição do algoritmo**

3.2.1. Eliminações na 1º posição

Na primeira posição da frase podem ser eliminadas as seguintes categorias:

PAR REL AUI MD1 PA2 CMP OPR

se elas fizerem parte de um conjunto, isto é, se se tratar de uma palavra homógrafa.

3.2.2. Eliminações com respeito aos vizinhos

Cada palavra está sendo testada qual das suas funções não é possível no conceito atual. Para este fim, é criada uma

tabela com os possíveis vizinhos de esquerda e de direita de cada função. Se uma função não encontra um desses vizinhos possíveis de um lado, ela é eliminada. Na medida em que um homógrafo é "reduzido", isto é, o número das funções possíveis diminui, fica mais fácil no próximo passo resolver, também, as homografias vizinhas.

Este passo resolve aproximadamente 60% das homografias de um texto normal — e, o que é muito importante, ele não pode tomar uma decisão que não seja segura.

3.2.2.1. Homografia ATT/NOM

Tenta-se, depois deste passo, avançar na desambiguação com algumas regras destinadas a certas combinações de funções, por exemplo a homografia ATT/NOM que é um caso especialmente difícil em português porque todos os adjetivos podem funcionar como substantivos e quase todos os substantivos podem ser acrescentados a outro substantivo como qualificador — a mesma distribuição como os adjetivos atributivos ("Programa-Padrão", "Imposto-Calamidade", etc.). Além disso, tem muitos adjetivos que costumam preceder os substantivos. Os gramáticos comentam esse fato como pertencente à estilística: o adjetivo que precede tem um valor emotivo, etc. Existem poucos adjetivos que nunca podem preceder o substantivo, como, p. ex., cores.

Tenta-se resolver este problema com algumas regras simples e com uma lista dos adjetivos que freqüentemente (nos textos tratados) precedem os substantivos.

3.2.2.2. Homografia com VRB

As combinações que são tratadas com regras especiais são organizadas em dependência de algumas categorias principais.

O primeiro grupo são as homografias que contêm a função do verbo conjugado (VRB). São consideradas as seguintes combinações:

VRB/QAT	Sua
VRB/CON	Como
VRB/PRP	Entre, Para,...
VRB/VBM	Vou, Vamos,...
VRB/NOM	Trabalho, Precisa (Ver Cap. 1.1.)
VRB/QAT/NML	Esta, Estas,...

Deve-se insistir no fato que até agora não se trata de palavras isoladas; todas essas regras ainda valem para qualquer palavra que tenha exatamente tais funções. Assim, a regra vale também para novas palavras que podem ser inseridas no dicionário o que vai ser o caso durante a construção do sistema. É claro que se trata de classes fechadas de palavras que, depois de processar muitos textos, não vão crescer mais.

3.2.2.3. Outras combinações

As outras combinações tratadas com regras especiais são as seguintes:

PRP/DET	A, As, ...
PRP/NOM	Pelo, Pelos, Para, ...
CON/SUB	Como, ...
SUB/OPR	Se
ATT/VRB/PRP	Salvo, ...

Logicamente, algumas palavras podem aparecer em várias regras se elas têm como funções possíveis o conjunto comum das funções tratadas em cada regra.

Neste parágrafo chama-se a atenção sobre tudo para as palavras que somente são homógrafas por causa das maiúsculas e da falta de acentos; muitas homografias serão resolvidas logo se o texto for escrito com sinais, os quais são codificados dentro do sistema de maiúsculas com números que seguem e vogal:

PARA	PRP
PA1RA	VRB
PARA1	NOM
ETC.	

3.2.2.4. Palavras explícitas

Algumas palavras, porém, têm que ser tratadas explicitamente por terem muitas funções na língua portuguesa ou, pelo menos, terem funções muito diferentes sintaticamente. São estas:

Que	REL/AVP/SUB/OPR
Vão	NOM/ATT/VRB/VBM
Caso	NOM/VRB/SUB

São também estas palavras que ainda apresentam algumas dificuldades na resolução das homografias (ver cap. 3.3.2.).

3.2.3. Afirmações (prioridades)

Como vemos no anexo 5.2., a maioria das homografias é resolvida já nos passos até agora descritos.

Porém, temos que tomar uma decisão também para o restante das palavras. Esta decisão se baseia simplesmente na ordem numérica das funções, onde as funções com valor semântico precedem as outras.

Assim, quando sobra uma homografia NOM/ATT será excluída a função ATT (Adjetivo em função atributiva) para não perder o candidato a descritor no caso da aplicação na indexação automática.

Neste caso, atualmente são listados somente os substantivos como candidatos a descritores. Para a criação de descritores compostos que está previsto como próximo desenvolvimento, será necessário diferenciar em todos os casos entre ATT e NOM. Esta tarefa será realizada junto com a análise dos grupos nominais, seguindo, assim, numa pequena parte a filosofia mencionada do SATAN (cap. 2.2.1.), só que o número de possibilidades será muito menor.

Em alguns casos específicos muda-se a preferência da ordem numérica:

ATT/VRB	VRB
VRB/PRP	PRP
INF/PRP	PRP
NOM/PRD	PRD

Estas regras são conseqüências de contagens estatísticas em vários textos; a língua portuguesa manifesta — assim parecem indicar as regras 2 e 3 uma preferência pelo grupo nominal com vários complementos proposicionais, pelo menos nas variantes de textos científicos.

3.2. **Tese**

O programa LINGA que contém a desambiguação das homografias foi testado até agora com 3 textos de natureza científica e jurídica: um texto geral do próprio autor sobre o processamento de textos em linguagem natural ("artigo"), normas legais do Banco Nacional de Habitação ("BNH") e um extrato do regulamento de caça ("MBI"), que contém muitas palavras homógrafas. Foi provavelmente por esta qualidade que ele foi

selecionado como teste também no sistema SPIRIT nas tentativas de adaptação e língua portuguesa.

3.3.1. Erros

A maioria dos erros constatados na versão atual do programa se concentra nas seguintes áreas:

— Palavra "E"; como os textos vêm exclusivamente em minúsculas. É difícil decidir se a palavra "E" é um verbo auxiliar ou uma conjunção; as regras de preferência estatística não oferecem ajuda.

— Homografias com verbo: em textos portugueses não se pode confiar nas vírgulas, isto é, muitas vezes falta a vírgula onde deveria ser colocada e vice-versa:

"Antes que entre na área de caça(,) o caçador pára obrigatoriamente no posto."

Nesses casos será necessário "abrandar" as regras de exclusão da função VRB.

— Abreviações: como o programa ainda não trabalha com um reconhecimento das abreviações, várias vezes a frase termina com um ponto depois de "ETC.;" neste caso, as regras de exclusão podem levar a uma falsa decisão.

— Formas verbais com hífen ("devolvê-la") ainda não são tratadas e são analisadas como substantivos (NOM).

Todas estas classes de erros não têm muita influência para o primeiro passo da aplicação na indexação automática, o qual é a extração automática de candidatos a descritores em forma normalizada. Porém, será possível eliminar todas elas sem grande esforço na próxima versão do programa.

3.3.2. Casos difíceis

Em alguns casos específicos, os erros acontecem por conta da última decisão que se toma só à base da ordem numérica das funções; mesmo depois de passar por todos os algoritmos descritos, ficam várias funções atribuídas a uma palavra. Isto acontece sobretudo no caso de grupos nominais mais complicados (por exemplo com vários atributos conectados), onde o contexto imediato não é suficiente para tomar uma decisão segura:

"Cobra venenosa. É perigosa"

Neste caso, nem o leitor humano pode decidir qual é a função do "E" sem olhar para o resto da frase.

Estes casos serão deixados à mencionada análise dos grupos nominais de uma frase a qual resolverá também casos como este:

"Muitas regras semelhantes aplicadas..."

Será possível também examinar os resultados morfológicos para cada palavra e testar as concordâncias dentro destes grupos.

4. CONCLUSÃO

Temos na análise descrita um valioso instrumento para o processamento de textos em linguagem natural, mais particularmente para a resolução do problema das palavras sintaticamente ambíguas.

Este instrumento é indispensável em várias aplicações da análise automática de textos:

— Correção automática de erros da datilografia (fornecendo listas alfabéticas de todas as palavras não existentes nos dicionários e/ou não tendo terminação possível em português).

— Indexação automática (fornecendo candidatos a descritores em forma normalizada com a freqüência dentro do texto ou de um grupo de textos).

— Tradução automática (fornecendo as palavras lexicalizadas com informações morfológicas e sintáticas para a construção da frase na língua estrangeira.

É possível realizar esta tarefa com a aplicação de regras lingüísticas com um esforço econômico relativamente pequeno, sem um dicionário explícito dos elementos da língua e, em conseqüência, depois de um tempo de desenvolvimento bastante reduzido.

O programa foi testado em 1983 com bases de dados bem maiores, o que levará a um aperfeiçoamento, junto com a incorporação das medidas descritas no capítulo 3.3.1.