Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

GistSumm – *GIST SUMMarizer*: Extensões e Novas Funcionalidades

Thiago Alexandre Salgueiro Pardo

NILC-TR-05-05

Fevereiro 2005

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Apresenta-se, neste relatório, a descrição das extensões e das novas funcionalidades do GistSumm (*GIST SUMMarizer*), um sumarizador automático de textos. Em sua versão inicial, o GistSumm era um sumarizador extrativo, intersentencial, monodocumento, para produção de sumários genéricos. As extensões e novas funcionalidades do sistema referem-se à sua adaptação para (a) realização de sumarização multidocumento, (b) produção de sumários focados nos interesses da audiência a que se destina e (c) realização de sumarização intra-sentencial.

Este trabalho conta com o apoio da FAPESP, da CAPES e da Comissão Fulbright.

ÍNDICE

1. INTRODUÇÃO **2**

2. A SUMARIZAÇÃO NO GISTSUMM **2**

2.1. SEGMENTAÇÃO SENTENCIAL **2**
2.2. RANQUEAMENTO DE SENTENÇAS **3**
2.3. SELEÇÃO DE SENTENÇAS **3**

3. GISTSUMM: EXTENSÕES E NOVAS FUNCIONALIDADES **4**

4. CONSIDERAÇÕES FINAIS **5**

REFERÊNCIAS **6**

1. Introdução

Apresenta-se, neste relatório, a descrição das extensões e das novas funcionalidades do GistSumm – *GIST SUMMarizer* (Pardo, 2002; Pardo et al., 2003). O GistSumm é um sumarizador automático de textos que se baseia na idéia principal destes para a produção dos sumários. Em sua versão inicial, o GistSumm possui as seguintes características:

- realização de sumarização monodocumento, ou seja, para produção do sumário, um único texto-fonte é dado como entrada para o sistema;
- realização de sumarização extrativa, isto é, produção do sumário (ou extrato) de um texto-fonte pela seleção de sentenças inteiras do texto e posterior junção delas;
- pela razão anterior, é de natureza intersentencial, ou seja, não se realiza sumarização no interior das sentenças;
- produção de sumários genéricos, que são sumários voltados para uma audiência qualquer, sem interesses especificados.

As extensões e novas funcionalidades do sistema referem-se à sua adaptação para:

- realização de sumarização multidocumento, em que vários textos-fonte são fornecidos ao sistema para produção do sumário;
- produção de sumários focados nos interesses da audiência, isto é, sumários que tentam, de alguma forma, responder perguntas ou apresentar fatos relevantes sobre um tópico especificado pelo usuário;
- realização de sumarização intra-sentencial, isto é, sumarização no interior das sentenças.

Na próxima seção, o processo de sumarização do GistSumm é brevemente revisto. Na Seção 3, as extensões e novas funcionalidades do sistema delineadas acima são descritas. Alguns comentários finais são apresentados na Seção 4.

Para uma introdução à sumarização automática e revisão das técnicas mais relevantes da literatura, sugere-se a leitura de Martins et al. (2001), Rino e Pardo (2003), Mani e Maybury (1999) e Mani (2001).

2. A sumarização no GistSumm

O GistSumm tenta simular a forma como a sumarização humana acontece. Inicialmente, procura-se pela idéia principal do texto-fonte para, então, complementá-la com informações adicionais relevantes.

O processo de sumarização no GistSumm em sua versão inicial consiste, basicamente, em três etapas: segmentação sentencial, ranqueamento e seleção de sentenças. Cada uma destas etapas é introduzida nas subseções seguintes.

2.1. Segmentação sentencial

Inicialmente, as sentenças do texto-fonte são identificadas por meio de regras simples baseadas na ocorrência de sinais de pontuação, como o ponto e os sinais de interrogação e de exclamação. Verifica-se, também, a presença de abreviaturas (por

meio de uma lista de abreviaturas) para diferenciar o ponto que segue as palavras desta classe do ponto delimitador de sentenças.

2.2. Ranqueamento de sentenças

Esta etapa consiste em atribuir uma pontuação às sentenças identificadas na etapa anterior e produzir um ranque destas sentenças. Essa etapa compreende vários passos, como delineados a seguir:

a) *case folding*: todas as letras das sentenças são transformadas em letras minúsculas;
b) *stemming*: por meio de um *stemmer*, as palavras do texto são substituídas pelas suas respectivas raízes (*stems*);
c) remoção de *stopwords*: as *stopwords* (que são palavras muito comuns e, portanto, irrelevantes para o processamento em questão) são removidas do texto;
d) pontuação das sentenças: a pontuação das sentenças pode ocorrer por um de três métodos estatísticos simples: método *keywords* (Black and Johnson, 1988), método *average keywords*, isto é, o método *keywords* com normalização em função do tamanho das sentenças (medido em número de palavras) e método TF-ISF (*Term Frequency – Inverse Sentence Frequency*) (Larocca Neto et al., 2000);
e) ranqueamento das sentenças em função da pontuação obtida no passo anterior: as sentenças são ranqueadas e a sentença de maior pontuação é eleita como *gist sentence*, isto é, a sentença que melhor representa a idéia principal do texto.

Os passos (a), (b) e (c) são para fins de uniformização dos dados e para produção de melhores resultados. Em relação ao passo (b), utiliza-se o *stemmer* de Porter (1980); para o português, em particular, utiliza-se uma adaptação desse *stemmer* (Caldas Jr. et al., 2001). Sobre o passo (c), utiliza-se uma lista de *stopwords* (isto é, uma *stoplist*) para se identificar essas palavras.

2.3. Seleção de sentenças

Nesta etapa, são selecionadas as sentenças que formarão o sumário. Selecionam-se as sentenças que (a) contenham pelo menos um *stem* em comum com a *gist sentence* selecionada na etapa anterior e (b) tenham uma pontuação maior do que um *threshold*, que é a média das pontuações das sentenças. Por (a), procura-se selecionar sentenças que complementem a idéia principal do texto; por (b), procura-se selecionar somente sentenças relevantes.

O número de sentenças selecionadas para formar o sumário, por sua vez, depende da taxa de compressão especificada pelo usuário do sistema. A taxa de compressão é uma medida que determina o tamanho do sumário em relação ao tamanho do texto-fonte.

Para mais detalhes sobre o GistSumm e seu processo de sumarização, sugere-se a leitura de Pardo (2002) e Pardo et al. (2003). A próxima seção descreve a nova versão do sistema.

3. GistSumm: extensões e novas funcionalidades

Em sua versão inicial, o GistSumm foi implementado em Borland Delphi. A nova versão do sistema foi desenvolvida em C e é, portanto, portável para diferentes ambientes computacionais.

Em linha de comando, a execução do sistema deve observar a seguinte sintaxe:

GistSumm.exe [-is] [-ak] [-nM] texto1 [... textoM] [consulta] taxa_compressão

em que:

- a opção -is determina que se deve realizar sumarização intra-sentencial;
- a opção -ak determina que o método de ranqueamento de sentenças a ser utilizado deve ser o método *average keywords*; como opção padrão, o método *keywords* é utilizado; o método TF-ISF não foi incorporado à essa nova versão do sistema devido ao seu desempenho ruim (vide Pardo, 2002);
- a opção -nM especifica que se deve realizar sumarização multidocumento considerando-se M textos, os quais devem ser listados depois dessa opção; como opção padrão, a sumarização monodocumento é realizada;
- texto1 [... textoM] indica o(s) arquivo(s) com o(s) texto(s) que deve(m) ser sumarizado(s); se a opção -nM for utilizada, M textos devem ser indicados em seqüência;
- a opção consulta é utilizada para geração de sumários focados nos interesses do usuário; neste caso, a consulta na qual o sumário deve ser focado deve ser especificada como uma seqüência de palavras entre aspas duplas;
- taxa_compressão especifica a taxa de compressão para a geração do sumário e deve ser um número entre 0 e 1; se for maior do que 1, considera-se que esse número indica o número máximo de palavras a serem incluídas no sumário; se for menor do que 0, somente a *gist sentence* é incluída no sumário.

Alguns exemplos de execução do sistema são mostrados abaixo:

- para sumarizar o texto contido no arquivo meutexto.txt com uma taxa de compressão de 80%

  GistSumm.exe meutexto.txt 0.80

- fazendo o mesmo, mas habilitando a sumarização intra-sentencial

  GistSumm.exe -is meutexto.txt 0.80

- usando, agora, o método *average keywords* para ranqueamento de sentenças

  GistSumm.exe -is -ak meutexto.txt 0.80

- realizando sumarização multidocumento de dois textos, com a especificação de uma consulta pelo usuário

  GistSumm.exe -n2 meutexto1.txt meutexto2.txt “meu assunto preferido” 0.80

A sumarização intra-sentencial, quando especificada, é realizada em todas as sentenças pela exclusão das *stopwords*. Apesar das sentenças resultantes terem a legibilidade prejudicada, o tamanho das sentenças é significativamente reduzido.

Para a realização da sumarização multidocumento, todos os textos apresentados ao sistema são justapostos, como se fossem um único texto, e o processo tradicional de sumarização do GistSumm é realizado. Note que, no estágio atual de desenvolvimento, questões complexas da sumarização multidocumento não são tratadas, como o reconhecimento e eliminação de informações redundantes provenientes de diferentes textos e a ordenação temporal dos eventos relatados nos textos.

Para a produção de sumários focados nos interesses do usuário, procura-se pela *gist sentence* que mais se assemelhe ao interesse/consulta do usuário (em vez da sentença com maior pontuação). A busca desta *gist sentence* se dá pelo cálculo da medida do cosseno (Salton, 1989) entre as sentenças do texto-fonte (ou textos-fonte, no caso de sumarização multidocumento) e a consulta especificada. Com base nessa medida, a sentença mais próxima da consulta especificada é determinada e escolhida como *gist sentence*. O processo de seleção de sentenças é executado como relatado na seção anterior.

O Quadro 1 sintetiza as principais características das duas versões do GistSumm. Na próxima seção, apresentam-se algumas considerações finais.

Quadro 1 – Principais características das duas versões do GistSumm

Características	**Versão inicial**	**Nova versão**
Linguagem de implementação	Borland Delphi	C
Ambiente computacional	MSWindows	MSWindows, UNIX/Linux
Sumarização mondocumento	Sim	Sim
Sumarização multidocumento	Não	Sim
Sumarização intersentencial	Sim	Sim
Sumarização intra-sentencial	Não	Sim
Geração de sumários genéricos	Sim	Sim
Geração de sumários focados nos interesses do usuário	Não	Sim
Métodos de ranqueamento de sentenças	*Keywords*, *Average Keywords* e TF-ISF	*Keywords* e *Average Keywords*

4. Considerações finais

Neste relatório, foram descritas as extensões e novas funcionalidade do GistSumm. Atualmente, encontram-se disponíveis para *download*[1] versões do sistema para a língua portuguesa e inglesa.

Futuramente, uma avaliação abrangente do sistema deverá ser realizada, verificando-se o desempenho do método de sumarização do sistema e a utilidade do sistema em diversas tarefas, como recuperação de informação e perguntas e respostas.

[1] http://www.nilc.icmc.usp.br/~thiago/GistSumm.html

Referências

Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, N. 3. Department of Computation. University of Manchester Institute of Science and Technology.

Caldas, Jr., J.; Imamura, C.Y.M.; Rezende, S.O. (2001). Evaluation of a stemming algorithm for the Portuguese language (in Portuguese). In the *Proceedings of the 2nd Congress of Logic Applied to Technology*, Vol. 2, pp. 267-274.

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, Atibaia-SP.

Mani, I. and Maybury, M.T. (1999). *Advances in Automatic Text Summarization*. MIT Press.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Relatório Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos.

Pardo, T.A.S. (2002). *GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos*. Série de Relatórios do NILC. NILC-TR-02-13.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, Vol. 14, N. 3.

Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA)*, pp. 203-245. Campinas-SP.

Salton, G. (1989). *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.