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

AVALIAÇÃO DE ALGORITMOS DE SUMARIZAÇÃO EXTRATIVA DE TEXTOS EM PORTUGUÊS

Carolina F. Reis de Souza

Maria das Graças Volpe Nunes

NILC-TR-01-9

Outubro, 2001

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

ALGORITMOS DE SUMARIZAÇÃO EXTRATIVA DE TEXTOS EM PORTUGUÊS[*]

Carolina Fátima Reis de Souza
carol@grad.icmc.sc.usp.br
Maria das Graças Volpe Nunes
mdgvnune@icmc.sc.usp.br
NILC - Núcleo Interinstitucional de Lingüística Computacional
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - São Carlos

RESUMO

A sumarização automática utiliza técnicas para gerar sumários automaticamente a partir de um dado texto. Um sumário é um resumo que tem o objetivo de passar a idéia principal de um determinado texto em poucas linhas. É evidente a grande utilidade da sumarização automática em várias áreas. Temos exemplos de sumários em textos científicos, artigos de jornais e revistas, pesquisas de dados na Internet, etc. Neste trabalho é apresentado um ambiente para testes de estratégias de sumarização automática extrativa de português, o SUMEX, seu funcionamento, as técnicas utilizadas e os resultados obtidos.

ABSTRACT

Automatic summarization is relevant in many computer applications such as summary generation of papers, news and documents. Information retrieval is also an area where this application is getting more useful. This paper presents the SUMEX, an environment of tests on extractive automatic summarization of Portuguese texts.

1 INTRODUÇÃO

A sumarização automática utiliza técnicas para gerar sumários automaticamente a partir de um dado texto. Um sumário é um resumo que tem o objetivo de passar a idéia principal de um determinado texto em poucas linhas. Ao contrário da língua inglesa, ainda são poucos os estudos de sumarização automática para a língua portuguesa.

É evidente a grande utilidade da sumarização automática em várias áreas. Temos exemplos de sumários em textos científicos, artigos de jornal e revistas, pesquisas de dados na Internet, etc.

Existem duas abordagens principais do tema: a fundamental e a empírica. Na abordagem fundamental, a estruturação profunda do discurso e a interferência de questões estilísticas do português são consideradas (Rino, 2001). Na empírica, é utilizado o método

[*] Trabalho de IC com apoio do PIBIC/CNPq.

de sumarização automática por extração de sentenças a partir do texto-fonte a ser sumarizado (Martins et al., 2001).

Neste trabalho, enfocamos a abordagem empírica, usando técnicas de sumarização automática extrativa. As técnicas utilizadas consideram a seleção de sentenças relevantes a partir da existência de palavras-chave geradas automaticamente e/ou palavras do título, de sua localização, de palavras sinalizadoras e palavras-chaves fornecidas pelo autor do documento, quando disponíveis. Isoladamente ou combinadas entre si, elas foram avaliadas num ambiente próprio para isso, o SUMEX, descrito na Seção 3.

Na próxima seção será introduzida a sumarização automática extrativa. Serão detalhadas também as técnicas utilizadas na seleção de sentenças e serão apresentados os programas necessários para a utilização do SUMEX. Na seção 3, será apresentada a avaliação das técnicas implementadas no ambiente SUMEX. A Seção 4 apresenta as conclusões e os comentários finais.

2 SUMARIZAÇÃO AUTOMÁTICA EXTRATIVA

A sumarização automática extrativa consiste da extração de sentenças relevantes do texto-fonte para a formação do sumário. Uma das principais vantagens é a simplicidade (baixo custo) de geração do resultado. Uma vez sabendo quais são as palavras-chave do texto, todas as sentenças que contiverem essas palavras, serão selecionadas para o sumário. Apesar da simplicidade deste método, não existiam trabalhos relativos ao português relatados na literatura.

As dificuldades surgem em garantir ao sumário gerado (a) uma boa textualidade, ou seja, se são coesos e coerentes, e (b) uma boa proximidade, ou seja, se ocorre a preservação da idéia principal. Um exemplo de problema em (a) seria a seleção de uma sentença contendo um pronome, sendo que aquilo a que ele se refere está em sentença anterior e não selecionada. Este é um problema de coesão textual típico na sumarização automática.

As palavras-chaves costumam ter um papel relevante para a seleção superficial de sentenças importantes do texto. Por isso o método de seleção de palavras-chave do texto tem grande importância no resultado final do sumário.

São utilizados no SUMEX dois métodos de palavras-chaves, o EPC-P e o EPC-R (Pereira, 2001). O primeiro, EPC-P (Extrator de Palavras-chaves baseado em padrões), encontra as palavras simples e compostas mais freqüentes que se encaixam em um dos padrões pesquisados de palavras-chaves. Os padrões de palavras-chaves são: Nome, Nome Preposição Nome, Nome Adjetivo, Nome Adjetivo Preposição Nome, Nome Preposição Nome Adjetivo. O segundo, EPC-R (Extrator de Palavras-chaves baseado em radicais), é fundamentado no *Extractor* (Turney, 1999), que é um programa que acha palavras-chaves de textos em inglês. Conforme a análise dos resultados dos dois métodos, concluiu-se que o EPC-P gera melhores palavras-chaves visando a sumarização automática e , portanto, este será utilizado para a avaliação do SUMEX.

Para a implementação das técnicas no SUMEX, foram necessários dois outros programas, o *Tokennizer* e o *Tagger* (Aluísio e Aires, 2000). O primeiro transforma o arquivo texto de entrada em outro contendo o mesmo texto com sua pontuação separada por espaços (conservam-se apenas os pontos de abreviações). O segundo utiliza o arquivo gerado pelo *Tokennizer* como entrada e gera outro arquivo de saída com o texto etiquetado, isto é, à frente de cada palavra é acrescentada a sua classe morfossintática.

O SUMEX também utiliza um programa de extração de radicais do português (*stemmer*), baseado no algoritmo de *Porter* (Porter, 1980). Este programa é utilizado quando as palavras-chave já foram encontradas. A partir, daí elas serão transformadas nos seus radicais, e estes serão utilizados para a seleção de sentenças para o sumário. Ou seja, os radicais das palavras-chave é que serão as palavras-chave propriamente ditas.

Por exemplo, se a palavra "educação" é selecionada como palavra-chave, o seu radical "educ" será utilizado pelo sumarizador como palavra-chave na extração de sentenças. Portanto, todas as sentenças que contiverem alguma palavra cujo radical também seja "educ", serão selecionadas para o sumário.

2.1 Estratégias de Seleção de Sentenças

As estratégias utilizadas no SUMEX para selecionar sentenças são descritas abaixo. Estas técnicas foram exploradas principalmente nas décadas de 70 e 80.

- Palavras-chave + Título

Nessa técnica, serão selecionadas todas as sentenças que contiverem alguma palavra-chave ou alguma palavra do título.

- Palavras-chave + Título + Localização

Além de selecionar as sentenças com alguma palavra-chave ou alguma palavra do título, serão adicionadas ao sumário a primeira e a última frase de cada parágrafo.

- Palavras-chave + Título + Sinalizadoras

Seleciona sentenças com palavras-chave, palavras do título e sentenças que possuem alguma das frases sinalizadoras abaixo:

objetivo
resultado
neste artigo
este artigo
neste texto
este texto
a conclusão
as conclusões
neste trabalho
este trabalho

Essas palavras foram escolhidas pois foi observado que normalmente elas estão presentes em frases com conteúdo explicativo, indicando algo importante sobre o texto.

- Palavras-chave do Autor

Nesta técnica, simplesmente serão utilizadas como palavras-chave as palavras que o usuário fornecer ao programa, podendo assim selecionar as palavras-chave do próprio autor do texto para analisar o sumário resultante.

2.2 Proposta de um Sumarizador Extrativo

Foi criado um ambiente, utilizando Visual C++ 6.0, para testes das estratégias de sumarização automática extrativa apresentadas anteriormente, chamado SUMEX. Sua interface é apresentada na Figura 1.

Antes de utilizar o SUMEX, deve-se passar o arquivo contendo o texto a ser sumarizado em dois programas. Primeiramente, utiliza-se o programa *Tokennizer*, que separa as pontuações das palavras. Isto deve ser feito para que o final das sentenças seja detectado. Este programa devolve um arquivo texto com extensão *new*.

Após isso, deve-se utilizar o programa *Tagger* tendo como entrada o arquivo *new*. Este programa etiqueta todas as palavras do texto. Por exemplo a frase "a construção de um projeto na rede pública de ensino" é transformada em "a_ART construção_N de_PREP um_ART projeto_N na_PREP+ART rede_N pública_ADJ de_PREP ensino_N" no arquivo texto de saída, que terá a extensão *tag*. Este arquivo etiquetado é necessário para o cálculo das palavras-chaves do texto, através do método EPC-P.

Em seguida, deve-se delimitar o texto a ser sumarizado no arquivo *new* colocando os caracteres */i* no início e no fim do texto. Se for desejável que as palavras do título façam parte das palavras-chave, deve-se selecionar no mesmo arquivo o título com os caracteres */t* no início e fim do título. Se não, deve-se somente inserir os caracteres delimitadores de título com um espaço entre eles (/t /t), no início do arquivo.

Um problema que ocorre no arquivo *new* é que ele perde as marcações de fim do parágrafo. Portanto deve-se rearranjar os parágrafos do texto dando Enter no fim dos parágrafos e tirar caracteres como □ que sobram entre as frases.

Iniciando o uso do SUMEX, o usuário deve entrar com o nome do arquivo de entrada (arquivo *new*) e o nome do arquivo de saída onde estará o resultado da sumarização.

Após isso, deve-se escolher a estratégia de sumarização. Se a escolha for Palavras-chaves + Título, ou Palavras-chaves + Título + Localização, ou Palavras-chaves + Título + Sinalizadoras, o usuário deverá escolher o método de extração de palavras-chaves, EPC-P ou EPC-R e digitar o número de palavras-chaves desejadas. Se o método escolhido for EPC-P, deve-se digitar no campo arquivo etiquetado o nome do arquivo com extensão *tag*. Após o cálculo das palavras-chave, o usuário deverá pressionar o botão "sumarizar". A partir daí, a sumarização será realizada e ao seu final será apresentada uma caixa de mensagem indicando o resultado da sumarização, como na Figura 2.

Se a escolha do tipo de sumarização for Palavras-chave do Autor, o usuário deverá digitar as palavras-chave desejadas, simples ou compostas, separadas por vírgulas, no campo indicado e pressionar o botão "sumarizar". O resultado da sumarização também será exibido como na Figura 2.

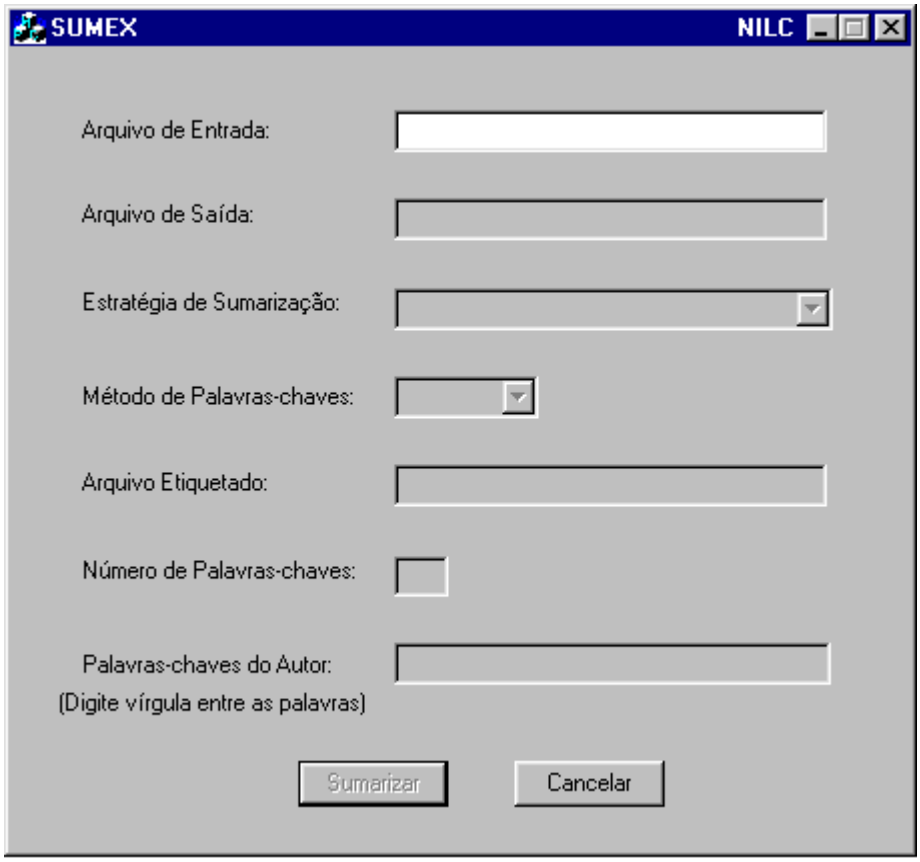

Figura 1 - Interface do SUMEX

Figura 2 - Caixa de mensagem que indica o fim da sumarização

3 AVALIAÇÃO DAS TÉCNICAS IMPLEMENTADAS

A avaliação do SUMEX foi feita utilizando-se dezoito artigos científicos de computação retirados da *Revista Brasileira de Informática na Educação* e dos anais do *Simpósio Brasileiro de Informática na Educação - 1998*. O sumário foi feito apenas da introdução dos artigos e não foi utilizado o título como palavra-chave, pois observou-se que praticamente em todas as frases da introdução, alguma palavra do título aparecia e o sumário gerado era então praticamente o texto original. Portanto, os delimitadores do título foram colocados no início do arquivo contendo somente um espaço em branco entre eles,

da seguinte forma: */t /t*, e os delimitadores de texto */i* foram colocados no início e no fim da introdução do artigo.

Gerou-se um sumário para cada estratégia de sumarização apresentada na seção 2.1 e foi feita uma análise da textualidade (coesão e coerência) e da proximidade (preservação da idéia principal do texto) de cada sumário. Para as estratégias Palavras-chaves + Título, ou Palavras-chaves + Título + Localização e Palavras-chaves + Título + Sinalizadoras, o número de palavras-chave utilizado foi cinco. Os sumários resultantes foram comparados com o texto original, com o sumário feito pelo autor do artigo científico e com o sumário feito pela ferramenta *AutoResumo* do *Word*. O percentual de sumarização utilizada no *AutoResumo* foi de 25% para todos os textos.

A Tabela 1 apresenta o percentual de sumarização de cada texto, para cada estratégia utilizada.

Estratégia 1: Palavras-chaves
Estratégia 2: Palavras-chaves + Localização
Estratégia 3: Palavras-chaves + Sinalizadoras
Estratégia 4: Palavras-chaves do Autor

Obs.: Lembrando que as palavras do título não foram utilizadas na sumarização dos textos.

Tabela 1 - Percentual do Original (número de setenças do sumário/ número total de sentenças do texto-fonte)

	Estratégia 1	Estratégia 2	Estratégia 3	Estratégia 4
Texto 1	33.3	91.7	41.7	8.3
Texto 2	52.6	84.6	63.2	31.6
Texto 3	63.6	90.9	63.6	45.5
Texto 4	22.2	88.9	66.7	44.4
Texto 5	19.0	61.9	33.3	4.8
Texto 6	38.9	83.3	38.9	22.2
Texto 7	54.5	81.8	54.5	36.4
Texto 8	54.5	72.7	59.1	9.1
Texto 9	50.0	71.9	65.6	9.4
Texto 10	37.5	81.3	50.0	25.0
Texto 11	9.5	52.4	23.8	23.8
Texto 12	88.9	100	88.9	44.4
Texto 13	51.9	74.1	51.9	11.1
Texto 14	30.8	61.5	30.8	15.4
Texto 15	0	50	12.5	0
Texto 16	66.7	66.7	66.7	66.7
Texto 17	33.3	33.3	50	83.3
Texto 18	57.1	71.4	71.4	28.6

Analisando a tabela 1, é notado que as estratégias 1 e 3 possuem, muitas vezes, resultados iguais. Isso acontece pois a diferença entre as duas estratégias é que na segunda,

palavras sinalizadoras também fazem parte da seleção de sentenças, e na maioria das vezes, quando estas surgem no texto, estão em frases que já possuem alguma palavra-chave.

A estratégia que obteve o pior resultado foi a estratégia 2, pois os sumários resultantes são muito extensos, não funcionando como resumo do texto-fonte. Isto ocorreu pois a maioria dos textos sumarizados possui parágrafos pequenos, com poucas sentenças.

A estratégia 4 também apresentou bons resultados, apesar dos percentuais de sumarização dos textos 1, 5, 8, 9 e 15 terem sido muito baixos.

A Tabela 2 apresenta o percentual de erros existente nos sumários de cada texto, para de cada estratégia e para o sumário feito pelo *AutoResumo*.

Tabela 2 - Percentual de erros de Coesão e Coerência (número de sentenças problemáticas / número total de sentenças do sumário)

	Estraté-gia 1	Estraté-gia 2	Estraté-gia 3	Estraté-gia 4	Auto Resumo
Texto 1	25	0	40	100	33.3
Texto 2	20	11.7	8.3	33.3	42.8
Texto 3	0	0	0	0	0
Texto 4	100	0	16.6	25	0
Texto 5	50	0	71.4	100	42.8
Texto 6	57.1	0	57.1	100	0
Texto 7	14.3	0	14.3	25	25
Texto 8	8.3	0	7.6	100	50
Texto 9	0	9.1	14.2	0	55.5
Texto 10	83.3	25	87.5	0	100
Texto 11	0	18	20	20	25
Texto 12	0	0	0	0	100
Texto 13	7.1	0	7.1	100	40
Texto 14	25	0	25	50	0
Texto 15	0	0	0	0	50
Texto 16	25	25	25	25	0
Texto 17	0	0	0	0	0
Texto 18	0	20	20	0	100

Analisando os dados da Tabela 2, observa-se que a estratégia que apresentou menor índice de erro foi a estratégia 2. Isto ocorre pois, analisando juntamente com a Tabela 1, o sumário resultante é praticamente o texto original, portanto, praticamente não apresenta erros de coesão e coerência. As estratégias 1 e 3 apresentaram resultados razoáveis; poucos textos apresentaram alto percentual de erros.

A maioria dos sumários resultantes do *AutoResumo* do *Word*, apresenta erros de coesão e coerência, mas poucos são os completamente incoerentes.

Na estratégia 4, 27.7% dos textos resultaram completamente incoerentes (obtendo o pior resultado entre todas as estratégias) mostrando que muitas vezes o autor não escolhe boas palavras-chave para seu artigo.

A Tabela 3 mostra se cada sumário manteve a idéia principal do texto-fonte ou não.

Tabela 3 - Preservação da idéia principal, comparando sumário e seu texto-fonte

	Estraté-gia 1	Estraté-gia 2	Estraté-gia 3	Estraté-gia 4	Auto Resumo
Texto 1	Sim	Sim	Sim	Não	Não
Texto 2	Sim	Sim	Sim	Sim	Sim
Texto 3	Sim	Sim	Sim	Sim	Sim
Texto 4	Não	Sim	Não	Não	Não
Texto 5	Não	Sim	Sim	Não	Não
Texto 6	Não	Sim	Não	Não	Sim
Texto 7	Sim	Sim	Sim	Não	Não
Texto 8	Sim	Sim	Sim	Não	Sim
Texto 9	Sim	Sim	Sim	Sim	Não
Texto 10	Sim	Sim	Sim	Sim	Não
Texto 11	Sim	Sim	Sim	Sim	Sim
Texto 12	Sim	Sim	Sim	Sim	Não
Texto 13	Sim	Sim	Sim	Não	Não
Texto 14	Sim	Sim	Sim	Não	Não
Texto 15	Não	Sim	Não	Não	Não
Texto 16	Sim	Sim	Sim	Sim	Sim
Texto 17	Sim	Sim	Sim	Sim	Sim
Texto 18	Sim	Sim	Sim	Não	Não

Observando a Tabela 3, conclui-se que todos os sumários obtidos pela estratégia 2 mantiveram a idéia principal do texto. Mas isso aconteceu pois os sumários são muito extensos e praticamente iguais aos textos originais.

As estratégias 1 e 3 apresentaram bons resultados, pois a maioria dos seus sumários preserva o tema principal. Já nos resultados da estratégia 4 e do *AutoResumo*, a maioria dos sumários não manteve a idéia principal do texto-fonte.

Analisando juntamente as três tabelas, observamos que a estratégia que apresentou melhores resultados relativos foi a estratégia 1, Palavras-chaves, pois na Tabela 1 sua porcentagem de sumarização foi média, na Tabela 2 seu percentual de erros, na maioria dos sumários, não foi muito alto e na Tabela 3 a maioria dos sumários preservou a idéia principal do texto-fonte.

3.1 Um exemplo de sumarização

Neste exemplo, são apresentados testes das estratégias de sumarização do ambiente SUMEX e do *AutoResumo* do *Word* aplicados para a introdução de um artigo científico. Os resultados são comparados com o texto original e com o sumário do autor do artigo.

Texto: Educação e Informática: a construção de um projeto na rede pública de ensino

É inconteste a importância assumida pela Informática nos dias de hoje. Dados os avanços científico-tecnológicos que se processam velozmente em todos os quadrantes do planeta, a Informática passou a ser um bem dos mais disputados pelos diversos países que buscam assegurar o seu lugar no contexto sócio-político mundial. Daí, a relevância que vem sendo atribuída aos sistemas de ensino, reconhecidos, universalmente, como instâncias que potencializam o domínio do conhecimento nesse campo, ao mesmo tempo, em que também incorporam os recursos da informática para melhor desempenho de suas próprias finalidades e objetivos.

Da mesma forma que a Informática por si só não resolve as questões candentes da sociedade, também não constitui uma panacéia para resolver todos os problemas do sistema de ensino e do cotidiano escolar. A sua utilização para surtir efeitos positivos e duradouros no plano educacional terá que ser efetivada no âmbito de uma programação de caráter pedagógico, o que requer, sobretudo, planejamento e estratégias adequados. E isto não será feito sem esforço e parcerias entre todos que desejam elevar o nível de produtividade do sistema escolar: governo, escola, comunidade.

No Brasil, o reconhecimento da Informática como um meio poderoso para alterar os indicadores educacionais que colocam o país como um dos últimos no *ranking* dos sistemas educacionais da América Latina, ainda é recente. Somente a partir de meados dos anos 80, a Informática ultrapassou, de fato, os muros da universidade e passou a ocupar espaço na agenda dos governos. Assim, pode-se observar que aos poucos vão sendo redirecionadas as políticas da área, constituindo-se a informatização dos sistemas de ensino um item prioritário.

Nesse sentido, várias secretarias estaduais e municipais de educação têm procurado desenvolver ações voltadas para a implantação de redes informáticas e para o desenvolvimento de programas de capacitação de pessoal docente. Em geral, essas iniciativas têm encontrado inúmeros obstáculos, não apenas devido ao custo elevado para aquisição do equipamento tecnológico, como também, pela ausência de profissionais com formação na área de educação e informática.

Examinar os processos desencadeados pelas secretarias de educação no país, portanto, constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear as políticas voltadas para a Informática na Educação. Nessa ótica, a experiência que vem sendo desenvolvida na rede municipal de ensino do Recife merece ser aqui analisada.

(A) Sumário do artigo fornecido pelo autor na seção RESUMO da publicação:

Este artigo trata da implantação da Informática no âmbito da rede municipal de ensino do Recife, destacando-se as ações voltadas para a criação de infra-estrutura, associada a um programa de formação de recursos humanos e desenvolvimento de projetos pedagógicos. Uma análise crítica e perspectivas para a continuidade do processo são apresentadas.

(B) Sumário gerado pelo *Auto-Resumo* do *Word* (25% do texto original):

Da mesma forma que a Informática por si só não resolve as questões candentes da sociedade, também não constitui uma panacéia para resolver todos os problemas do sistema de ensino e do cotidiano escolar. No Brasil, o reconhecimento da Informática como um meio poderoso para alterar os indicadores educacionais que colocam o país como um dos últimos no *ranking* dos sistemas educacionais da América Latina, ainda é recente. Examinar os processos desencadeados pelas secretarias de educação no país, portanto, constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear as políticas voltadas para a Informática na Educação.

Análise: A ligação entre a primeira e a segunda senteças não é boa, ocorrendo um problema de coesão e de coerência. O assunto principal, a implantação da Informática no âmbito da rede municipal de ensino do Recife, não é sequer citado.

(C) Sumários gerados no SUMEX (lembrando que as palavras do título não serão utilizadas):

Os sumários feitos pelo SUMEX, apresentados abaixo, possuem espaços em branco entre as pontuações devido ao uso do programa *Tokennizer*, como explicado na seção 2.

- **Estratégia 1: Palavras-chaves**

Número de Palavras-chaves: 5

Resultado:

Radicais das palavras-chaves:

informat na red municipal
red municipal de ensin
projet
informat na educ
program

Sumário (33.3%):

A sua utilização para surtir efeitos positivos e duradouros no plano educacional terá que ser efetivada no âmbito de uma programação de caráter pedagógico , o que requer , sobretudo , planejamento e estratégias adequados .

Nesse sentido , várias secretarias estaduais e municipais de educação têm procurado desenvolver ações voltadas para a implantação de redes informáticas e para o desenvolvimento de programas de capacitação de pessoal docente .

Examinar os processos desencadeados pelas secretarias de educação no país , portanto , constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear

as políticas voltadas para a Informática na Educação . Nessa ótica , a experiência que vem sendo desenvolvida na rede municipal de ensino do Recife merece ser aqui analisada .

Análise: No primeiro parágrafo ocorre um erro de coesão, pois no início da frase há referência a frase anterior que não consta no sumário. Mas se esse problema for resolvido teremos um sumário bom, que contém a idéia principal do texto.

- **Estratégia 2: Palavras-chaves + Localização**

Número de Palavras-chaves: 5

Resultado:

Radicais das palavras-chaves:

informat red municipal
red municipal de ensin
projet
informat educ
program

Sumário (91.7%):

É inconteste a importância assumida pela Informática nos dias de hoje. Dados os avanços científico-tecnológicos que se processam velozmente em todos os quadrantes do planeta , a Informática passou a ser um bem dos mais disputados pelos diversos países que buscam assegurar o seu lugar no contexto sócio-político mundial . Daí , a relevância que vem sendo atribuída aos sistemas de ensino , reconhecidos , universalmente , como instâncias que potencializam o domínio do conhecimento nesse campo , ao mesmo tempo , em que também incorporam os recursos da informática para melhor desempenho de suas próprias finalidades e objetivos .

Da mesma forma que a Informática por si só não resolve as questões candentes da sociedade , também não constitui uma panacéia para resolver todos os problemas do sistema de ensino e do cotidiano escolar . A sua utilização para surtir efeitos positivos e duradouros no plano educacional terá que ser efetivada no âmbito de uma programação de caráter pedagógico , o que requer , sobretudo , planejamento e estratégias adequados . E isto não será feito sem esforço e parcerias entre todos que desejam elevar o nível de produtividade do sistema escolar : governo , escola , comunidade .

No Brasil , o reconhecimento da Informática como um meio poderoso para alterar os indicadores educacionais que colocam o país como um dos últimos no ranking dos sistemas educacionais da América Latina , ainda é recente . Assim , pode-se observar que aos poucos vão sendo redirecionadas as políticas da área , constituindo-se a informatização dos sistemas de ensino um item prioritário .

Nesse sentido , várias secretarias estaduais e municipais de educação têm procurado desenvolver ações voltadas para a implantação de redes informáticas e para o desenvolvimento de programas de capacitação de pessoal docente . Em geral , essas iniciativas têm encontrado inúmeros obstáculos , não apenas devido ao custo elevado para

aquisição do equipamento tecnológico , como também , pela ausência de profissionais com formação na área de educação e informática .

Examinar os processos desencadeados pelas secretarias de educação no país , portanto , constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear as políticas voltadas para a Informática na Educação . Nessa ótica , a experiência que vem sendo desenvolvida na rede municipal de ensino do Recife merece ser aqui analisada .

Análise: O resultado não é eficiente pois o sumário é praticamente igual ao texto-fonte. Isto ocorreu pois o texto possui parágrafos pequenos, com poucas frases.

- **Estratégia 3: Palavras-chaves + Palavras Sinalizadoras**

Número de Palavras-chaves: 5

Resultado:

Radicais das palavras-chaves:

informat red municipal
red municipal de ensin
projet
informat educ
program
objet
result
nest artig
este artig
nest text
este text
a conclusã
conclusõ
nest trabalh
este trabalh

Sumário (41.7%):

Daí , a relevância que vem sendo atribuída aos sistemas de ensino , reconhecidos , universalmente , como instâncias que potencializam o domínio do conhecimento nesse campo , ao mesmo tempo , em que também incorporam os recursos da informática para melhor desempenho de suas próprias finalidades e objetivos .

A sua utilização para surtir efeitos positivos e duradouros no plano educacional terá que ser efetivada no âmbito de uma programação de caráter pedagógico , o que requer , sobretudo , planejamento e estratégias adequados .

Nesse sentido , várias secretarias estaduais e municipais de educação têm procurado desenvolver ações voltadas para a implantação de redes informáticas e para o desenvolvimento de programas de capacitação de pessoal docente .

Examinar os processos desencadeados pelas secretarias de educação no país , portanto , constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear as políticas voltadas para a Informática na Educação . Nessa ótica , a experiência que vem sendo desenvolvida na rede municipal de ensino do Recife merece ser aqui analisada .

Análise: O sumário resultante é semelhante ao apresentado na primeira estratégia. A diferença é que a primeira sentença foi acrescentada pois possui a palavra sinalizadora "objetivo". Ocorrem erros de coesão textual na primeira e na segunda sentenças. A primeira sentença ainda gera um problema de coerência pois não se encaixa no contexto do sumário.

- **Estratégia 4: Palavras-chaves do Autor**

Radicais das palavras-chaves:

polit de informat educ
formaçã de recurs human
informat educ

Sumário (8.3%):

Examinar os processos desencadeados pelas secretarias de educação no país , portanto , constitui um exercício fundamental para o debate sobre as perspectivas que deverão nortear as políticas voltadas para a Informática na Educação .

Análise: O sumário apresentado possui problemas de coesão, pois possui a palavra "portanto" que indica a conclusão de uma idéia, mas que neste sumário não foi antes apresentada.

Análise Geral: O sumário que mais se aproximou do sumário do autor do artigo científico foi o resultante da Estratégia 1, Palavras-chaves, que apesar de conter um problema de coesão, preservou a idéia principal do texto. O sumário produzido pelo *AutoResumo* do *Word* apresentou problemas de coerência, não sendo um bom sumário.

4 CONCLUSÕES

O estudo feito mostra que também para a língua portuguesa problemas de textualidade e proximidade são muito evidentes nos sumários produzidos por métodos que utilizam técnicas de sumarização automática extrativa. Vale a pena, no entanto, continuar a investigação de técnicas de sumarização de textos em português, extrativas ou fundamentais, a fim de colocar a língua portuguesa em pé de igualdade com outras, quanto a existência de ferramentas de processamento de línguas naturais.

Ressalta-se que para determinadas aplicações, apesar de suas limitações, técnicas simples como as apresentadas aqui podem ser suficientes e vantajosas.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro do Programa PIBIC-CNPq (USP) e a colaboração dos pesquisadores do NILC.

REFERÊNCIAS BIBLIOGRÁFICAS

ALUÍSIO, S.M.; AIRES, R.V. *Etiquetação de um Corpus e Construção de um Etiquetador de Português.* Relatórios Técnicos do ICMC-USP, 107 (NILC-TR-00-2). Março 2000, 18p.

KUPIEC, J.; PEDERSEN J.; and CHEN F. 1995. *A trainable document summarizer*. In Proceedings of the 18th ACM-SIGIR Conference, pages 68--73.

MARTINS, C.B.; PARDO, T.A.S.; ESPINA, A.P.; RINO, L.H.M. *Introdução à Sumarização Automática.* Rel. Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos. Abril, 2001. 38p.

PEREIRA, M. *Algoritmos de Extração de Palavras-chaves em Português*. NILC-TR-01-06, Setembro, 2001.

PORTER, M. F. *An algorithm for suffix stripping*. Program, 14(3):130--137, 1980.

RINO, L.H.M. *Exploração de métodos diversos para a sumarização automática*. Projeto de Pesquisa Fapesp. 2001.

TURNEY, Peter (1999). *Learning to Extract Keyphrases from Text*, Tech. Report Number NRC-41622, National Research Council Canada, Institute for Information Technology.

TURNEY, Peter. *Extraction of keyphrases from text: Evaluation of four algorithms*.
Technical report, Institute for Information Technology, 1997.