COMUNICAÇÕES

INDEXAÇÃO AUTOMÁTICA BASEADA EM MÉTODOS LINGUÍSTICOS E ESTATÍSTICOS E SUA APLICABILIDADE À LÍNGUA PORTUGUESA*

Alexandre Andreewski

Centre National de Ia Recherche Scientifique, Université de Paris Sud, França.

Vitoriano Ruas

Departamento de Informática, Pontifícia Universidade Católica, Rio de Janeiro.

RESUMO

Considera-se neste artigo a indexação automática usando o processamento de documentos em linguagem natural, que é obtido com o auxílio de métodos linguísticos combinados com métodos estatísticos permitindo uma indexação ponderada. A título ilustrativo descreve-se em linhas gerais, um sistema de indexação desse género denominado SPIRIT, o qual foi desenvolvido para o idioma francês por uma equipe de pesquisadores do CNRS. Enfim, são tratados aspectos essenciais de sua adaptação à língua portuguesa.

Descritores: Ambiguidade; Análise sintética; Entropia; Estatística; Filtros; Indexação automática; Indexação ponderada; Linguística; Matrizes de precedência; Método de aprendizado; Proximidade; Relações léxico-semânticas.

1 - INTRODUÇÃO

Define-se indexação automática como a técnica de processamento eletrônico de documentos visando a sua recuperação a partir de informações relativas ao seu conteúdo. Trata-se mais especificamente de obter os documentos que contêm o maior número de informações relativas a uma dada pergunta do usuário.

Vamos considerar neste artigo, técnicas de indexação automática baseadas em métodos linguísticos e estatísticos, com o objetivo de processar os documentos em linguagem natural. Tomamos como referência o sistema SPIRIT**, que é um sistema de indexação automática desse tipo, desenvolvido pelo primeiro autor em colaboração com P. Binquet, F. Debili, C. Fluhr e B. Pouderoux do Centre National de Ia Recherche Scientifique (CNRS) francês. Ele permite assim o armazenamento e a interrogação em linguagem natural e, com os tratamentos linguísticos a todos os níveis dos textos introduzidos no sistema, aliados a tratamentos estatísticos, permite ainda a realização de uma indexação ponderada dos documentos. Desta forma, em resposta a uma pergunta formulada ao sistema, também em linguagem natural, os documentos — resposta são classificados segundo um critério de proximidade semântica. As únicas intervenções manuais são as que dizem respeito à correção dos erros tipográficos, que, aliás, são detectados automaticamente pelo sistema.

* Trabalho parcialmente financiado pela F INEP.

** Systéme Syntaxique et Probabiliste d'Indexation et de Recherche d'Informaticos Textuelles.

2 - COMPONENTES DO SISTEMA

O sistema SPIRIT contém os componentes seguintes:

1º) Um dicionário (com ± 250.000 formas em francês) que permite a análise morfológica dos textos. Em particular, ele permite o reconhecimento da sinonímia, das variações paradigmáticas de uma palavra, tais como as formas conjugadas dos verbos, as variações em género e número de adjetivos, substantivos, etc., além de expressões idiomáticas como "por causa de". Além disso, ele fornece todos 'os valores gramaticais de uma dada palavra como por exemplo "para", que tanto pode ser um verbo conjugado como uma preposição.

2º) Algoritmos de análise sintática: Lembrando que entende-se por sintaxe o conjunto de regras que

estabelecem as configurações de categorias de palavras consideradas corretas, a análise sintática é fundamental para determinar a categoria correta no texto de uma palavra ambígua como "para".

Observação: Trata-se, portanto, de um nível inferior de análise sintática, sendo que os níveis mais elevados são menos importantes para o sistema SPIRIT.

3º) **Algoritmos de análise semântica:** Lembra-se que, de uma maneira geral, a semântica trata do conjunto de sistemas conceituais da língua. No sistema SPIRIT ela se reduz à identificação correta da relação palavra-designado que tenta-se determinar em função do contexto. Este último é levado em conta graças às relações ditas **léxico-semânticas,** tais como, sujeito-verbo, verbo-complemento, substantivo-complemento, substantivo-adjetivo, etc., as quais são obtidas automaticamente por métodos ditos de **"filtragem".** Desta forma, por exemplo, "papel de carta" e "papel de vilão" serão automaticamente identificados como dois conceitos diferentes.

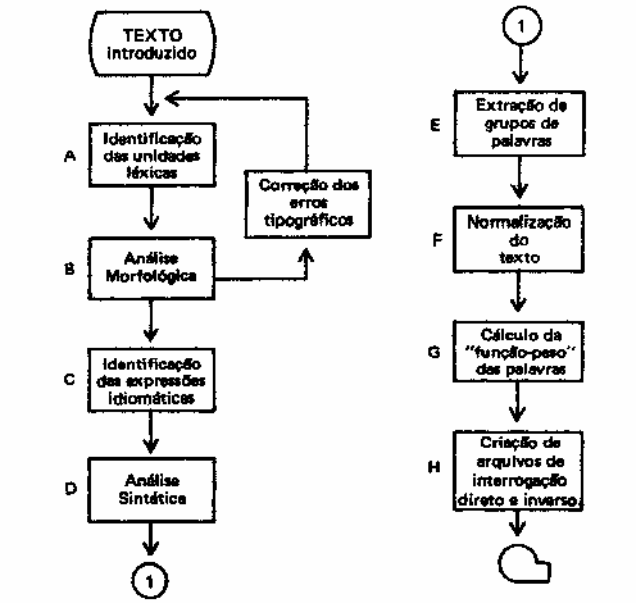
3 - FLUXOGRAMAS DE PROCESSAMENTO DE UM CORPUS* E DE UMA QUESTÃO

A cada etapa do processamento com o sistema SPIRIT representado no fluxograma abaixo, associamos uma letra que permitirá a referência à mesma no resto do artigo. Exceto as etapas D, E, G e I que constituem, juntamente com as **hipóteses de trabalho** que se adotou tratamentos específicos do sistema SPIRIT, as demais são comuns a todos os sistemas de indexação automática, a menos de variações não essenciais.

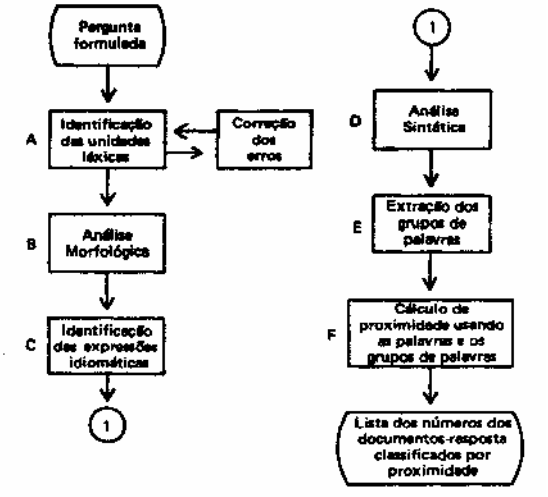
Vamos então nos deter em explicações sobre os pontos D, G e I. Sobre o ponto E daremos apenas alguns aspectos essenciais, já que sua descrição pormenorizada exige considerações bastante extensas, explanadas nos artigos das referências bibliográficas 1 e 2. Salientamos entretanto, que mesmo sem levar em conta a análise semântica, é possível obter respostas bastante pertinentes e satisfatórias.

* Um conjunto de documentos.

3.1 Tratamento de um Corpus



3.2 Tratamento de uma pergunta em linguagem natural



Observa-se que, exceto o ponto 1 no fluxograma acima, todas as outras etapas do tratamento da pergunta são idênticas às do texto introduzido.

No quadro a seguir fornecemos sucintamente o objetivo e o meio usado pelo sistema SPIRIT para executar cada etapa do processamento:

	Objetivo	Meio
A	. Determinar as unidades léxicas como "guarda-chuva", "dar" e "se" em "dar-se"	Separadores: branco, hífen, apóstrofo, etc.
B	. Obter todas as categorias possíveis de uma palavra, reconhecer sinônimos e obter o representante de classe no caso de variação paradigmática de uma palavra.	Dicionário
	. Reconhecer os grupos de palavras que correspondem a uma expressão idiomática como "por causa de"	Dicionário
O	. Determinar a classe (categoria gramatical) correta das palavras no texto introduzido, preparar e facilitar a análise semântica.	Regras de precedência de classes de palavras, de tipo binário, ternário, etc. (essas regras são armazenadas sob a forma de matrizes) *
E	. Extrair do texto os grupos de palavras que que têm valor de conceito.	As matrizes mencionadas acima e de um modo geral, as relações léxicas mais importantes representadas por filtros lingüísticos*
F	. Eliminar as palavras supérfluas (isto é, as que não têm valor informativo) tais como artigos, preposições, conjunções, etc. . Substituir uma palavra de variação paradigmática pelo seu representante de classe.	. Dicionário
G	. Calcular o peso informativo de uma palavra.	. Contagem de freqüência das palavras em cada documento. . Função de entropia*
H	. Criar os arquivos de interrogação. Direto: classificado por documento Inverso: classificado por palavra.	. Classificação das palavras por ordem alfabética dentro do documento no direto, e incluindo sua entropia e os documentos em que figura, no inverso.

Nota: Os elementos assinalados com asterisco serão vistos de maneira mais detalhada no decorrer do artigo.

4 - ANÁLISE SINTÁTICA

Ela é fundamental para levantar todas as ambigüidades existentes numa língua, por causa do emprego de uma mesma palavra em funções distintas. Esse é essencialmente o fenômeno da homografia.

Como o computador encontrará no dicionário duas ou mais categorias para tais palavras, será necessário levantar a ambigüidade, procurando-se determinar a categoria gramatical correta da palavra que corresponde ao seu emprego no texto.

Vejamos um caso típico:

"Como são as travas para a sua cerca?"

Cada uma das palavras da frase acima é ambígua *e,* consultando o dicionário se encontrará:

Como	: advérbio, conjunção, verbo conjugado	(3)
são	: verbo conjugado, substantivo, adjetivo	(3)
as	: artigo definido, pronome	*(2)*
travas	: substantivo, verbo conjugado	(2)
para	: preposição, verbo conjugado	(2)
a	: artigo definido, pronome, preposição	(3)
sua	: pronome possessivo, verbo conjugado	(21
cerca	: substantivo, verbo conjugado, preposição*	(3)

Como seria preciso proceder para poder decidir que categoria correta tem cada palavra na frase em termos de processamento?

A abordagem mais chegada aos lingüistas, que preferem em geral uma análise exaustiva da língua, corresponderia a examinar todas as combinações possíveis das categorias e a eliminar as que são falsas por comparação com uma espécie de repertório de construções corretas da língua. Mas basta observar um exemplo simples como a frase acima para se constatar que tal procedimento conduziria ao exame de 1.296 combinações diferentes! Fica portanto evidente que tal abordagem é proibitiva do ponto de vista computacional, ainda mais porque as frases encontradas normalmente nos textos podem ser muito mais longas. Ademais, armazenar, todas as combinações corretas de categorias gramaticais é completamente absurdo.

Por outro lado, é um fato incontestável que qualquer observador que conheça suficientemente a língua entende a frase acima corretamente, o que faz supor que em seu cérebro são ativados mecanismos para levantar a ambigüidade de cada palavra da cadeia, em função do contexto, isto é, das que a precedem ou a seguem. No sistema SPIRIT esse fato essencial foi levado em conta e de certa forma simulado. Na realidade, fez-se dele uma hipótese de trabalho da maneira ilustrada a seguir. Observamos que, por motivo de economia e de eficiência, somente o "contexto imediato" de cada palavra é considerado.

* Para ser rigoroso esta categoria é elemento da locução prepositiva "cerca de".

Seja então o exemplo:

'Tu nos deste uma grande ajuda nos momentos difíceis deste fim de mês".

As palavras "nos" que aparecem na frase acima são ambíguas. Entretanto a presença do pronome "tu" (não ambíguo) antes do primeiro "nos" e do substantivo "momentos" (não ambíguo) depois do segundo "nos" é suficiente para que se deduza que o primeiro é pronome (objeto indireto) e o segundo é a contração "em + os".

Se além disso tivéssemos que levantar a ambigüidade das palavras "deste" (além de "ajuda", etc.), constataríamos que não é preciso mais do que as "uma ou duas" palavras vizinhas para determinar suas categorias corretas.

É por esse motivo que matrizes de precedência foram introduzidas no sistema. Nessas matrizes cada linha corresponde à categoria gramatical da palavra que precede e cada coluna corresponde à categoria da palavra que sucede. Na sua forma mais simples, o termo da matriz será um ou zero. No primeiro caso indica que a relação de vizinhança entre as duas categorias que correspondem à sua linha e à sua coluna é verdadeira, e no segundo caso, que é falsa.

Exemplo: Da frase acima concluímos que

PRONOME SUJEITO-PRONOME OBJETO INDIRETO =1
PRONOME SUJEITO-CONTRAÇÃO PREPOSIÇÃO/ARTIGO = 0
CONTRAÇÃO PREPOSIÇÃO/ARTIGO-SUBSTANTIVO = 1

Nota-se que para podermos escrever o termo da matriz em termos de duas opções, fomos obrigados a criar duas categorias diferentes de pronomes. Na realidade o que ocorre é que, em lingüística, decidir entre "verdadeiro" e "falso" pode ser um procedimento demasiadamente estrito, pois pode-se estar rejeitando uma possibilidade de precedência, entre categorias, verdadeira, embora pouco freqüente na língua. É por isso que a alternativa 1 ou O foi substituída por freqüências de ocorrência da precedência calculada em textos suficientemente longos, mas muito mais curtos do que os que serão efetivamente submetidos ao processamento, uma vez definido um conjunto de categorias de trabalho em número razoável (um número muito grande de categorias poderia se revelar pouco prático, ao passo que um número muito reduzido poderia fazer o método perder a eficiência) após estudo da língua em questão.

Aliás, o uso desses textos serve para estabelecer todas as regras de precedência armazenadas no sistema e dá-se o nome de método de aprendizado ao processo correspondente. É justamente esse método que representa assim um papel essencial no procedimento usado para a construção do sistema SPIRIT, que vamos descrever a seguir.

5 - 0 MÉTODO DE APRENDIZADO

Esse método é uma tentativa de reproduzir automaticamente os mecanismos que se supõe serem criados no cérebro humano, quando do aprendizado de uma língua natural e que, entre outras coisas, permitem decidir se uma construção de frase é correta ou não. Seu uso para levantar ambigüidades no sistema SPIRIT justifica-se pela qualidade dos resultados obtidos.

Procede-se da forma seguinte:

Inicialmente é preciso definir um conjunto de categorias gramaticais em função de estudos lingüísticos prévios, o qual pode ser aumentado tanto incluindo novas categorias, como subdividindo as já existentes, segundo a experiência adquirida e os resultados obtidos. Na sua versão francesa atual há 176 categorias no sistema SPIRIT. O método então funciona da seguinte maneira:

19) Fornece-se ao computador um texto de aprendizado qualquer, contendo inicialmente cerca de 5.000 palavras. Esse texto deve estar resolvido gramaticalmente, o que significa que cada palavra está associada à sua categoria correta no texto.

2º) Constrói-se um dicionário de acúmulo onde as palavras do texto de aprendizado são classificadas por ordem alfabética e seguida de todas as categorias em que aparece no dito texto.

3º) Constrói-se a seguir um texto ambíguo que nada mais é do que o texto de aprendizado com cada palavra seguida de todas as categorias diferentes em que aparece neste último texto.

49) Compara-se enfim o texto ambíguo com o texto de aprendizado e obtém-se automaticamente as regras corretas de precedência (ponderadas por

frequências ou não) binárias, ternárias*, etc. segundo o número de palavras envolvidas na regra.

O processamento das quatro etapas acima pode ser obtido, por exemplo, aumentando-se o texto de aprendizado ou alterando-se a lista de categorias gramaticais, o que permite a obtenção progressiva ou o abandono de regras de precedência.

A título ilustrativo do método, damos abaixo um exemplo de um pequenino texto de aprendizado. Sob cada palavra aparece sua categoria gramatical correta e abaixo desta aparece(m) sua(s) outra(s) categoria(s) no texto entre parênteses, se for o caso.

Eu	me	caso	entre	a	Páscoa
PRONPES	PRONREFL	VERBCONJ (CONJ)	PREP (VERCONJ)	ARTDEF (PREP)	SUBST

e	o	fim	de	maio,	caso
CONJ	ARTDEF	SUBST	PREP	SUBST	CONJ (VERBCONJ)

entre	a	curto	prazo	para	essa	firma.
VERBCONJ (PREP)	PREP (ARTDEF)	ADJ	SUBST	PREP	PRONDEM	SUBST

As palavras ambíguas neste texto são então "caso", "entre" e "a". As abreviaturas usadas para as categorias são, cremos, de interpretação evidente.

Eis o dicionário de acúmulo:

a : ARTDEF, PREP

caso : CONJ, VERBCONJ

curto : ADJ

de : PREP

e :CONJ

entre : PREP, VERBCONJ

essa : PRONDEM

eu : PRONPES

fim : SUBST

maio : SUBST

me :PRONREFL

o : ARTDEF

para : PREP

páscoa: SUBST

prazo: SUBST

O texto ambíguo é o texto de aprendizado, quando se tiram os parênteses e se considera, indiferentemente, uma categoria ou outra para as palavras ambíguas.

Por comparação dos dois textos obtém-se assim as regras binárias, nas quais as resoluções compatíveis aparecem sublinhadas da mesma maneira,tais como:

PRONREFL X (CONJ, VERBCONJ)
(CONJ, VERBCONJ) x (PREP, VERBCONJ)
(PREP, VERBCONJ) x (ARTDEF, PREP)

De maneira análoga, obtém-se as regras ternárias:

(PRONREFL x (CONJ, VERBCONJ) x (PREP, VERBCONJ)
(CONJ, VERBCQNJ) x (PREP, VERBCONJ) x (ARTDEF,PREP)
(PREP, VERBCONJ) x (ARTDEF, PREP) x ADJ

Nota-se que as duas últimas regras binárias apresentadas são ambíguas.

No entanto a 1ª e a 3ª regras ternárias acima levantam completamente essas ambiguidades. Nota-se ,que a 2ª regra acima, embora ternária é ambígua. Mas tal ambiguidade, por sua vez, será levantada pela aplicação em cadeia das três regras ternárias seguindo o texto. Aliás, como a aplicação das regras é sempre encadeada, podem-se levantar, praticamente, todas as ambiguidades nos textos a serem processados,

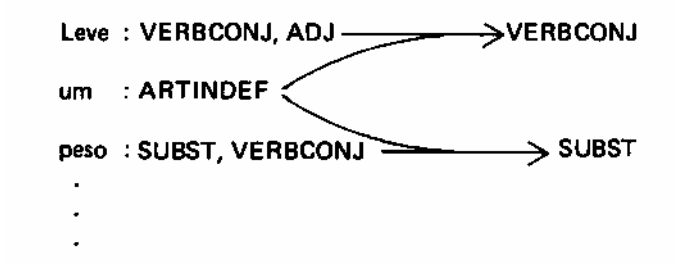
* A experiência mostra que não é necessário construir regras de ordem superior. Na realidade são as regras ternárias que são usadas no sistema SPIRIT, por terem melhores propriedades de inferência que as binárias.

desde que o texto de aprendizado utilizado para a formação das regras seja suficientemente longo (+- 30.000 palavras em francês foram suficientes).

Imaginemos agora que se aplica a matriz de precedência binária, correspondente às regras obtidas por aprendizado ao texto:

"Leve um peso mais leve, que sua cara não sua tanto".

Obtém-se, em sucessão, usando-se indiferentemente a precedência e a sequência:



As setas curvas acima partem da categoria que permite levantar a ambiguidade da palavra vizinha, cuja categoria correta é indicada pela seta reta.

6 - AS RELAÇÔES LÉ X I CO-SEMÂNTICAS

Para se obter respostas ainda mais pertinentes, é conveniente extrair não somente dos textos introduzidos como também das perguntas formuladas ao sistema, os grupos de palavras que têm valor de conceito. Com esse objetivo pode-se proceder também por aprendizado da maneira seguinte: fornecem-se ao computador análises semânticas corretas que correspondem a uma cadeia de categorias sujeita a certas restrições pré-estabelecidas, e ao mesmo tempo usam-se as regras de precedência obtidas pela análise sintática já descrita.

Na maioria dos casos, trata-se de extrair do texto cadeias do tipo C1 P1C2 P2..... CiPi, onde os Pi são palavras vínculo tais como as preposições e os Ci são, quer categorias gramaticais dadas, quer grupos de palavras ou ainda outras cadeias com estrutura fixa. Em geral as cadeias estudadas devem-se situar entre dois separadores como os sinais de pontuação, as conjunções, etc. As cadeias-modelo C1P1C2P2 • • • CiPi, que poderão ou não corresponder a um todo semântico pertinente no texto examinado, chamam-se filtros linguísticos.

Observa-se que os Ci e Pi podem ser contíguos ou não no texto, e caberá aos algoritmos incorporados ao sistema torná-los contíguos. Esse é um problema bastante complexo em geral, mas no caso de termos contíguos pode-se usar técnicas simples como a de marcação com parênteses e com setas. O objetivo dos parênteses é o de agrupar palavras que tem um elo estreito entre si e o das setas é o de indicar as palavras-vínculo. Uma vez conhecidas as categorias gramaticais associadas às palavras envolvidas na cadeia em estudo, obtém-se as regras de formação binárias, ternárias e de ordem superior. Como frequentemente durante o processamento essas regras criam situações contraditórias, apela-se para a Estatística. De fato, aqui também, se se deseja trabalhar com regras demasiado exclusivas, tais situações podem deturpar os resultados.

Vejamos o exemplo da cadeia: "tratamento

em hospital de um município".

Suponhamos que somente a regra seguinte se encontra armazenada:

"Da cadeia "$SUBST_1$ em ARTIGO $SUBST_2$ de ARTIGO SUBST3" extraem-se as relações semânticas: "$SUBST_1$, em ARTIGO $SUBST_2$" "$SUBST_2$ de ARTIGO $SUBST_3$".

É claro que no caso do exemplo acima, obter-se-iam relações verdadeiras. Mas o que ocorreria no caso:

"tratamento em hospital de um ferimento"

A relação verdadeira "tratamento de um ferimento" não será identificada e em seu lugar figurará a relação falsa "hospital de um ferimento".

A melhor solução é pois a de armazenar as duas regras que tem a possibilidade de fornecer relações verdadeiras com pesos respectivos, que são calculados com base nas frequências de ocorrência no texto de aprendizado. Essas frequências, que poderão ser atualizadas a qualquer momento, em função do corpus com que se vai trabalhar, permitem a tomada da decisão acertada no momento da interrogação.

Por exemplo, se a pergunta é:

"Em que casos se pode receber gratuitamente tratamento em hospital de um município de acidente automobilístico"?

Suponhamos que se extraiu a cadeia preposicional seguinte, identificada por meio de um filtro linguístico adequado:

. .. tratamento em hospital de um município de acidente ...

e que se estabeleçam as relações: "tratamento-hospital" e "hospital-município".

A relação verdadeira "tratamento-acidente" não será extraída do texto se apenas relações contíguas *são* usadas (em seu lugar sairia "município de acidente"). Entretanto, com a estatística pode-se passar, a seguir, a uma pesquisa baseada em uma segunda regra, *e* a relação acima será encontrada, embora seu peso seja inferior.

7 - A FUNÇÃO DE ENTROPIA

A entropia H de uma palavra normalizada* p com respeito a um conjunto de N documentos d1, d2, ..., dN, é uma quantidade destinada a avaliar o caráter discriminativo dessa palavra, no sentido que, quanto mais sua entropia é baixa mais informativa é a palavra.

A entropia é dada por:

$$H(d/p) = -Z \sum_{N} P(di/p) \log_2 P(di/p)$$

onde P(dj/p) é a probabilidade de se obter o documento dj, dado que ele contém a palavra p. Tal probabilidade é dada pela fórmula de Bayes:

$$P(dj/p) = \frac{P(P^{/d_i})(P^{rd}l)}{\underline{N}\ 2\ P(p/dj)P(dj)}$$

onde

$$p/n/d-) = \frac{\text{freqüência de p em dj}}{\text{cardinal de dj}}$$

Dois casos extremos facilitam a compreensão da função H:

19) Se p só pertence a um certo documento dj então:

P(dj/p) = 1, P(dj/p) = O se j^i = » H = 0 (mínimo)

2°) Se p pertence a todos os documentos então:

$$(P(d_2/p) = \ldots = P(d_N/p) = \_L = \frac{}{N}$$

H = log2N (máximo)

* Diz-se que uma palavra está normalizada se foi substituída por seu representante de classe, no caso em que ela pode ser considerada como uma variação paradigmática lo mesmo.

A finalidade principal da função H é a de comparar os graus de presença de duas palavras diferentes em um corpus.

8 - CÁLCULO DA PROXIMIDADE DE UM DOCUMENTO COM A PERGUNTA

Com intuito de simplificar o problema, não vamos considerar o caso em que a proximidade é calculada, usando grupos de palavras também.

Como a pergunta é submetida aos mesmos tratamentos que o corpus, todas as suas eventuais ambigüidades se encontram levantadas. O cálculo de proximidade com um documento é então efetuado com base no número de palavras normalizadas em comum com a pergunta, onde cada palavra é ponderada por uma função do complemento da entropia com relação a)og2N e de sua freqüência no documento. Mais especificamente, no sistema SPIRIT usa-se a função PROX definida da seguinte maneira:

Denotando-se por q a pergunta e por CP o cardinal ponderando,

$$PROX(dj) = \frac{CP(djUq)}{n}$$

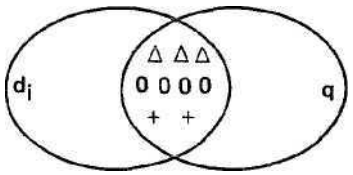
onde $CP(S) = 2\,[\,)og_2N - H(pj) + 1\,] =$

$$= n + S\,[\,\log_2 N - H\,(_{Pj})\,]$$

sendo que Pi é a i-ésima palavra de um conjunto Se *n* = cardinal de S.

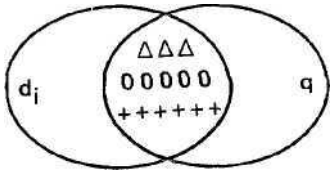
Esquematicamente, pode-se representar o cardinal de di n q com:

Sem considerar a entropia



duas palavras +
quatro palavras O
três palavras A

Considerando a entropia



A figura acima indica que a palavra + é rara, ao passo que A pertence a todos os documentos e O a um número de documentos superior a um.

9 - O PROBLEMA DA ADAPTAÇÃO DO SISTEMA À LÍNGUA PORTUGUESA

Para se proceder a uma adaptação do sistema SPIRIT a uma outra língua, constituem aspectos essenciais, a análise das ambigüidades existentes nessa língua, assim como a possibilidade de levantá-las por métodos de aprendizado que levem em conta as propriedades posicionais de classes de palavras.

Essa adaptação ao inglês e ao espanhol revelou-se perfeitamente factível e está sendo realizada atualmente. No intuito de iniciar uma eventual adaptação à língua portuguesa, realizamos uma pesquisa de vocabulário, visando a determinação dos tipos de pares, trincas, etc. de categorias gramaticais que correspondem a uma palavra ambígua. Redigimos também um texto de aproximadamente 300 palavras, o qual, embora gramaticamente correto, contém um número deliberadamente elevado de palavras ambíguas. As categorias associadas a cada palavra desse texto foram determinadas, a que já permite a construção automática de um número apreciável de regras de precedência, envolvendo um número de categorias básicas.

Quanto aos tipos de ambigüidades que se pode encontrar na língua portuguesa, observamos que como em francês, as mais freqüentes são as do tipo:

SUBSTANTIVO/ADJETIVO como "armada"
SUBSTANTIVO/VERBO CONJUGADO como "arma(s)"
ADJETIVO/VERBO PARTICfPIO como "armado"

Além disso, após consulta de alguns textos de gramática e de dicionários, constatou-se que, considerando-se apenas as categorias de palavras usuais, há apenas cerca de 40 palavras de uso corrente que correspondem a ambigüidades de outros tipos. Esse número é da ordem do dobro, se não se levar em conta acentos gráficos.

Damos abaixo o texto de aprendizado proposto, no qual as palavras ambíguas se encontram sublinhadas uma vez. Se a palavra é empregada no próprio texto em categorias diferentes ela é sublinhada duas vezes.

ESTAÇÃO DE CAÇA 1981

A partir de hoje está aberta a_ estação de caça na reserva da mata municipal. Um e x t r a t o do regulamento a respeito é fornecido abaixo, conforme lei federal n° 9.876 aprovada a 5.7.81.

1. Antes que entre na área de caça o caçador pára obrigatoriamente no posto de fiscalização, onde deve apresentar | guarda sua autorização de caça. Esta a guarda para"devolvê-la ao caçador se este volta são e salvo antes do pôr do_ sol, ou caso contrário, ã farnília da vítima.

2. Está obrigado o uso de corrente de aço de tipo A5 a ser atada à presa no pescoço com mais de uma volta e meia e no. direito, caso ela seja animal perigoso não decapitado.

3. Toda presa cujo peso seja superior j 100 kg ou

4. Se o animal capturado morto é destinado ao curtume *g* declaração ao partir é obrigatória, independentemente do peso.

5. Esta fiscalização se reserva o direito de confiscar toda presa que não esteja conforme as disposições acima, caso esta o consinta.

6. Pelo não respeito de cada item deste regulamento se cobra uma multa de 8500, a ser depositada na conta corrente da administração municipal no Banco Regional, de número 6789 (a menos de acordo entre cavalheiros pois quem não tem cão caça com gato).

O regulamento completo encontra-se afixado no local da fiscalização, assim como na estação mais próxima da mesma, no quadro ao lado do banco de espera reservado aos agentes funerários.

Lembra-se ainda aos distintos caçadores que não se mata ser animal sem o respeito das regras elementares que são aplicadas por eles em caso semelhante.

muito obrigada em nome da fauna e FELIZ CAÇADA!

A PREFEITURA MUNICIPAL

No apêndice deste artigo é fornecida a lista de resoluções gramaticais corretas do texto acima. Nessa lista aparecem entre parênteses as demais categorias possíveis da palavra em questão. Salienta-se entretanto que o uso dessas categorias para a construção do texto ambíguo pode acelerar a obtenção das regras de precedência, embora, como vimos no parágrafo 5, não seja esse o procedimento usado pelo sistema SPIRIT.

10 - CONCLUSÃO

O simples exame do texto de aprendizado acima permite conjecturar que, como o francês, o português é posicionai, no sentido que um grande número de ambigüidades pode ser levantadas tendo em conta unicamente o contexto imediato.

Observa-se, de passagem, que essa conclusão não se aplica a qualquer língua. De fato, citando apenas alguns exemplos, constata-se que as propriedades posicionais são muito menos marcantes em russo* e em certas línguas asiáticas como o japonês. Aliás, no caso desta última língua, é interessante observar que quando se usa a escrita fonética, o número de ambigüidades se revela extremamente elevado. Isso obviamente torna um sistema como o SPIRIT menos eficiente.

Voltando ao português, pode-se concluir que, muito provavelmente, os métodos estatísticos e de filtragem aqui descritos serão aproximadamente os mesmos que para o francês, podendo-se com isso esperar obter resultados comparáveis. Aliás, o texto de aprendizado dado acima, permite urna iniciação do trabalho lingüístico que se faz mister para se construir para o português um sistema de indexação automática, baseado na metodologia aqui apresentada. Tal trabalho consistiria então na definição de categorias gramaticais, na construção do dicionário associado e na verificação da qualidade dos resultados *e* pode-se estimar que bem dirigido ocuparia cerca de dois lingüistas durante dois a três

Para finalizar, gostaríamos de observar que um sistema como o SPIRIT nos parece particularmente bem adaptado a corpus normativos como os jurídicos, legislativos, médico-terapêuticos, etc.

REFERÊNCIASBIBLIOGRÁFICAS

1 ANDREEWSKY, A.; DEBILI, F. & FLUHR, C.; Apprentissage-sifvtaxe semantique lexicale, **Revue du Palais de la Découverte,** 9(83): 17-40, 1980.

[2] ANDREEWSKY, A., COMBRISSON, F. & FLUHR, C., Le problème de l'identification automatique des concepts. Note du Service de documentation du **Centre d'Etudes Nucleaires** de Saclay, 8 CEA-N-1816,1975.

[3] ANDREEWSKY, A. & FLUHR, C. Apprentissage-Analyse automatique du langage, application à la documentation, **Dunod-documents de Linguistique quantitative,** n? 21, 1973.

ABSTRACT

This paper deals with automatic indexing based on linguistic and statistical methods, which aims to allow the processing of documents in natural language. The main lines of a system called SPIRIT, that uses such methods, and that was developed for the French Languages by a group of researchers of the CNRS, including the first author, is described. Some basic aspects of the applicability of those methods to the Portuguese Language are considered.

Língua materna do primeiro autor.

APÊNDICE

RESOLUÇÃO GRAMATICAL DO TEXTO DE APRENDIZADO

Significado de algumas siglas e símbolos não evidentes:

PRONVER : Pronome verbal
ELOC : Elemento de locução
* : Indica categoria a excluir se os acentos gráficos são representados.
ARTGEN : Artigo generalizado

A	: ELOCPREP, (PRONVER, ART, PREP, ELOCADJ, ELOCADV, PREPART*)
partir	ELOCPREP, (VERINF, VERCONJ,SUBST)
de	ELOCPREP, (PREP, ELOCADJ, VERCONJ*)
hoje	ADV
está	VERAUX, (VERCONJ, PRONDEM*, ARTDEM*)
aberta	VERPARTP, (ADJ)
a	ART, (PRONVER, PREP, ELOCPREP, ELOCADJ, ELOCADV, PREPAR*)
estação	: SUBST
de	: PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
caça	: SUBST, (VERCONJ\
na	:PREPART
reserva	: SUBST, (VERCONJ)
da	: PREPART, (VERCONJ*)
mata	: SUBST, (VERCONJ)
municipal	: ADJ
Um	: ART, (ADJNUM)
extrato	: SUBST, (ADJ)
do	: PREPART, (SUBST*)
regulamento	: SUBST
a	: ELOCADJ, (PRONVER, ART, PREP, ELOCPREP, ELOCADV, PREPART*)
respeito	ELOCADJ, (SUBST, VERCONJ)
é	VERAUX, (VERCONJ, CONJ*)
fornecido	VERPARTP, (ADJ)
abaixo	ADV
conforme	PREP, (ADJ)
lei	SUBST
federal	ADJ
NO	ELOCADJ
9876	ELOCADJ, (NÚMERO)
aprovada	: ADJ, (VERPARTP)
a	: PREP, (PRONVER, ART, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
5/7/81	: DATA
Antes	: ELOCCONJ, (ADV, ELOCPREP)
que	: ELOCCONJ, (CONJ, PROM, ELOCPRON, PRONREL, SUBST*)
entre	: VERCONJ, (PREP)
na	:PREPART
área	:SUBST
de	: PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
caça	:SUBST, (VERCONJ)
o	: ART, (PRON)
caçador	:SUBST
passa	: VERCONJ, (SUBST)
e	: CONJ, (VERAUX*, VERCONJ*)
pára	: VERCONJ (PREP)
obrigatoriamente	: ADV
no	: PREPART, (SUBST*)
posto	:SUBST, (VERPARTP)
de	: PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
fiscalização	:SUBST
onde	: PRONREL, (ADV)
deve	: VERAUX, (VERCONJ)
apresentar	•. VERINF, (VERCONJ
	: PREPART, (PRONVER*, ART*, PREP*, ELOCPREP* ELOCADJ*, ELOCADV*)
guarda	: SUBST, (VERCONJ)
sua	: PRONPOSS, (VERCONJ)
autorização	:SUBST
de	: PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
caça	: SUBST, (VERCONJ)
Esta	: PRONDEM, (ARTDEM, VERAUX*, VERCONJ*)
	: PRON, (ART, PREP, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
guarda	VERCONJ, (SUBST)
para	PREP, (VERCONJ)
devolve	VERINF, (VERCONJ)
la	VERINFPRON, (SUBST*)
ao	PREPART
caçador	SUBST, (ADJ)
se	CONJ, (PRONVER, SUBST' PRQNIND)
este	PRONDEM, (ARTDEM,

SUBST)
volta : VERCONJ, (SUBST)
são : ADJ, (VERAUX. VERCONJ)
e : CONJ, (VERAUX*, VERCONJ}
salvo : ADJ, (VERPARTP, PREP)
antes : ELOCPREP, (ELOCCONJ, ADV)
do : ELOCPREP, (SUBST*)
pôr : VERINF*, (PREP)
do : PREPART, (SUBST*)
sol : SUBST
ou : CONJ
caso : ELOCADV, (SUBST, VERCONJ, CONJ)
contrário : ELOCADV, (SUBST, ADJ)
à : PREPART, (PREP*, PRON*, ART*, ELOCPREP*, ELOCADJ*, ELOCADV*)
família SUBST
da PREPART, (VERCONJ*)
vítima SUBST, (VERCONJ*)
Está VERAUX, (VERCONJ, PRONDEM*, ARTDEM*)
obrigado VERPARTP, (ADJ)
o ART, (PRONVER)
uso SUBST, (VERCONJ)
de PREP, (ELOCPREP, E LOCADJ, VERCONJ*)
corrente SUBST, (ADJ)
de PREP, (ELOCPREP, E LOCADJ, VERCONJ*)
aço : SUBST
de : ELOCADJ, (PREP, ELOCPREP, VERCONJ*)
tipo : ELOCADJ, (SUBST)
A5 : ELOCADJ
a : PREP, (PRONT, ART, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
ser : VERAUX, (VERINF, SUBST)
atada : VERPARTP, (ADJ)
á : PREPART, (PRON*, ART*, PREP*, ELOCADJ*, ELOCADV*, ELOCPREP*)
presa : SUBST, (ADJ)
no : PREPART, (SUBST*)
pescoço : SUBST
com : PREP
mais : ADVCOMP, (ADV, ELOCCONJ, ELOCADV, SUBST)
de : PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
uma : ADJNUM, (ART)
volta : SUBST, (VERCONJ)

e CONJ, (VERAUX*, VERCONJ*)
meia ADJ, (SUBST)
e CONJ, (VERAUX*, VERCONJ*)
nó SUBST, PREPART*
direito ADJ, SUBST, ADV
caso CONJ, (SUBST, VERCONJ, ELOCADV)
ela PRON
seja VERCONJ, VERAUX
animal SUBST, ADJ
perigoso ADJ
não ADVNEG
decapitado ADJ, VERPARTP
Toda ARTGEN
presa SUBST, ADJ
cujo PRONREL
peso SUBST, (VERCONJ)
se/a *VERCONJ, (VERAUX)*
superior ADV, (SUBST, ADJ)
a PREP, (PRON, ART, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
100 ADJNUM
Kg : SIGLA
ou :CONJ
cobra : SUBST, (VERCONJ)
venenosa :ADJ
deve : VERAUX, (VERCONJ)
ser : VERAUX, VERINF, SUBST
declarada : VERPARTP, (ADJ)
à PREPART, (PREP*, PRON*, ART*, ELCOPREP*, ELOCADJ*, ELOCADV*)
guarda : SUBST, (VERCONJ)
de -. PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
caça : SUBST, (VERCONJ)
se CONJ, (PRONVER, SUBST*, PRONIND)
capturada : VERPARTP, (ADJ)
viva : ADJ, (VERCONJ, INTERJ}
salvo : PREP, (ADJ, VERPARTP)
ausência : SUBST
da : PREPART, (VERCONJ*)
mesma : PRONDEM, (ADJ)
Se : CONJ, (PRON, PRONREFL, SUBST*)
o : ART, (PRON)
pelo : SUBST, (PREPART)
do : PREPART, (SUBST)
animal : SUBST, (ADJ)
capturado : VERPARTP, (ADJ)
morto : ADJ, (SUBST, VERPARTP)
é : VERCONJ, (VERAUX,

CONJ*)
destinado : ADJ, (VERPARTP)
ao :CPREPART
curtume :SUBST
a : ART, (PRQN, PREP, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
declaração :SUBST
ao :CPREPART
partir : SUBST, (VERINF, VERCOIMJ, ELOCPREP)
: VERCONJ, CVERAUX, COW *)
obrigatória : ADJ
independenternente : ADV
do : CPREPART, (SUBST*)
peso : SUBST, (VERCONJ)
Esta : ARTDEM, (VERBAUX, VERBCOIMJ, PRONDEM)
fiscalização :SUBST
se : PRONREFL, (PRONVER, CONJ, SUBST*)
reserva : VERCONJ, (SUBST)
o : ART, (PRON)
direito : SUBST, (ADJ, ADV)
de : PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
confiscar : VERINF, (VERCONJ)
toda :ARTGEN
presa : SUBST, (ADJ, VERPARTP)
que :PRONREL,(PRON,CONJ, ELOCCONJ, ELOCPRON, SUBST*)
não : ADVNEG
esteja : VERCONJ, (VERAUX)
conforme : ADJ, (PREP)
às : PREPART, (PRON*, ART*, SUBST*)
disposições :SUBST
acima : ADJ, (ELOCADV)
caso : CONJ, (SUBST, ELOCADV, VERCONJ)
esta : PRONDEM, (ARTDEM, VERAUX:, VERCONJ)
: PRONVER, (ART)
consinta : VERCONJ
Pelo : PREPART, (SUBST)
não : ADJNEG, (ADVNEG)
respeito : SUBST, (VERCONJ, ELOCADJ)
de : PREP, (ELOCPREP, ELOCADJ, VERCONJ*)
cada :ARTGEN
item :SUBST
deste : PREPPRONDEM, (VERCONJ)
regulamento :SUBST
se : PRONIMP, (PRONVER, CONJ, SUBST*)
cobra : VERCONJ, (SUBST)
uma : ART, (ADJNUM)
multa : SUBST, (VERCONJ)
de : PREP, (ELOCPREP, ELOCADJ,VERCONJ*)
S : SIGLA
500 : ADJNUM
a : PREP, (PRON, ART. ELECPREP,ELOCADJ, ELOCADV. PREPART*)
ser : VERAUX, (VERINF, SUBST)
depositada : VERPARTP, (ADJ)
na : PREPART
conta : SUBST, VERCONJ
corrente : ADJ, SUBST
da : PREPART, VERCONJ*
administração : SUBST
municipal : ADJ
no : CPREPART, (SUBST*)
banco : SUBST, (VERCONJ)
regional : ADJ
de : ELOCADJ. (PREP, ELOCPREP, VERCONJ*)
número : ELOCADJ, (SUBST)
6789 : ELOCADJ, (NÚMERO)
a •. ELOCPREP, (PRON, ART, PREP, ELOCADJ, ELOCADV, PREPART*)
menos : ELOCPREP, (ADV, SUBST, ELOCCONJ)
de ; ELOCPREP, (PREP, ELOCADJ, VERCONJ*)
acordo : SUBST, (VERCONJ)
entre : PREP, (VERCONJ)
cavalheiros : SUBST
pois : CONJ
quem : PRON
não : ADVNEG
tem : VERCONJ, (VERAUX)
cão : SUBST
caça : VERCONJ, (SUBST)
com :PREP
gato : SUBST
O : ART, (PRON)
regulamento : SUBST
completo : ADJ
encontra : VERAUX, (VERCONJ)
se : PRONREFL. (PRON, CONJ)
afixado : VERPARTP, (ADJ)
no : PREPART
local : SUBST, (ADJ)
da : PREPART, (VERCONJ)
fiscalização : SUBST
assim : ELOCCONJ, (ADJ)

como :ELECCONJ, CADV, VERCONS)
na : PREPART
estação : SUBST
mais : ADV, (ELOCCONJ, E LOCADV, SUBST, ADVCOMP)
próxima : ADJ
da : PREPART
mesma : PRONDEM, (ADJ)
no .•PREPART, SUBST*
quadro :SUBST
ao : ELOCPREP, (PREPART)
lado :ELOCPREP, (SUBST)
do : ELOCPREP, (PREPART, SUBST*)
banco : SUBST, (VERCONJ)
de : PREP, (ELOCPREP, ELOCADJ, VERBCONJ*)
espera :SUBST, (VERCONJ)
reservado : ADJ, (VERPARTP, SUBST)
aos : PREPART
agentes : SUBST
"funerários :ADJ
Lembra :VERCONJ
se : PRONIND, (PRONVER, CONJ, SUBST*)
ainda : ADV, (ELOCCONJ)
aos : PREPART
distintos : ADJ, (SUBST)
caçadores : SUBST, (ADJ)
que : CONJ, (PRONREL, PRON, ELOCCONJ, ELOCPRON, SUBST*)
nato :ADVNEG
se : PRONIND, (PRONVER, CONJ, SUBST*)
mata : VERCONJ, (SUBST)
ser : SUBST, (VERAUX, VERCONJ)
animal ADJ, (SUBST)
sem PREP
o ART, (PRONVER)
respeito SUBST, (VERCONJ)
das PREPART
regras SUBST
elementares ADJ
que PRONREL, (CONJ, PRON, ELOCCONJ, E LOCPRON, SUBST*)
são : VERAUX, (VERCONJ, ADJ, SUBST)
aplicadas : VERPARTP, (ADJ)
por : PREP, (VERINF)
eles : PRON
em : PREP, (ELOCPREP)
caso : SUBST, (VERCONJ, CONJ, E LOCADV)
semelhante : ADJ, (SUBST)
Muito : ADV, (ADJ)
obrigado : INTERJ, (ADJ, VERPARTP)
em : ELOCPREP, (PREP)
nome : ELOCPREP, (SUBST)
da :ELOCPREP, (PREPART, VERCONJ*)
fauna :SUBST
e : CONJ, (VERAUX*, VERCONJ*)
feliz :ADJ
caçada : SUBST, (VERPARTP, ADJ)
A : ART, (PRONVER, PREP, ELOCPREP, ELOCADJ, ELOCADV, PREPART*)
Prefeitura :SUBST
Municipal :ADJ