OTHERO, Gabriel de Ávila. Algumas considerações sobre a importância da continuidade tópica na classificação automática de documentos digitais. *Revista Virtual de Estudos da Linguagem – ReVEL*. V. 2, n. 3, agosto de 2004. ISSN 1678-8931 [www.revel.inf.br].

ALGUMAS CONSIDERAÇÕES SOBRE A IMPORTÂNCIA DA CONTINUIDADE TÓPICA NA CLASSIFICAÇÃO AUTOMÁTICA DE DOCUMENTOS DIGITAIS

Gabriel de Ávila Othero[1]

gabriel_othero@terra.com.br

Introdução

Hoje em dia, empresas, universidades e centros de pesquisa estão utilizando cada vez mais documentos digitais – tais como livros virtuais, artigos on-line, bancos virtuais de teses etc. Além de resolver o problema do espaço físico (criando, em contrapartida, o problema do espaço virtual), o uso de textos digitais possibilita o intercâmbio praticamente instantâneo de documentos entre instituições de qualquer ponto do planeta. Cada vez mais estão aparecendo bibliotecas inteiras na rede, contendo exclusivamente livros virtuais (ou e-books)[2], além de *sites* de jornais e revistas que disponibilizam todos seus textos on-line[3].

No entanto, à medida que mais documentos digitais vão surgindo na rede e entrando no dia-a-dia de empresas e instituições acadêmicas, surge um grande desafio: conseguir classificar e organizar esses documentos digitais de maneira eficiente. Utilizar para isso uma classificação manual pode ser bastante dispendioso, lento e oneroso para as

[1] Aluno do Programa de Pós-graduação em Letras, área de Lingüística Aplicada, pela Pontifícia Universidade Católica do Rio Grande do Sul; bolsista de pesquisa do CNPq.
[2] Veja Virtualbooks (www.virtualbooks.com.br) e Project Gutenberg (www.gutenberg.net), por exemplo.
[3] Veja o *site* dos jornais Folha de S. Paulo (www1.folha.uol.com.br/fsp/) e Zero Hora (www.zh.com.br), por exemplo.

instituições. Uma tarefa e – um desafio – que se apresenta para os pesquisadores da área de Processamento de Linguagem Natural (PLN) e da Lingüística Computacional é a classificação automática de documentos digitais. Afinal, uma vez estando no computador, por que não utilizar *softwares* que façam o penoso trabalho de ler toda uma obra antes de classificá-la de acordo com o tema ou assunto? Ou seja, já que temos livros e documentos virtuais, por que não termos também classificadores virtuais e programas que possam trabalhar como verdadeiros "bibliotecários digitais"?

Os classificadores automáticos ou semi-automáticos parecem ser a chave para solucionar o problema de se lidar com enormes quantidades de textos e documentos digitais. Além de serem ágeis e rápidos, programas desse tipo em geral não representam grandes custos para instituições, empresas, ou mesmo para usuários particulares. No entanto, a classificação automática ainda está em seus estágios iniciais, uma vez que ela tenha surgido para solucionar um problema relativamente recente, da era digital.

Neste trabalho, discutiremos a importância que estudos em sintaxe funcional sobre manutenção de tópico no discurso podem ter para aplicações em PLN, especificamente com relação aos classificadores automáticos de textos e documentos digitais. Na seção 1, analisaremos os métodos empregados nos trabalhos de classificação automática de textos digitais apresentados em Langie (2003) e Langie & Lima (2003), tentando mostrar que, se fossem desenvolvidos alguns recursos que levassem em conta teorias funcionais sobre a manutenção de tópicos discursivos no processo de classificação automática, a precisão dos classificadores automáticos de texto poderia ser ainda mais acurada.

Na seção 2, veremos como se dá a manutenção tópica no discurso, baseando-nos nas de idéias de Givón (1979, 1992) e Ariel (1988), principalmente, e iremos propor que essas teorias sejam levadas em consideração no desenvolvimento de programas que classifiquem textos digitais quanto ao tema ou assunto. Mostraremos que palavras importantes na manutenção tópica (em especial os pronomes pessoais do caso reto) são subestimados ou mesmo ignorados por programadores no desenvolvimento de ferramentas de classificação automática de textos.

1. Sobre a Classificação Automática de Textos

A Classificação Automática de Textos (CT), como vimos, poderá ajudar instituições que lidam com grandes quantidades de documentos digitais. Programas automáticos ou semi-automáticos de classificação de documentos poderão auxiliar bibliotecas virtuais e convencionais, além de refinar grandes sites de busca como o Yahoo![4] e o Google[5].

De acordo com Sebastiani (2002)[6], a CT pode ser definida da seguinte forma:

> A categorização de textos é a tarefa de atribuir um valor booleano {T, F} para cada par $(d_j, c_i) \in D \times C$, onde $D = \{d_1, ..., d_{|D|}\}$ é um conjunto de documentos e $C = \{c_1, ..., c_{|C|}\}$ é um conjunto pré-definido de categorias.

Em outras palavras, a classificação automática consiste em classificar um documento digital em uma determinada categoria, de acordo com critérios estabelecidos *a priori* por programadores humanos. Essa classificação geralmente respeita uma estrutura hierárquica, utilizando "uma árvore de categorias, permitindo que os documentos sejam classificados tanto nas folhas quanto nos nodos intermediários" (Langie & Lima, 2003: 3).

Langie & Lima (2003: 2) explicam que,

> com a utilização de uma estrutura hierárquica de categorias, o processo de classificação pode ser decomposto em subprocessos menores, nos quais a quantidade de variáveis envolvidas é reduzida. Conforme Koller e Sahami, em [Koller,1997][7], categorias que se encontram próximas, dentro da estrutura hierárquica, possuem, em geral, mais características em comum do que outras categorias.

Dessa forma, uma palavra como *lingüista* pode não ser muito esclarecedora para ajudar a classificar um texto entre as categorias *Sintaxe*, *Semântica* ou *Fonologia*, mas poderá ser um bom atributo para diferenciar textos entre categorias maiores, como *História*, *Medicina* ou *Lingüística Aplicada*.

Veja um exemplo de uma árvore de categorias utilizada por Langie & Lima (2003: 5):

---

[4] www.yahoo.com
[5] www.google.com
[6] SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1, apud Langie (2003: 6).
[7] KOLLER, D.; SAHAMI, M. (1997). Hierarchically classifying documents using very few words. *Proceedings of ICML-97, 14th International Conference on Machine Learning*.

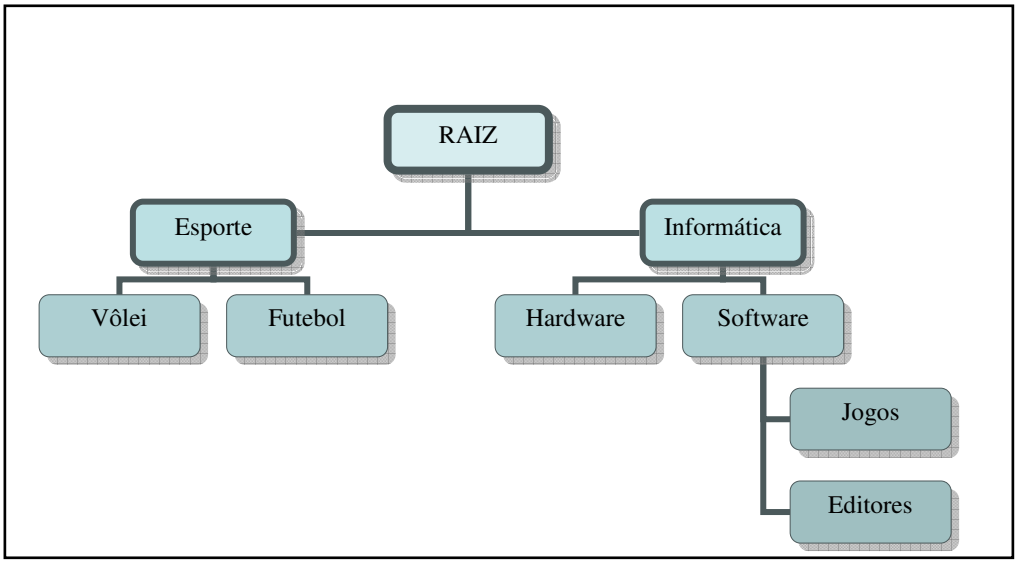

Figura 1: Exemplo de árvore de categorias

Basicamente, a tarefa de um classificador automático é "ler" os textos e classificá-los em categorias ou subcategorias pré-estabelecidas pelos programadores humanos. Durante a "leitura" do texto, o programa irá comparar uma série de palavras que aparecem no texto-alvo com as palavras que ele já conhece. A partir daí, ele irá tentar classificar este texto-alvo dentro de uma categoria, baseando-se no número de ocorrência das palavras-chave que apareceram no documento. Dentre as palavras que o classificador conhece, há palavras com mais valor semântico e outras com menor valor semântico. Isso quer dizer que algumas palavras recebem valores mais altos do que outras na escala de importância na classificação do texto: "o peso na composição dos vetores é a combinação de fatores conhecida como TFC [Salton e Buckley, 1988][8]. Nessa combinação, termos mais importantes recebem peso próximo a 1 e termos menos importantes recebem valores próximos a zero" (Langie & Lima, 2003: 8).

Ainda de acordo com Langie & Lima (2003: 8), estas são as principais etapas por que passam os documentos digitais até serem classificados devidamente:

> Antes de terem suas representações geradas, os documentos passam por uma etapa de *pré-processamento*. Nesta etapa os termos dos documentos são

[8] SALTON, G.; BUCKLEY, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing e Management*, Vol. 24 n. 5.

> colocados em minúsculas, e caracteres de pontuação e dígitos são eliminados. Além disso, aplica-se uma seleção de atributos com a remoção de *stopwords* (artigos, advérbios, conjunções, numerais, preposições, pronomes, e verbos de ligação). A lista de *stopwords* contém 365 termos. Atributos também são selecionados com a remoção de termos cuja freqüência do documento (FD) é inferior a 3. A FD de um termo t corresponde ao número de documentos nos quais t ocorre pelo menos uma vez. Conforme [Yang e Pederson, 1997][9] **(sic)** o princípio da seleção de atributos usando FD é que termos raros não são informativos para predizer a categoria de um documento e também não influenciam no desempenho do classificador.

Não iremos nos concentrar no funcionamento dos classificadores automáticos de textos em si (para isso, remetemos o leitor para Langie, 2003, Langie & Lima, 2003, e Garside, Leech & McEnery, 1997). O que nos chama atenção aqui são basicamente dois fatores sobre o procedimento desses classificadores: (a) a lista das chamadas *stopwords*; e (b) o critério utilizado pelos programadores para classificar uma palavra em "relevante" ou "não-relevante" na determinação do assunto do texto.

Na lista de *stopwords*, encontram-se todas aquelas palavras que aparentemente não desempenham um papel importante na classificação automática dos textos de acordo com o tema ou assunto. São palavras que não deveriam apresentar carga semântica relevante, tendo, em geral, um papel meramente funcional no texto. Nessa lista, entram palavras como artigos, advérbios, conjunções, numerais, preposições, *pronomes*, e verbos de ligação. No entanto, como veremos na próxima seção, acreditamos firmemente que os pronomes são o principal mecanismo de manutenção tópica em um texto, tendo, por isso, um valor fundamental para qualquer classificador que se baseie no número de ocorrência de determinado objeto-de-discurso, ou entidade, no texto para proceder com a sua classificação.

Outro fator interessante, que também analisaremos na próxima seção, diz respeito ao critério de classificação das palavras em *relevantes* ou *não-relevantes*. Essa questão está intrinsecamente relacionada à anterior, já que, como veremos, um tópico discursivo é raramente repetido: normalmente ele é retomado por meio de anáforas lexicais, ou, mais freqüentemente, por meio de anáforas pronominais ou elípticas.

[9] YANG, Y.; PEDERSEN, J. (1997). A Comparative Study on Feature Selection on Text Categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*.

2. Sobre a Continuidade Tópica

Dentro do paradigma funcionalista, a sintaxe é percebida como uma estratégia de organização frasal para um determinado fim comunicativo. Nós, falantes, organizamos a informação do texto de frase em frase, por meio de expressões nominais, que podem servir de tópicos, e suas retomadas anafóricas, que funcionam normalmente para manter a continuidade tópica. Nessa organização textual, as expressões nominais nos ajudam a armazenar, arquivar e organizar a informação nova que vai "entrando", vinda do texto, enquanto as retomadas anafóricas (principalmente através de elipse e pronomes) servirão para manter o arquivo do tópico discursivo aberto na memória de trabalho.

Ainda que haja algumas controvérsias sobre o que é ser o tópico, ou sobre se a topicalidade é escalar ou não (cf. Ariel, 1988, Givón, 1979, 1992, e Pontes, 1986), definiremos o tópico como *aquilo sobre o que se vai falar*. A essa sucinta definição, porém, devemos acrescentar uma outra, baseada nas idéias de Givón: *um tópico é caracterizado como tópico se for retomado e designado como tal em um número sucessivo de orações*.

De acordo com Barbisan & Machado (2000: 72), "os tópicos são etiquetas de arquivo para a armazenagem na memória episódica. As etiquetas são tornadas assim perceptualmente salientes pelo emissor, que nelas embasa a informação para o receptor." Enquanto a expressão nominal "abre esses novos arquivos" em nossa memória, introduzindo novos tópicos, normalmente os pronomes servem para retomar e manter a cadeia tópica, garantindo assim a *continuidade e manutenção do tópico* no discurso.

Confira o gráfico visto em Othero (2004: 24), sobre o processo de referenciação e continuidade tópica:

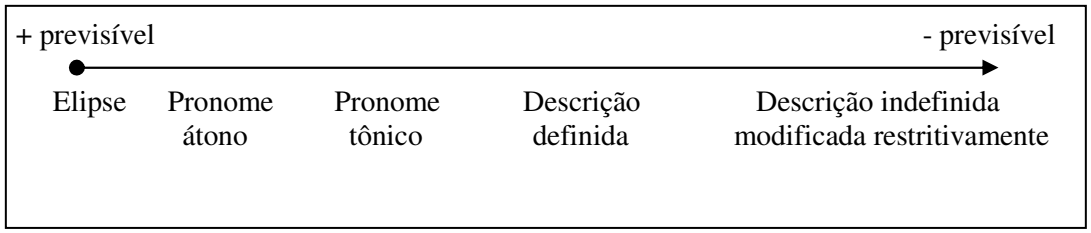

Gráfico 1: Escalas quantitativas de identificação referencial (recursos contextuais para designar o referente)

No processo de continuidade tópica, utilizamos diversas estratégias de retomada do tópico. Ariel (1988) mostra que, apesar dessa diversidade de estratégias de manutenção tópica, a distribuição dos marcadores de topicalidade nas retomadas anafóricas não é exatamente flexível. Além disso, ela argumenta que a retomada anafórica pronominal é amplamente usada, especialmente quando o tópico tem acessibilidade máxima (ou grande) e distância mínima (ou muito curta). Diversos estudos (Ariel, 1988, Barbisan & Machado, 2000, Givón, 1992, entre outros) têm mostrado que a retomada anafórica pronominal é definitivamente a estratégia mais utilizada na manutenção tópica no discurso. Isso remete-nos imediatamente às duas questões abordadas na seção 1:

a) A lista de *stopwords* contém pronomes, não os levando em consideração durante o processo de classificação do texto.

b) Uma palavra é considerada relevante se ela não é rara e não se encontra na lista de *stopwords*. Porém, se um tópico (que é relevante ao texto) for retomado exclusivamente por pronomes, ele será considerado relevante? A resposta é um incrível "NÃO" para qualquer ferramenta de classificação automática de textos que não leve em conta as teorias funcionais sobre a manutenção tópica discursiva.

A questão que queremos levantar aqui é a seguinte: como um classificador automático de documentos digitais pretende classificar textos baseando-se na freqüência das palavras mais importantes e/ou relevantes que encontra no texto se, ao mesmo tempo: (a) as expressões consideradas mais importantes – os tópicos – são normalmente retomadas por anáforas pronominais[10] (pelo menos a curta e média distância); e (b) a ferramenta, além

[10] Não mencionaremos as retomadas por anáfora zero ou elipse (que também são muito freqüentes), porque isso iria complicar e comprometer ainda mais o processamento textual dos classificadores automáticos que sequer cogitam a hipótese de trabalhar com esse tipo de retomada que está, de certa forma, "invisível" na superfície do texto.

de não reconhecer a relação anafórica pronominal, deixa todos os pronomes em uma lista de palavras raras e sem relevância, chamada de *stopwords*.

3. Possíveis Caminhos

Foge à proposta específica deste trabalho testar as hipóteses aqui discutidas em um trabalho prático. Para isso, necessitaríamos de mais tempo e espaço: teríamos de preparar ao menos três corpora (um de teste, um de controle e um de treinamento); explicitar os procedimentos da anotação manual (feitas por anotadores humanos) e da anotação automática (feita por classificadores automáticos); além de confrontar os resultados para averiguar a validade de nossa hipótese.

No entanto, acreditamos que esse seja o caminho necessário que as ferramentas de classificação automática de documentos digitais deverão tomar. Afinal, como sabemos, um texto não é composto por expressões nominais e sua exaustiva repetição. Ele antes é elaborado de outra maneira: o tópico é normalmente introduzido na forma de expressão nominal (normalmente um sintagma nominal indefinido) e retomado por diferentes tipos de anáfora, desde a elipse, passando pela anáfora pronominal (casos mais freqüentes), até a repetição da própria expressão nominal (muitas vezes, através de uma expressão definida).

Apesar de parecer ser o caminho ideal, nossas considerações levam a um outro grande problema: o do processamento anafórico pronominal[11]. Esse é com certeza um verdadeiro desafio às ferramentas automáticas que trabalham com textos em linguagem natural. Porém, acreditamos firmemente que, com esforços conjuntos entre lingüistas e programadores, tais dificuldades poderão no futuro ser solucionadas. Desse trabalho, o resultado imediato será provavelmente a melhoria significativa dos classificadores automáticos de documentos digitais.

[11] Cf. De Rocha (1996), Fligelstone (1992) e Garside, Leech & McEnery (1997), especialmente o capítulo 5 – Discourse annotation: anaphoric relations in corpora.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ARIEL, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24.
2. BARBISAN, L. B.; MACHADO, R. F. (2000). O tópico no texto argumentativo. *Letras de Hoje*, v. 35, n. 3.
3. COULSON, Mark. Anaphoric reference. (1996). In: GREENE, Judith; COULSON, Mark. *Language understanding: current issues*. Buckingham: Open University Press.
4. De ROCHA, M. (1996). A corpus-based study of anaphora in English and Portuguese. In: BORTLEY, S. P. & McENERY A. M. (eds.). *Corpus-based and computational approaches to discourse anaphora*, London: UCL Press.
5. FLIGELSTONE, S. (1992). Developing a scheme for annotating text to show anaphoric relations. In: LEITNER, G. (ed). *New directions in English language corpora: methodology, results, software*. Berlin: Mouton de Gruyter.
6. GARSIDE, Roger; LEECH, Geoffrey; McENERY, Anthony. (1997*). Corpus annotation: linguistic information from computer text corpora*. London / New York: Longman.
7. GIVÓN, T. (1979). From discourse to syntax: grammar as a processing strategy. In: GIVÓN, T. (ed.) *Discourse and syntax*. New York: Academic Press.
8. GIVÓN, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30.
9. HALLIDAY, M. A. K.; HASAN, R. (1976). *Cohesion in English*. London: Longman.
10. LANGIE, L. C. (2003). *Um estudo sobre a aplicação do algoritmo* k*NN à categorização hierárquica de textos*. Dissertação de Mestrado. Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS.
11. LANGIE, L. C.; LIMA, V. L. S. (2003). Classificação hierárquica de documentos textuais digitais usando o algoritmo *k*NN. *I Workshop em Tecnologia da Informação e Linguagem Humana*. São Carlos-SP, Brasil, 12 de outubro.
12. OTHERO, Gabriel de Ávila. (2004). *A anáfora e a tessitura do texto: um estudo do uso anafórico das descrições definidas*. Pará de Minas: Virtualbooks.

13. PONTES, Eunice. (1986). A noção de tópico. In: PONTES, Eunice. *Sujeito: da sintaxe ao discurso*. São Paulo: Ática; Brasília: INL, Fundação Nacional Pró-Memória.
14. SILVA, Magda Teresinha. (2003). A codificação do tópico como SN DEF em textos narrativos e em textos argumentativos. *Revista Virtual de Estudos da Linguagem – ReVEL*. Vol. 1, n. 1. [www.revel.inf.br]

APÊNDICE

. Lista de *stopwords* utilizada por Langie (2003):

abaixo	antes	automaticamente	como
acima	ao	b	conciso
acola	aonde	basicamente	conforme
agora	aos	bastante	conosco
ai	aparentemente	bem	consequentement
ainda	apenas	bilhao	e
al	apesar	bilhoes	consigo
alem	apos	bonito	contigo
algo	apropriado	c	contra
alguem	aquela	cada	contudo
algum	aquelas	catorze	correntemente
alguma	aquele	cedo	cr
algumas	aqueles	cem	cuja
alguns	aqui	certamente	cujo
ali	aquilo	cinco	d
alias	as	cinquenta	da
ambos	assim	claramente	daquilo
andar	ate	cm	das
anteontem	atraves	com	de
anteriormente	atualmente	comigo	debaixo

definitivamente
defronte
dela
dele
demais
dentro
depois
desde
dessa
dessas
desse
desses
desta
destas
deste
destes
dez
dezeseis
diante
dificilmente
diretamente
disponivel
disso
disto
do
dois
dos
doze
durante
duzentos
e
ela
elas
ele
eles
em
embora
enquanto
entao
entre
entretanto
especialmente
essa
essas
esse
esses
esta
estar
estas
este
estes
et
etc
eu
ex
exatamente
exceto
exclusivamente
f
facilmente
fazer
felizmente
ficar
finalmente
fora
frequentemente
g
geralmente
h
ha
haver
i
inicialmente
inteiramente
isso
isto
j
ja
jamais
jr
junto
k
km
l
la
las
lo
logo
longe
los
m
maioria
mais
mal
mas
me
mediante
melhor
menor
menos
mesmo
meu
meus
mil
milhao
milhares
milhoes
mim
minha
minhas
ml
muita
muitas
muito
muitos
n
na
nada
nao
naquilo
necessario
nela
nele
nem

nenhum
nenhuma
nessa
nessas
nesse
nesses
nesta
nestas
neste
nestes
ninguem
nisso
nisto
no
nono
normalmente
nos
nossa
nossas
nosso
nossos
novamente
nove
novecentos
noventa
nunca
o
oitenta
oito
oitocentos
onde
ontem
onze
opcionalmente
os
ou
outra
outro
p
para
parecer
pela
pelas
pelo
pelos
perante
permanecer
pior
pois
por
porem
porquanto
porque
portanto
possivelmente
posteriormente
pouca
poucas
pouco
poucos
praticamente
primeiramente
primeiro
principalmente
prontamente
provavelmente
q
quais
qual
qualquer
quando
quanta
quantas
quanto
quantos
quarenta
quase
quatorze
quatro
que
quem
quinto
quinze
r
rapidamente
realmente
s
se
segundo
seis
seiscentos
seja
sem
sempre
senao
ser
sessenta
sete
setecentos
setenta
seu
seus
si
sim
so
sob
sobre
sp
sua
suas
subito
t
tais
tal
talvez
tambem
tanta
tantas
tanto
tantos
tao
te
tel
ter

terceiro
teu
ti
toda
todas
todavia
todo
todos
tras
tres
treze
trezentos
trinta
tu
tua
tudo
u
ultimamente
um
uma
umas
uns
v
varias
varios
vem
vez
vezes
vinte
voce
w
x
z
zero
ª
º
ÿ