Universidade de Évora

Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Sumarização Automática de Texto

Luís Filipe Romão Rodrigues

Orientador: *Prof. Doutor Paulo Quaresma*

Setembro 2011

Universidade de Évora

Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Sumarização Automática de Texto

Luís Filipe Romão Rodrigues

Orientador: *Prof. Doutor Paulo Quaresma*

Setembro 2011

Resumo

Sumarização Automática de Texto

Sumarizar é uma actividade frequentemente realizada pelo ser humano. Quando se narra um evento, em geral, é costume fazer um sumário do que aconteceu e não fazer uma narração completa e detalhada.

A sumarização automática de texto é uma técnica que utiliza um programa de computador para gerar estruturas sintéticas que contêm as informações mais relevantes de um textos. O texto original é passado ao programa sendo transformado numa versão condensada. Esta área das ciências da computação tem a sua origem no final dos anos 50 e tem vindo a ser investigada desde então. O aumento exponencial de informação disponível hoje devido principalmente à Internet, coloca a sumarização automática de novo em voga. Assim, é essencial o desenvolvimento de novas metodologias e técnicas de forma a ser possível a rápida consulta e fácil acesso a toda informação disponível ao ser humano.

A dissertação proposta apresenta o estudo de uma abordagem e a implementação de um sistema simbólico (em oposição à abordagem estatística) de sumarização automática para a língua portuguesa. Os sistema utiliza a teoria da estrutura retórica para o reconhecimento de relações entre segmentos, fazendo uso do modelo desenvolvido no sistema AuTema-Dis (Leal, 2008). Uma arquitectura modular em quatro etapas que processa um texto desde a sua forma original até a geração do sumário final.

No final deste trabalho é feita uma avaliação que compara a performance dos vários sistemas de sumarização para a língua portuguesa. É feita uma avaliação qualitativa do sistema desenvolvido neste projecto recorrendo a juízes humanos falantes nativos do português de Portugal.

Abstract

Automatic Text Summarization

Summarization is an activity often performed by humans. When an event is narrated, in general, it is customary to make a summary of what happened not detailed narration.

Automatic text summarization is a technique that uses a computer program to generate synthetic structures that contain the most relevant information of a text. The original text is used as input for the computer program and is transformed into a condensed version. This area of computer science has its origins in the late 50's and has been continuously researched since then. The exponential growth of information available due mainly to the Internet, puts the automatic summarization once again in vogue. It is therefore essential to develop new methodologies and techniques so as to be possible an easy access to all information available to humans.

This thesis addresses an approach and the implementation of a symbolic system (as opposed to statistical approach) for automatic summarization for the Portuguese language. The system uses the theory of rhetorical structure for recognizing relationships between segments, using also the model developed in AuTema-Dis (Leal, 2008). This system defines a modular four steps architecture which processes a text from its original form into the final summary.

At the end of this thesis there is a evaluation that compares the performance leading systems for automatic summarization for the Portuguese language. The system presented in this thesis is evaluated resorting to human judges (all native speakers of Portuguese from Portugal).

Agradecimentos

Quero agradecer a um conjunto de pessoas que me ajudaram a desenvolver o trabalho aqui apresentado, todas elas parcialmente responsáveis pelo seu conteúdo final.

Primeiramente, quero expressar os meus sinceros agradecimentos ao Professor Paulo Quaresma pelo incentivo, acompanhamento e disponibilidade sobre a evolução deste trabalho. Agradeço também à Ana Luísa sem a qual este trabalho não teria sido possível.

Agradeço também aos meus pais e a toda a minha família pelo grande apoio que me deram ao longo deste trabalho.

Quero também agradecer à Marta e a todos os meus colegas e amigos que não poderia aqui nomear e que me apoiaram ao longo deste trabalho, em projectos académicos e em tantos outros.

Não poderia claro deixar de agradecer à Universidade de Évora, a todas as pessoas que lá estão e que marcaram a minha passagem pelo Alentejo.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Esta dissertação descreve um trabalho de investigação numa área especifica de processamento de língua natural: os sistemas de geração automática de sumários. Ao longo dos vários capítulos são apresentados, de uma forma muito geral, diversos sistemas e algumas técnicas e teorias usadas nesta área. O principal foco deste trabalho é um sistema de sumarização para a língua portuguesa que foi desenvolvido na Universidade de Évora como parte do projecto AuTema-Dis (Leal, 2008). É importante referir que apesar deste sistema ser apresentado como um sistema de sumarização de texto não o é exactamente. O sistema realiza a compressão textual de frases aglutinando-as para a geração do sumário final. É um sistema de sumarização na medida que reduz a quantidade de informação de um dado texto.

O projecto AuTema-Dis propõe uma arquitectura que, implementada computacionalmente, realiza a análise textual, considerando as informações mais relevantes dispostas na superfície de um texto, bem como, as relações de significação que se estabelecem entre os elementos linguísticos que a compõe. Observando os vários módulos que compõem o sistema percebeu-se que seria possível, com algumas alterações, utilizar a Teoria RST (Mann e Thompson, 1988) para a realização de sumários de texto.

1.1 Contexto

Fazer sumários é uma das actividades mais comuns na comunicação através de linguagem natural. Quando um indivíduo conta os acontecimentos de uma reunião, os comentários que alguém fez de uma outra pessoa, o tema principal de um filme ou livro, ou quais são as últimas notícias, ele certamente expressará

de forma condensada as partes importantes dessas informações (Hutchins, 1987) e não fará uma descrição detalhada dos factos. Inconscientemente, estamos sempre a fazer sumários. Os mais frequentes são os sumários escritos, como, por exemplo, notícias em jornais, artigos em revistas ou resumos de textos científicos.

Com os grandes avanços nas Tecnologias de Informação e Comunicação surgiram grandes depósitos digitais de textos das mais diversas naturezas. A Internet como meio de comunicação global aumentou o número de textos disponíveis *online*. Temos um ambiente universal e heterogéneo onde a informação é transmitida em grandes volumes e em curtos períodos de tempo. Um estudo da Universidade de Berkeley (Lyman e Varian, 2003) indica que foram criados 5 milhões de terabytes de novas informações (livros, revistas, filmes, música, etc) em 2002, só a *world wide web* continha cerca de 170 terabytes de informação. Estes valores são cerca do dobro da informação gerada em 1999. Na altura, esperava-se um crescimento de 30% ao ano. A rede mundial de computadores acabou por mudar o conceito de informação: o mais informado não é aquele que possui a maior quantidade de informação, mas aquele de dispõe dos melhores recursos para a obter, analisar e utilizar da forma mais rápida possível.

Armazenamento	Terabytes 2002	Terabytes 1999/2000	Alteração
Papel	1634	1200	36%
Filme	420254	431690	-3%
Magnético	5187130	2779760	87%
Óptico	103	81	28%
Total	5609121	3212731	74.5%

Tabela 1.1: Produção mundial de informação original (se armazenada digitalmente) em terabytes em 2002 (Lyman e Varian, 2003).

Neste sentido a sumarização automática tornou-se uma tarefa de grande relevância na sociedade moderna devido aos problemas associados à sobrecarga de informação. O excesso de informação textual compromete a eficiência do tratamento e utilização da mesma pelos seres humanos, o que acarreta perdas de precisão nos resultados desejados, aumento tempo necessário para a realização de tarefas e frequentemente o incremento do custo com os procedimentos. Recuperar, manipular e consumir informação em linguagem natural são tarefas muito complexas mas ao

mesmo tempo extremamente importantes.

Cada vez mais torna-se evidente a necessidade de sumários automáticos. Os jornais fazem um uso intenso de cabeçalhos das notícias para lhes dar destaque, são feitas sinopses das principais notícias. Nas revistas técnicas ou cientificas que frequentemente contêm temas condensados. No meio académico, seria útil para os estudantes utilizarem versões abreviadas de obras literárias para melhor assimilar, em períodos reduzidos de tempo, os principais aspectos de um determinado tema. Ao navegar pela *world wide web*, um utilizador pode reduzir o tempo e o esforço necessários para localizar e assimilar o que é essencial (Rino e Pardo, 2003). A geração de um sumário de forma automática tem ainda a vantagem do seu tamanho poder ser controlado e o seu conteúdo ser determinístico.

Esta área do processamento de língua natural tem vindo a ser objecto de estudo desde os primórdios das ciências da computação. No final da década de 1950 começaram a surgir métodos estatísticos para extrair as frases principais de um texto. Desde então até hoje esta área nunca parou, os actuais investigadores sugerem que o desenvolvimento e a avaliação de sistemas de sumarização automática constituem ainda um tema muito promissor.

1.2 Conceitos

Torna-se importante fazer uma apresentação dos conceitos fundamentáveis desta área do processamento da língua natural. Para (Radev, Hovy e McKeown, 2002) um sumário é "um texto que é produzido a partir de um ou mais textos, que transmite a informação importante do(s) texto(s) original(ais) e que não é maior do que metade do seu tamanho (sendo usualmente menor)". Esta definição simples engloba três aspectos importantes que caracterizam a sumarização automática:

- sumários podem ser produzidos a partir de um ou mais textos;
- sumários devem preservar a informação mais importante do texto original;
- sumários devem ser tão curtos quanto possível.

Segundo (Rino e Pardo, 2003), para a criação de um sumário é necessário que se verifiquem algumas características referentes ao texto original, as quais são denominadas por "premissas da sumarização":

- deve haver um texto-fonte para ser sumarizado;

- deve existir, por ser um texto o objecto da sumarização::

    1. uma ideia, ou assunto, central a partir da qual é construída a trama textual;
    2. um conjunto de unidades de informação que possuam uma relação nítida com a ideia central do texto-fonte;
    3. um objectivo comunicativo central que orienta tanto a selecção de unidades de informação quanto a escolha da estrutura textual, para estabelecer a ideia pretendida;
    4. um enredo, elaborado em função das escolhas supracitadas, que tem por objectivo transmitir coerentemente a ideia central, para que os objectivos comunicativos sejam atingidos.

- a principal premissa para a sumarização de texto é a tarefa de identificar o conteúdo relevante de um texto e utilizar esta informação para construir um novo enredo, utilizando o conteúdo disponível e preservando a ideia central no sumário correspondente. A sumarização nunca deve modificar a ideia proposta no texto original.

Existem inúmeros trabalhos que se dedicam identificar os vários tipos de sumários existentes, é de destacar, no entanto, o trabalho de (Hutchins, 1987). Para ele, existem três tipos de sumários: indicativos, informativos e sumários de crítica (*evaluative*).

- os sumários indicativos contêm apenas os assuntos essenciais de um texto não contendo necessariamente detalhes de resultados, argumentos e conclusões. Estes sumários podem servir como indexadores, isto é, o leitor percebe se o tema lhe interessa, no caso afirmativo irá consultar mais informação ao texto original;

- sumários informativos são auto-contidos, ou seja, detêm informações suficiente sobre o texto inicial que podem ser considerados substitutos do mesmo. Estes sumários devem conter todos os aspectos principais;

- os sumários de crítica avaliam e apresentam uma análise comparativa do conteúdo do texto original com trabalhos relacionados na mesma temática.

Segundo (Hutchins, 1987), é mais fácil produzir automaticamente sumários indicativos devido à complexidade de modelar a sumarização humana para os outros tipos de sumários.

(Sparck Jones, 1995) também desenvolveu algum trabalho na nesta área, a autora esclarece a distinção entre sumários e índices. Segundo ela, sumários são textos que podem ser substitutos do documento original (seriam então equivalentes aos sumários informativos de Hutchins). No entanto, a autora considera que não poderiam ser utilizados como substitutos para os textos originais, pois não teriam necessariamente o que estes continham de mais importante, quer em termos de conteúdo quer de estrutura. Estes sumários apenas transmitiam uma ideia vaga do texto original, podendo até ser apresentados numa forma não textual (podiam ser apenas uma lista de itens). Para a autora, os índices em formato textual poderiam ser comparáveis aos sumários indicativos de (Hutchins, 1987).

Observa-se que quando um ser humano faz um sumário tem em conta os objectivos do texto a sumarizar, os objectivos ou interesses dos possíveis leitores, o seu próprio conhecimento da área em questão, os seus valores e a importância relativa (e bastante subjectiva) que atribui a cada uma das frases. Na escolha do título (como veremos mais à frente, o título de um artigo pode ser considerado um sumário) o autor de um artigo jornalístico que refere um acidente de um importante líder local pode, por exemplo referir a morte, assim o título seria: "Morreu o Presidente Paulo Silva num acidente na auto-estrada"; pode ainda considerar que referir a morte é extremamente desagradável e focar-se apenas no desastre, assim temos: "Acidente na A1 acaba em desastre"; Pode suavizar ainda mais a questão e apenas informar da ocorrência do acidente: "Grande acidente ontem na A1". Facilmente se percebe, com este exemplo, que existe uma grande multiplicidade de frases ou estruturas nos sumários e, portanto, é possível produzir mais do que um sumário para um mesmo texto original.

Segundo (Hutchins, 1987), a análise do conteúdo dos documentos é uma das actividades mais importantes para a geração de um sumário. Compreender como os sumarizadores humanos desempenham esta tarefa poderia levar a avanços consideráveis na automatização da mesma. Há, no entanto, uma grande dificuldade na compreensão dos processos utilizados devido à sua grande subjectividade. Nalguns casos, os profissionais seguem uma sequência directa de raciocínio e construção que é possível modelar. O processo como os profissionais da sumarização esquadrinham, seleccionam e constroem um sumário a partir de um texto pode ser analisado e criada uma caracterização do processos envolvidos nas diversas fases. Contudo, (Endres-Niggemeyer, 1990) observa que o modo como eles realizam cada

um desses passos pode variar de acordo com o indivíduo o que dificulta a modelação necessária. Hutchins argumenta ainda que, em textos expositivos, o leitor apenas se lembra da ideia central ou argumento principal *gist* do texto original, classificando as outras características como positivas ou negativas em função desta ideia.

Na sumarização automática existem duas abordagens: a superficial ou estatística e a profunda ou simbólica, que embora sejam métodos distintos de sumarização, não competem e podem ser utilizados de forma combinada. Ambas procuram identificar a ideia central dos textos originais para posteriormente estabelecer os elementos que constituirão o sumário. A diferença principal está na proposta de construção do sumário.

A abordagem superficial baseia-se em métodos experimentais e estatísticos, os quais são empregados principalmente na produção de extractos. Inicialmente são identificados os segmentos mais importantes de um texto-fonte e, utilizando estes, são produzidos extractos através da justaposição. Por se basear na extracção a partir de um texto original, as técnicas adoptadas na abordagem superficial também são denominadas técnicas extractivas. Nesta estratégia, raramente as frases escolhidas sofrem algum tipo de modificação.

No que se refere à abordagem profunda apoia-se em teorias linguísticas, sendo assim uma abordagem consideravelmente mais complexa, simula a reescrita integral do sumário e explora diversas características linguísticas e extra-linguísticas tais como os objectivos comunicativos do autor ao escrever o texto, relações semânticas e retóricas. Dessa forma, estes sumários podem conter informações não necessariamente existentes no documento original. A escolha de uma linguagem eficaz para representar o conhecimento contido nos textos é fundamental nesta abordagem. Sem uma representação adequada do significado do texto dificilmente serão produzidos bons sumários.

Sendo esta uma área com diversas abordagens torna-se necessário a criação de métodos de avaliação. Os métodos de avaliação podem ser classificados em duas categorias (Mani et al., 1999): intrínseca e extrínseca. Na avaliação intrínseca, a qualidade de um sumarizador é avaliada pela análise da própria qualidade dos sumários. O que pode ser conseguido recorrendo a um conjunto de directrizes ou normas como o julgamento humano do sumário e/ou a presença da ideia essencial do texto-fonte no sumário (Rino e Pardo, 2003). Outro método é a comparação de similaridade entre o sumário automático e um sumário de referência, denominado sumário ideal (Edmundson, 1969; Rino et al., 2004; Jing et al., 1998).

Segundo (Jing et al., 1998; Mani et al., 1999), a avaliação extrínseca tem como objectivo verificar a qualidade da sumarização em função dos resultados obtidos para a realização de outras tarefas, tais como: categorização, recuperação de informação, quão simples é apreender o assunto através da leitura do sumário ou uma actividade denominada *question-answering* (a qual tem como objectivo verificar se os sumários retêm informação suficiente para que os participantes respondam a sua série de perguntas sobre o tema apresentado no texto original).

Desde a sua génese a sumarização automática apenas se focou na geração de sumários para um único documento. Neste caso, o fluxo de informação não é uniforme o que implica que existem partes mais importantes do que outras. O desafio reside em criar um sistema que consiga distinguir quais as partes que contêm mais informação. A maioria do trabalho disponível na literatura utiliza a extracção de frases para a construção de sumários, no entanto, mais recentemente começaram-se a utilizar outro tipo de técnicas que utilizam conhecimentos mais complexos da análise de língua natural. Nos últimos anos, com o advento da Internet, surgiu a necessidade de gerar sumários a partir de um conjunto de múltiplos documentos, muitas vezes disponíveis em várias línguas. Esta área da sumarização automática não é discutida nesta trabalho, apenas é focada a criação de sumários para um único documento e na língua portuguesa nas variantes Português Europeu e Português do Brasil.

1.3 Objectivos

O objectivo principal deste trabalho é propor um sistema automático de geração de sumários para a língua portuguesa nas variantes Português Europeu e Português do Brasil.

Este trabalho tem ainda vários objectivos imediatos. Pretende-se fazer uma avaliação dos vários sistemas de sumarização automática que existem para a língua portuguesa. Com vista a criação do sistema supracitado irá ser proposta uma arquitectura modular para a geração de sumários. Há ainda a intenção clara de fazer a avaliação da solução proposta utilizando um corpus de textos "reais", para tal será utilizado o TeMário[1] (Pardo e Rino, 2003).

Como é detalhado em capítulos posteriores, a implementação deste protótipo usa algumas abordagens diferentes das que são usadas pela maior parte dos sis-

[1]TeMário é um corpus desenvolvido para a sumarização automática de textos, é composto por textos jornalísticos e seus sumários em português.

temas geração de sumários, sendo aplicados métodos inovadores em alguns dos componentes do sistema.

1.4 Contribuição

A principal contribuição deste projecto é a criação de uma arquitectura modular para a geração de sumários para a língua portuguesa. O sistema apresentado funciona de forma totalmente independente do utilizador e sem a necessidade de intervenção de um especialista no processo.

Outro grande foco desta tese é a descrição e avaliação comparativa das várias soluções existentes para sumarização automática de textos escritos em Português.

Uma última contribuição a destacar é a criação de uma aplicação *web based* recorrendo a tecnologias inovadoras nesta área, de forma a facilitar a interacção do utilizador com o sistema. Com esta aplicação um utilizador facilmente cria sumários, sendo possível ajustar os vários parâmetros de configuração de forma simples, utilizando uma interface "amigável".

1.5 Estrutura da tese

Esta dissertação descreve o trabalho de investigação realizado com vista ao desenvolvimento de um sistema geração automática de sumários para a língua portuguesa. Nesta secção são descritos os vários capítulos que compõem esta dissertação.

No capítulo 1 é feita uma introdução do trabalho efectuado, a sua importância e os objectivos pretendidos. É ainda contextualizada a necessidade da geração automática de sumários e definidos os conceitos basilares desta área do processamento de língua natural.

No capítulo 2 é analisada a história e evolução dos sistemas de sumarização automática ao longo dos anos. É feita uma apresentação geral do início desta área do processamento de língua natural, seguida de uma discussão mais aprofundada sobre os dois tipo de abordagens existentes: superficial ou estatística e profunda ou simbólica. São apresentados alguns projectos existentes, nesta área, para língua portuguesa. Finalmente é discutida a teoria da estrutura retórica e apresentado o trabalho de (Leal, 2008), ambos utilizados como fundamento teórico do trabalho apresentado nesta tese.

No capítulo 3 é descrito o sistema de sumarização implementado, descrevendo-se todos os pormenores do sistema concreto. São apresentados os vários módulos desde a gramática até ao sistema de geração de extractos, sendo exibidas diversas figuras de forma a facilitar a compreensão do funcionamento do sistema.

O capítulo 4 inicia com uma breve discussão da avaliação de sumários automáticos, sendo abordadas os métodos intrínsecos e extrínsecos. É feita uma breve avaliação de outros sistema para a língua portuguesa. De seguida, é apresentado o método escolhido e a sua fundamentação. Finalmente é feita uma avaliação crítica do sistema que foi implementado, mostrando-se os pontos fortes e fracos. Para a avaliação foram utilizados corpus compostos por artigos jornalísticos provenientes do Jornal Público (português europeu) e Folha de São Paulo (português do Brasil).

No capítulo 5 é apresentada a conclusão sobre todo o trabalho, analisando-se de uma forma muito geral os pontos fortes do sistema e os pontos fracos. Neste capítulo está também patente uma pequena referência a algum trabalho futuro que é possível desenvolver sobre este sistema.

1.6 Resumo do Capítulo

Neste capítulo foi feita uma leve apresentação do trabalho que se desenvolveu na Universidade de Évora com o objectivo de produzir um sistema de sumarização automático para a língua portuguesa. Inicialmente foi apresentado um dos grandes desafios da sociedade da informação e comunicação: o excesso e informação e a geração automática de sumários como um possível solução a longo prazo. Após a definição do que é um sumário foram discutidos os conceitos na base dos sumários de texto e a influência de vários autores na definição desta área do processamento da língua natural.

No próximo capítulo é analisada a história e evolução dos sistemas de sumarização automática. É feita uma apresentação geral da história deste campo das ciências da computação, seguida de uma discussão mais aprofundada sobre os dois tipo de abordagens existentes: superficial ou estatística e profunda ou simbólica. São apresentados alguns projectos existentes, nesta área, para língua portuguesa. Finalmente é discutida a teoria da estrutura retórica e apresentado o trabalho de (Leal, 2008).

Capítulo 2

Estado da arte

O fluxo de informação num texto não é uniforme, isto é, há partes com maior importância que outras. O maior desafio da sumarização automática consiste em distinguir as secções que contêm a informação mais importante. Apesar de não ser a única possibilidade, a maioria dos trabalhos apresentam soluções baseada na extracção *ipsis verbis*[1] de frases dos textos a sumarizar.

Neste capítulo, serão apresentadas algumas destas técnicas extractivas, primeiramente serão focados os trabalhos iniciais dos anos 50 e 60 que iniciaram a investigação na sumarização automática. Seguidamente serão tratados, em pormenor, os vários métodos superficiais e profundos. Por fim, será apresentado o estado da sumarização automática para a língua portuguesa nos dias de hoje.

2.1 Trabalhos iniciais

Segundo (Rino e Pardo, 2003), a sumarização automática pode ser vista de forma genérica como uma tarefa composta por três processos: análise, transformação e síntese. Na análise, é elaborada uma representação computacional do texto-fonte. De seguida, no processo de transformação, é modificado o resultado produzido na análise de forma a gerar a representação do sumário. Finalmente, na síntese, a estrutura representativa do sumário é convertida numa expressão linguística, isto é, no sumário. Esta arquitectura é ilustrada na figura 2.1.

O trabalho inicial na área da sumarização automática focou-se em documentos técnicos. Provavelmente o trabalho mais citado sobre sumarização é o de (Luhn,

[1] expressão em latim cujo significado é "com as mesmas palavras"

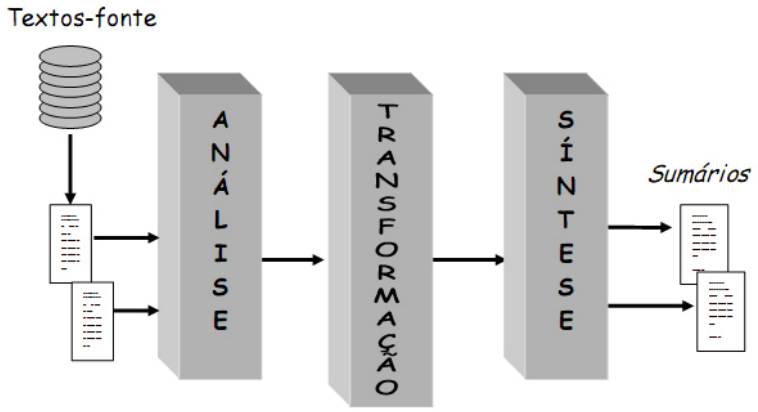

Figura 2.1: Arquitectura de um sistema de sumarização automática (Rino e Pardo, 2003).

1958), que descreve a investigação desenvolvida na IBM nos anos 50. Naquele artigo aparecem, pela primeira vez, vários conceitos-chave que mais tarde serão usados noutros trabalhos. No seu trabalho, Luhn propõe que através da frequência de uma dada palavra num artigo é possível apreender da sua importância. Como primeiro passo as palavras são transformadas para a sua forma radical e as *stop words*[2] são eliminadas.

Luhn compilou uma lista de palavras referentes ao conteúdo de um texto e procedeu a sua ordenação por frequência. Este índice, de palavras significantes, era utilizado para medir a significância de uma dada palavra. Numa frase, o nível de significância era derivado do número de ocorrências de palavras com significância e da distância linear entre elas considerando as palavras não significantes que se encontravam pelo meio. Por fim, as frases eram ordenadas pela sua significância, sendo seleccionadas como o sumário automático as que possuíam um nível de significância mais elevando.

Num trabalho relacionado, também desenvolvido na IBM, (Baxendale, 1958) apresenta uma característica particularmente útil para encontrar secções importantes de um texto, ele considerou a posição das frases no texto. Ao examinar uma amostra de 200 parágrafos, Baxendale mostrou que em 85% dos analisados, a frase que referia o assunto era a primeira e em 7% a última. Assim, Baxendale

[2]As *stop words* são palavras que são filtradas antes ou depois do processamento de um texto em língua natural. Apesar de não existir uma lista completamente bem definida é comummente aceite incluir substantivos comuns e artigos.

sugeriu que deveriam ser incluídas num sumário tanto a primeira como a última frase de cada parágrafo.

Outro sistema para a criação de extractos foi apresentado por (Edmundson, 1969). A sua contribuição principal foi desenvolver a estrutura típica para a criação de uma experiência de extracção de sumários. Primeiramente, desenvolveu um protocolo para criar extractos manuais o qual foi aplicado a 400 documentos técnicos. A frequência das palavras e o posicionamento das frases dos dois trabalhos anteriores foi tida em consideração, adicionou ainda duas novas características: a presença de *cue words*[3] e o esqueleto do texto (se uma frase é título, cabeçalho, etc.). Foram dados pesos a cada uma das características de forma a ser possível atribuir uma pontuação a cada frase. Durante a avaliação foi descoberto que 44% dos extractos automáticos coincidiam com os extractos manuais.

Durante muito tempo a exploração dos métodos explanados nestas obras seminais ficou estagnada devido à impossibilidade técnica da sua implementação (não só limitações de *hardware* e *software*, mas também de recursos electrónicos como dicionários ou repositórios linguísticos de grande porte). Na década de 90, vemos o ressurgimento destas abordagens devido aos computadores terem passado a ser de uso geral e os recursos linguísticos terem aumentado. O conhecimento de manipulações estatísticas mais elaboradas permitiu explorar textos dos mais variados domínios dando origem à metodologia baseada em corpus. Foram caracterizadas as diversas formas de transformação de uma dada entrada para a produção de extractos. Assim, a partir da arquitectura geral apresentada na figura 2.1, caracterizou-se a abordagem empírica como um processo de manipulação numérica/estatística de informação, ilustrado na figura 2.2.

2.2 Abordagem superficial

Nesta secção, serão apresentados exemplos de alguns métodos ou sistemas associados à abordagem superficial ou estatística. Serão inicialmente apresentados,

[3]As *cue words* ou palavras de pista, numa passagem a ser resumida, são as palavras ou frases que dão indicações importantes do conteúdo da referida passagem como um todo. Num paragrafo, a frase de assunto deverá conter *cue words* dado que esta frase geralmente indica o assunto do paragrafo. É muito difícil definir uma lista de *cue words* visto que estas dependem do conteúdo de uma passagem particular e, em geral, do tipo/assunto de texto que está a ser tratado.

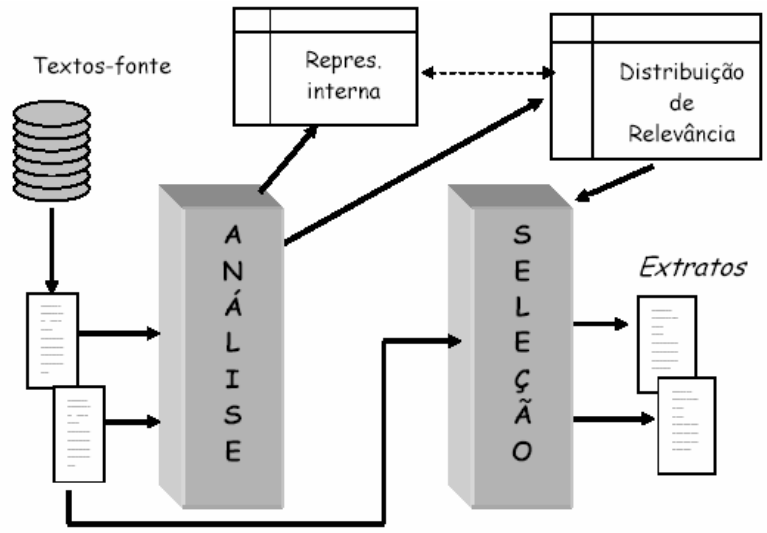

Figura 2.2: Arquitectura de um sistema de sumarização automática superficial (Rino e Pardo, 2003).

de uma forma mais detalhada, alguns dos métodos referidos na secção 2.1. Seguidamente serão discutidos os novos métodos criados na década de 90, métodos que surgiram com a utilização de algoritmos de aprendizagem máquina aplicados ao processamento de língua natural.

2.2.1 Método das palavras-chave

Este método parte do pressuposto que as ideias principais de um texto podem ser expressas por palavras-chave. É determinada a distribuição estatística das palavras-chave do texto original, as que possuam frequências no documento superiores a um dado valor mínimo são utilizadas para a criação de uma lista. Nesta lista, as palavras-chave são armazenadas com os seus pesos (os pesos foram calculados a partir da sua no texto original). A pontuação total de uma frase é a soma dos pesos de cada uma das palavras-chave que a constitui. As frases com maior peso e acima de um dado valor limite são extraídas e agrupadas de forma a constituir um sumário, sendo apresentadas na ordem em que aparecem originalmente no texto-fonte.

O programa desenvolvido por (Luhn, 1958) é baseado nestes mesmos princípios. De entre as palavras-chave encontradas, o programa considera somente aquelas de

classe aberta, ou seja, as que carregam significado[4].

(Earl, 1970) apresentou uma variação deste método, para ele os substantivos mais frequentes de um dado texto são responsáveis pela maioria das palavras-chave. Ele privilegiou os substantivos em relação aos verbos dado que os primeiros poderiam indicar a progressão temática do texto. Por exemplo, na frase "O jogador morreu", a palavra "jogador" seria mais representativa do que o verbo "morreu" porque indicaria o assunto sobre o qual se fala.

2.2.2 Método das palavras-chave do título

Neste método são utilizadas características estruturais do texto, tais como o título, os cabeçalhos e a formatação (Edmundson, 1969). Para aplicar este método é necessário assumir que os títulos, quer o principal quer os das secções foram correctamente formulados e que representam o conteúdo do texto (o que geralmente é verdade quanto temos documentos com bons títulos). Considera-se, como hipótese principal, que as palavras que compõem o título e os cabeçalhos das secções terão maior probabilidade de serem representativas do conteúdo do texto correspondente. É criado um glossário de termos contendo todas as palavras com significância presentes nas estruturas textuais referidas anteriormente, para cada uma destas é calculado o seu peso. De seguida, para cada frase é calculada a sua importância somando os pesos às palavras que a compõem e estão no glossário. Por fim, as frases com maior peso são extraídas para a geração do sumário.

2.2.3 Método das palavras sinalizadoras (*cue phrases*)

Neste método são utilizadas pistas do texto a sumarizar. frases em que ocorrem palavras como "significante","impossível" ou ''dificilmente" recebem um maior peso (Edmundson, 1969). É criado um dicionário, constituído por três sub-dicionários, contendo as palavras consideradas relevantes ao domínio do texto:

- *Bonus words*: palavras que pontuam positivamente a relevância das frases;
- *Stigma words*: palavras cuja ocorrência pontua de forma negativa (penaliza) a relevância das frases;
- *Null words*: palavras que não influenciam a medida de relevância das frases.

O valor (*cue weight*) final de cada frase é a soma dos pesos das suas palavras.

[4]As palavras cuja categoria lexical pertence ao conjunto de substantivos, verbos, advérbios e adjectivos.

Num texto científico, as palavras "conclusões" e "resultados" provavelmente serão altamente significativas, estando presentes no dicionário. A sua ocorrência numa dada frase implicará o aumento de relevância da mesma. Como facilmente se percebe pelo exemplo anterior, géneros distintos têm necessariamente dicionários distintos e outros marcadores da importância do conteúdo textual.

É importante referir que embora este método se assemelhe ao método das palavras-chave, eles diferem porque o dicionário, neste método, é composto por palavras consideradas significativas extraídas de outros textos e não do texto-fonte.

2.2.4 Método relacional

(Skorochod'ko, 1972) refere a dificuldade da utilização de uma única estratégia para lidar com os vários tipos de textos a sumarizar. Segundo o autor, para ser possível obter bons resultados, independentemente do tipo de texto, o método de sumarização deve variar de acordo com a estrutura do texto (com a organização das secções, subsecções e o fluxo das ideias). Ele descreve uma estratégia adaptativa que utiliza as relações entre frases. Utiliza uma representação gráfica do texto, cujas relações entre frases são criadas a partir das relações semânticas entre as palavras que as compõem. frases com maior número de outras frases semanticamente relacionadas recebem um maior peso tornando-se candidatas mais prováveis para extracção e formação do sumário final. A relação semântica entre frases é identificada, por exemplo, pela ocorrência de nomes comuns.

2.2.5 Método da frase auto-indicativa

Esta estratégia que foi descrita por (Paice, 1981), usa a indicação explicita da importância da frase a ser seleccionada para a criação de um sumário. Uma frase deste tipo apresenta uma estrutura cuja ocorrência é frequente no texto e indica explicitamente que a frase se refere a algo importante sobre o assunto do texto. Exemplos deste tipo de frases são: "O objectivo deste artigo é investigar..." e "Neste artigo, é descrito um método para...". Este método apenas permite a produção de sumários indicativos, isto é, aqueles que ajudam a identificar o assunto do texto sem, no entanto, apresentar uma discussão sobre o mesmo.

2.2.6 TF-IDF

A métrica TF-IDF (*Text Frequency-Inverse Document Frequency*) é uma medida estatística, baseada na frequência de termos. A sua utilização para a sumarização

de textos foi apresentada pela primeira vez por (Rau e Brandow, 1994). Este método parte do princípio que uma palavra será representativa num texto se ocorrer diversas vezes no texto em questão e for pouco frequente noutros textos. As palavras cujo peso indica que carregam informações importantes são separadas numa lista (lista de *signature words*). São ainda adicionadas a esta lista palavras do título, estas palavras ainda que pouco frequentes considere-se que carregam muita informação. De seguida, o peso das frases é calculado a partir do peso das palavras que as compõem. Ou seja, o peso resultante será a soma dos pesos das palavras da frase que também estiverem presentes na lista de *signature words*.

2.2.7 Métodos Bayes-ingénuo

(Kupiec, Pedersen e Chen, 1995) descreveram um método derivado de (Edmundson, 1969) que é capaz de "aprender" com os dados. A função de classificação categoriza cada frase como merecedora ou não de ser extraída usando um classificador Bayes-ingénuo. Seja s uma frase, S o conjunto de frases que perfazem o sumário e $F_1, ..., F_k$ as características relevantes para o sumário. Assumindo que as características são independentes temos:

$$P(s \in S|F_1, F_2, .., F_k) = \frac{\prod_{i=1}^{k} P(F_i|s \in S).P(s \in S)}{\prod_{i=1}^{k} P(F_i)}$$

As características eram compatíveis com as de (Edmundson, 1969) incluindo adicionalmente o comprimento da frase e a presença de palavras em maiúsculas. A cada frase era dada uma pontuação de acordo com a função supra-referida e apenas as frases com a maior pontuação eram extraídas.

2.2.8 Redes Neuronais

(Svore, 2007) propôs um algoritmo baseado em redes neuronais. Na sua abordagem treinou um modelo utilizando um conjunto de características das várias frases dos artigos (foram utilizados artigos da CNN). Este modelo conseguia inferir a ordenação (por importância) de cada frase no documento. Para a atribuição de pontuação era utilizando o algoritmo RankNet (Burges et al., 2005). Uma das características inovadoras foi a utilização dos registos de pesquisas do motor de pesquisa de notícias da Microsoft e entidades da Wikipédia. Para o autor, se uma frase continha palavras-chave utilizadas na pesquisa ou entidades presentes na Wikipédia haveria uma maior probabilidade dessa frase ser relevante.

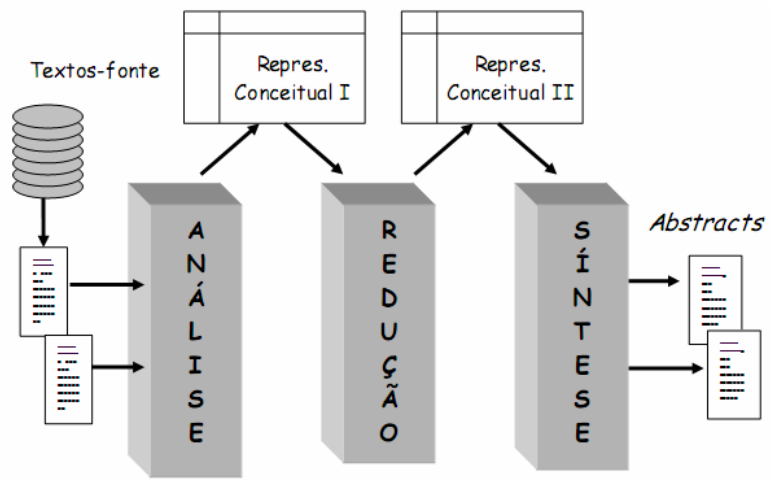

Figura 2.3: Arquitectura de um sistema de sumarização automática profunda (Rino e Pardo, 2003).

2.3 Abordagem profunda

Os métodos da abordagem superficial citados na secção anterior sofrem de variadas limitações. Estas limitações são impostas pela própria definição da abordagem superficial, estes métodos estão limitados à selecção e extracção de segmentos do texto original. Em oposição, temos a abordagem profunda que contempla o conhecimento linguístico associado ao texto original para compor o seus(s) possível(is) sumário(s). O conhecimento envolve, por exemplo, as relações semânticas e retóricas.

As principais dificuldades na abordagem profunda relacionam-se com a forma como é identificado e resumido o conteúdo relevante de um dado texto. A arquitectura de um sumarizador automático profundo foi apresentada por (Rino e Pardo, 2003), nela é sugerida uma abordagem que simula o processo humano de sumarizar. Contempla a compreensão do texto-fonte, condensação do conteúdo e reescrita textual. A arquitectura deste sistema pode ser observada na figura 2.3.

(Barzilay e Elhadad, 1997) descrevem um trabalho que utiliza a análise linguística para a realização da tarefa de sumarização. Para melhor compreender o seu método é necessário definir o conceito de cadeia lexical, a qual é considerada uma sequência de palavras relacionadas num texto.

No seu método é necessário segmentar o texto, identificar as cadeias lexicais e por fim usar as cadeias lexicais fortes[5] para identificar as frases a serem extraídas. O autores tentaram chegar a um meio termo entre (McKeown e Radev, 1995) e (Luhn, 1958). O primeiro utiliza a estrutura semântica do texto enquanto o segundo usa a estatística das palavras que compõem o texto. Neste trabalho os autores descrevem ainda a noção de coesão no texto como meio de relacionar as suas diferentes partes. Um exemplo é a coesão lexical, que utiliza palavras semanticamente relacionadas. Como se pode ver na frase seguinte, a palavra carro refere-se à palavra Jaguar:

O João comprou um Jaguar. Ele adora o carro.

Este fenómeno de coesão ocorre não só ao nível das palavras mas também com grupos de palavras, resultando em cadeias lexicais. Neste método após a identificação de palavras e grupos de palavras semanticamente relacionadas, são extraídas algumas cadeias que formam uma representação do documento. As cadeias extraídas são ordenadas por tamanho e homogeneidade. Por fim, são seleccionadas as que mais se adequam ao sumário utilizando um conjunto de heurísticas.

No trabalho (Ono, Sumita e Miike, 1994), os autores apresentam um modelo computacional de discurso para textos escritos em japonês, onde elaboram um procedimento prático para extrair a estrutura retórica do discurso. Os autores extraem uma árvore binária que representa as relações entre partes de frases (árvores de estrutura retórica são muito usadas em (Marcu, 1998), como veremos a seguir). Esta estrutura é extraída através de uma série de etapas: análise de frases, extracção de relações retóricas, segmentação, geração de candidatos e selecção dos candidatos mais adequados utilizando a importância relativa das relações retóricas. Na etapa seguinte, alguns nós da árvore da estrutura retórica são removidos para reduzir a frase, mantendo as suas partes importantes. Finalmente, o mesmo é feito para os vários parágrafos a fim de produzir o resumo.

(Marcu, 1998) descreve uma abordagem única para o resumo que, ao contrário da maioria dos trabalhos anteriores, não assume que as frases num texto têm uma estrutura plana. Utiliza uma teoria de discurso denominada teoria da estrutura retórica (ver secção 2.5) a qual incide em dois excertos não sobrepostos de texto: o

[5]cadeias lexicais fortes são definidas por (Barzilay e Elhadad, 1997) como duas palavras conectadas por uma relação (*WordNet*).

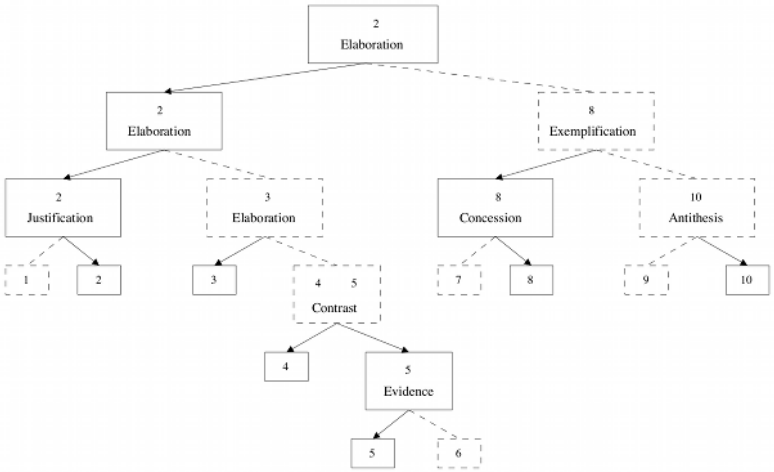

Figura 2.4: Exemplo de uma árvore discursiva de (Marcu, 1998). Cada um dos nós recebe um número, que equivalem aos números das frases no texto Os nós, na linha tracejada são satélites e na linha cheia são os núcleos.

núcleo e o satélite. De acordo com o autor, a distinção entre os núcleos e os satélites advém da observação empírica que o núcleo expressa melhor o que é importante para o escritor do que o satélite, sendo que o núcleo de uma relação retórica é compreensível sem o satélite mas o contrário já não é verdade. (Marcu, 1997) descreve como um analisador retórico produz uma árvore discursiva, na figura 2.4 pode ser observado um exemplo.

Uma vez criada a estrutura discursiva, pode ser derivada uma ordenação parcial de unidades com maior importância. Se for definido que o resumo deve conter $k\%$ do texto, as primeiras $k\%$ unidades da árvore são seleccionadas. Este método é descrito em (Marcu, 1997).

2.4 Língua Portuguesa

Nas secções 2.1, 2.2 e 2.3 foi feita uma contextualização histórica e apresentados vários métodos para a sumarização automática. Grande parte da pesquisa existente nesta área foca-se na língua inglesa, nesta secção serão apresentados alguns dos projectos focados na língua portuguesa. Todos os sistemas apresentados estão orientados para o Português do Brasil dado que não existem trabalhos significativos para Português Europeu.

2.4.1 TF-ISF-Summ

(Neto et al., 2000) desenvolveu uma ferramenta de *text mining*[6] composta por dois módulos (ver figura 2.5). O primeiro módulo calcula *clusters* dos documentos e gera uma lista de palavras-chave (por *cluster*) que os autores denominam de sumário ultra compacto. O segundo módulo tem uma relação directa com as técnicas de sumarização abortadas nesta tese. Faz uso da métrica TF-ISF (*Term-Frequency Inverse-Sentence-Frequency*) para classificar as frases de um dado texto e efectuar a extracção das mais relevantes (frases que têm a pontuação TF-ISF mais elevada). A TF-ISF é uma variação da medida estatística TF-IDF apresentada na secção 2.2.6 definida da seguinte forma: Seja $TF(w, s)$ a frequência da palavra w na frase s, isto é, o número de vezes que a palavra w ocorre na frase s. Quanto mais elevado o valor de $TF(w, s)$ mas representativa é a palavra na frase s. Seja $SF(w)$ a frequência da palavra w nas frases do texto e S o número de frases no documento, a frequência inversa da palavra é definida como: $ISF(w) = log(\frac{|S|}{SF(w)})$. Assim, a TF-ISF de uma palavra w numa dada frase s é dada por:

$$TF - ISF(w, s) = TF(w, s) * ISF(w)$$

Para a criação do sumário são necessárias várias fases:

1. pré-processamento do texto - todas palavras são convertidas para minúsculas e transformadas para a sua forma radical, são ainda removidas as *stop words*;
2. segmentação do texto em frases;
3. criação o vector com os pesos TF-ISF;
4. calcular o TF-ISF médio.

Finalmente, as frases com a pontuação TF-ISF média mais elevada e acima de um certo valor limite (definido pelo utilizador) são seleccionadas para compor o extracto final.

Os autores não fizeram uma avaliação profunda do seu sistema. Foi, no entanto, realizada uma avaliação subjectiva utilizando um sistema proveniente da conferência SUMMAC[7]. Da comparação subjectiva dos sumários produzidos pelo

[6] *Text mining* refere-se ao processo de obtenção de informação de qualidade a partir de texto em linguagem natural. É inspirado na mineração de dados, que consiste em extrair informação de bancos de dados estruturados; a mineração de texto extrai informação de dados não estruturados ou semi-estruturados.

[7] SUMMAC - TIPSTER Text Summarization Evaluation Conference

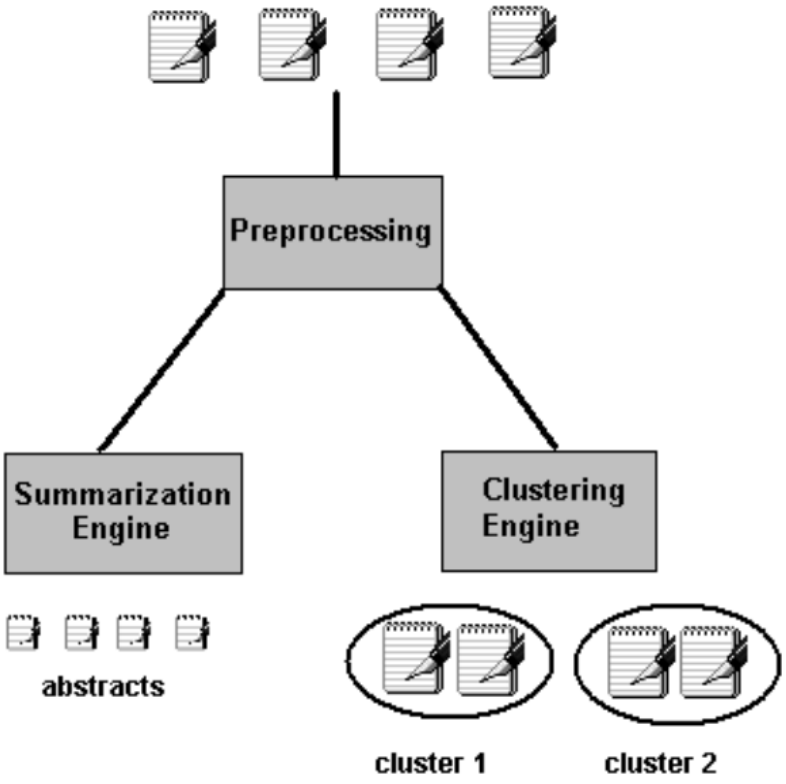

Figura 2.5: Arquitectura do sistema TF-ISF-Summ (Neto et al., 2000).

sistema TF-ISF-Summ e os sumários obtidos pelo sistema CGI/CMU (o sistema que obteve melhores resultados na tarefa *adhoc* do SUMMAC) os autores referem que "ambos os sistemas produzem resultados de grande qualidade, ambos conseguindo captar as ideias base do texto original".

2.4.2 GistSumm

O GistSumm (Pardo, 2002c) é um sistema de sumarização baseado método *gist*. O *gist* é a ideia principal pretendida pelo autor ou compreendida pelo leitor. Usando métodos estatísticos, o *gist* é identificado como a frase mais importante no texto original. De seguida, serve para identificar outras frases que irão fazer parte do extracto final. Para o método funcionar são assumidas as seguintes premissas:

- o texto é construído considerando de uma ideia principal *gist*;
- é possível identificar no texto apenas uma frase que expressa a ideia principal.

Baseado nas hipóteses apresentadas, o GistSumm identifica a frase *gist* utilizando estatísticas simples e por fim construí extractos coerentes com a frase *gist*.

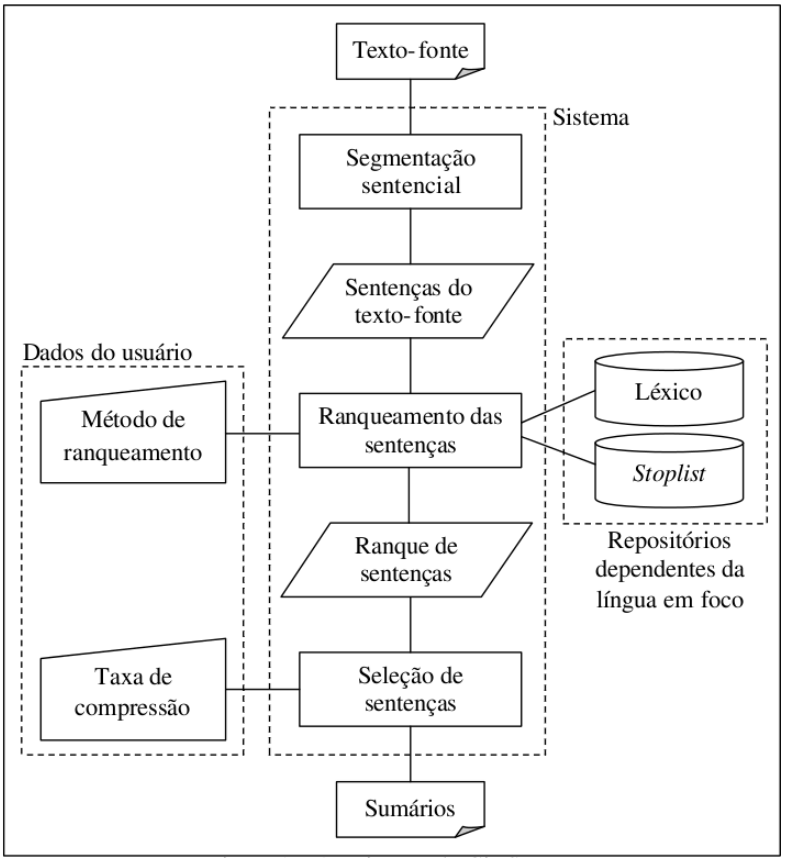

Figura 2.6: Arquitectura do sistema GistSumm (Pardo, 2002c).

O GistSumm é composto por três processos:

1. segmentação de texto - esta fase contêm também as etapas de conversão das palavras para minúsculas, a transformação para a sua forma radical e a remoção das *stop words*;
2. classificação das frases;
3. produção de extractos.

O GistSumm utiliza duas estratégias de classificação:

1. método das palavras-chave em que a pontuação de cada frase é a soma das frequências das palavras que a constituem;

2. métrica TF-ISF (apresentada na secção anterior).

A frase com a pontuação mais elevada será seleccionada como a frase *gist*. A produção de extractos é baseada na correlação das frases com a frase *gist* sendo o critério a coocorrência de palavras da frase *gist* nas frases candidatas. Para a produção do extracto final são seleccionadas todas as frases com uma avaliação acima de um dado valor limite.

Foi feita um avaliação utilizando um corpus com 20 artigos de jornal. Para cada texto foram gerados dois extractos (um para cada método de geração do *gist*), de seguida, foram apresentados a 12 juízes humanos para atribuir pontuações à preservação do *gist* e coerência textual. Foi observado que o método das palavras-chave supera o TF-ISF, quando comparados os extractos com os seus textos originais, tanto na coerência textual como na geração de *gist*. Este sistema foi ainda submetido a várias avaliações externas, entre as quais a DUC'2003 onde mostrou resultados acima da média.

2.4.3 NeuralSumm

O NeuralSumm (Pardo, Rino e Nunes, 2003) utiliza uma abordagem totalmente nova para a classificação, as redes neuronais. Este sistema é executado em quatro fases: segmentação de textos, extracção de características, classificação e produção de extractos. É baseado numa rede neuronal do tipo SOM[8] que é estilizada para classificar cada frase de um texto de acordo com o seu grau de importância. As frases podem ser classificadas como essenciais, complementares ou supérfluas. frases essenciais são aquelas que transmitem a ideia principal de um texto; frases complementares são as que acrescentam conteúdo à ideia principal, complementando-a; frases supérfluas, por sua vez, não acrescentam conteúdo algum. Uma vez classificadas as frases, o extracto correspondente é produzido utilizando a classificação obtida pelas frases, a taxa de compressão desejada para o extracto e a pontuação das frases dada pela distribuição de palavras do texto utilizando a medida TF-ISF. A pontuação das frases é utilizada apenas quando há empates na classificação proveniente da rede neuronal.

A avaliação deste sistema teve como primeiro objectivo medir o desempenho de uma rede neuronal do tipo SOM a classificar frases em essenciais, complementares e supérfluas e como segundo objectivo verificar a proximidade dos extractos gerados

[8]a rede utilizada possui uma arquitectura de 14x14 neurónios e foi treinada com a precisão de 1 milionésimo

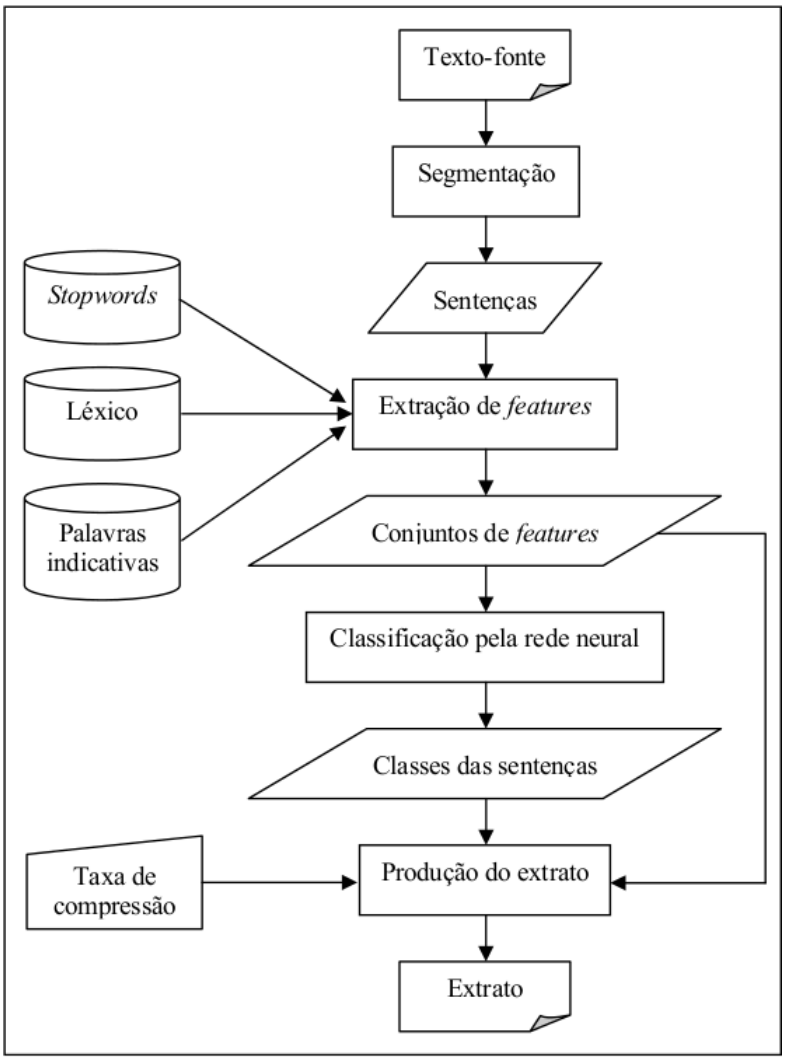

Figura 2.7: Arquitectura do sistema NeuralSumm (Pardo, Rino e Nunes, 2003).

automaticamente com os sumários autênticos. Relativamente ao primeiro objectivo os autores referem que a rede neuronal obteve menor taxa de erro em relação aos outros classificadores (bayseano, regras de decisão e árvores de decisão). Para o segundo objectivo foram utilizados 10 textos científicos (introduções de teses e dissertações) com os respectivos sumários autênticos, para os quais foram gerados automaticamente sumários ideias. Efectuada a comparação dos sumários ideais com o sumário gerado pelo NeuralSumm com uma taxa de compressão de 80% obtiveram-se os resultados 32% e 41% para a cobertura e precisão respectivamente.

2.4.4 ClassSumm

O ClassSumm (Neto, Freitas e Kaestner, 2002) utiliza uma abordagem de aprendizagem máquina para determinar segmentos relevantes para a produção de extractos de textos-fonte. Este sistema faz uso de dois módulos de classificação, um baseado numa versão ingénua de algoritmo de Bayes e outro no C4.5 (Quinlan, 1993). Tal como o TF-ISF-Summ (também do mesmo autor) este sistema é composto por várias etapas de processamento até gerar o sumário:

- pré-processamento do texto - todas palavras são convertidas para minúsculas e transformadas para a sua forma radical, são ainda removidas as *stop words*;
- são calculados os valores para um conjunto de características - TF-ISF média, comprimento da frase, posição da frase, semelhança com o título, etc. de cada frase;
- as frases são classificadas como extraíveis ou não - para cada frase o algoritmo deve "aprender" quais são relevantes para o sumário;
- são seleccionadas todas as frases acima de um valor limite definido pelo utilizador, isto é, as que têm mais probabilidade de pertencer ao extracto.

A avaliação foi feita utilizando documentos da colecção TIPSTER. Após o treino do sistema foram gerados sumários para taxas de compressão de 10% e 20%, foi observado pelos autores que a precisão e cobertura são significativamente mais altos com a taxa de 20% do que com a de 10%. Os melhores resultados foram obtidos pelo sistema quando este foi treinado utilizando o método bayes (para qualquer das taxas de compressão).

2.4.5 DMSumm

Todos os sistema de sumarização apresentados até aqui utilizam a abordagem superficial. Ao contrário, o DMSumm (Pardo, 2002a) utiliza a abordagem profunda para a resolução do problema da sumarização. Tem como objectivo a produção de sumários a partir de uma mensagem que corresponde à interpretação do texto que se pretende sumarizar. Segue a arquitectura clássica de geração automática de sumários de três passos (ver figura 2.8), isto é, a selecção de conteúdo, o planeamento textual e a realização linguística, sendo que o planeamento textual é, de facto, a implementação do modelo de discurso utilizado.

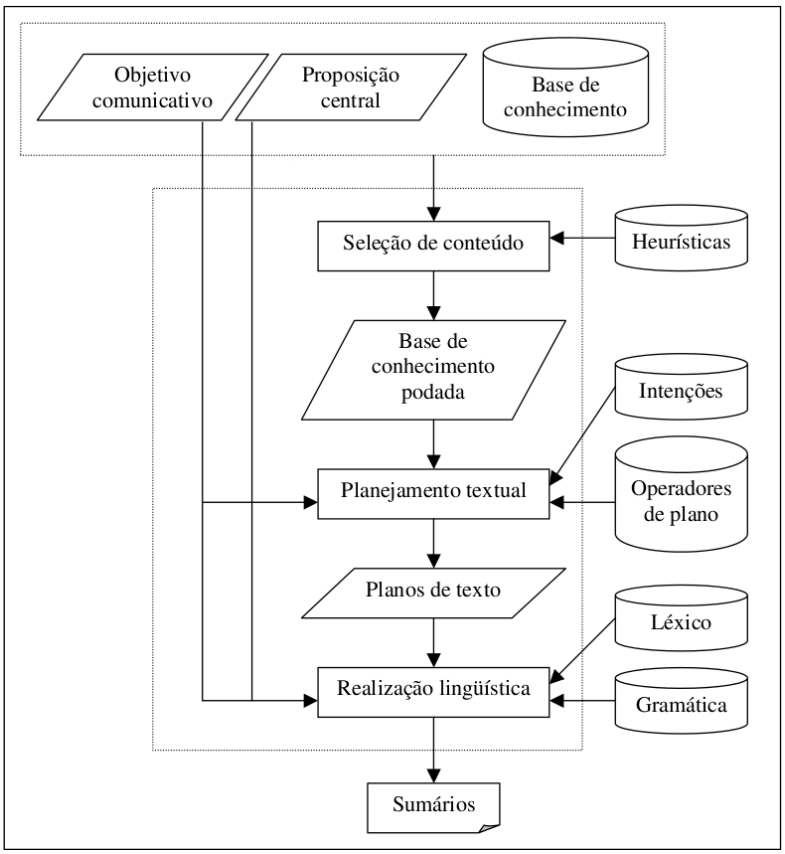

Figura 2.8: Arquitectura do sistema DMSumm (Pardo, 2002a).

O DMSumm propõe a implementação do modelo discursivo de (Rino, 1996), convém, no entanto, referir que o sistema não é um sumarizador automático de textos, propriamente dito, pois não parte de um texto-fonte para gerar seu sumário, mas sim de uma representação interna resultante da interpretação do texto que se quer sumarizar denominada mensagem-fonte. Este sistema é na realidade um gerador automático de sumários, já que não possui o processo de interpretação automática do texto-fonte na mensagem-fonte. Devido à necessidade de criação de uma mensagem-fonte (fase de pré-edição) é necessário um especialista para operar este sistema, ele tem de conhecer perfeitamente os modelos semânticos de forma a poder estruturar a base de conhecimento.

O DMSumm pretende, ao gerar o texto final, preservar a proposição central e satisfazer do objectivo comunicativo. A proposição central refere-se ao que se quer comunicar com o discurso, enquanto o objectivo comunicativo representa a

própria motivação para a existência de qualquer discurso. A entrada do sistema é a mensagem-fonte que é constituída pela proposição central, o objectivo comunicativo e uma base de conhecimento referente ao conteúdo do texto-fonte. O objectivo comunicativo é responsável por seleccionar os componentes textuais que se relacionarão à proposição central nos sumários e a base de conhecimento fornece informação, isto é, conhecimento para ser manipulado durante a sumarização.

A selecção de conteúdo recebe como entrada a mensagem-fonte, tendo a função de reduzir o conteúdo informativo disponível para a produção dos sumários. Esse processo é composto de duas tarefas:

1. podar a base de conhecimento por meio de heurísticas;

2. reproduzir tanto o objectivo comunicativo quanto a proposição central na base de conhecimento podada

A saída deste processo, a base de conhecimento podada, o objectivo comunicativo e a proposição central originais constituem os dados de entrada para o planeamento textual, denominados mensagem-fonte do sumário. O planeamento textual, cuja função é estruturar o discurso, recebe como entrada a mensagem-fonte do sumário. A partir desta, são construídos os planos de texto (estruturas retóricas) de possíveis sumários com base no modelo de discurso de (Rino, 1996), mapeando relações semânticas (da base de conhecimento) e intencionais em relações retóricas. O processo de realização linguística produz os sumários propriamente ditos a partir dos planos de texto, expressando estes últimos em língua natural pela aplicação de *templates*.

Dado que há interacção de um especialista no processo podem existir diferentes interpretações para um texto, a mensagem-fonte pode variar, dando origem a diferentes sumários.

Foram feitas várias experiências para avaliar o DMSumm, focando, principalmente, a satisfação do objectivo comunicativo e a preservação da proposição central. Outros critérios, como coerência textual, também foram considerados. Foi utilizado um corpus composto por 10 introduções de teses (mestrado e doutoramento) na área das ciências da computação. Os sumários automáticos foram julgados por 10 juízes linguistas computacionais e falantes nativos do português do Brasil. De acordo com os julgamentos 62% dos sumários foram considerados satisfatórios, 22% foram considerados aceitáveis e 16% foram considerados maus. Os autores observaram ainda que 67% dos sumários automáticos mantiveram a

coerência textual e 61% dos sumários automáticos preservaram somente parcialmente a ideia principal e 31% a preservaram totalmente. O DMSumm produziu sumários com 44% de precisão e 54% de cobertura, com uma *f-measure* de 0,48.

2.4.6 SuPor

O sistema SuPor (Rino e Módolo, 2004) é composto por dois módulos: o módulo de treino (figura 2.9) e o módulo de extracção (figura 2.10). Durante o treino é atribuído um peso cada uma das características do texto medindo a sua representatividade nos textos de origem e nos seus extractos ideais. Após o treino, um especialista poderá personalizar o SuPor para sumarizar qualquer texto-fonte utilizando vários parâmetros: o tipo de pré-processamento (remover *stop words*, reduzir as palavras à sua forma radical, transformar em *n-grams*), seleccionar o conjunto de características e o classificador que irá ser utilizado na sumarização, por fim terá de definir a taxa de compressão pretendida.

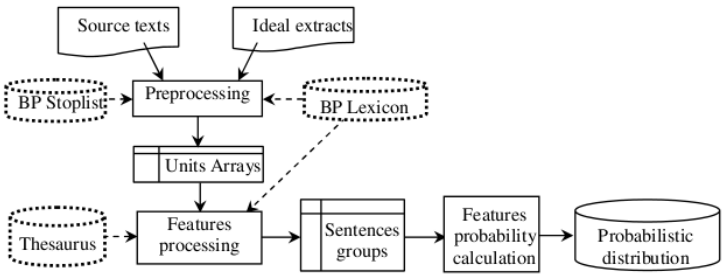

Figura 2.9: Arquitectura do sistema SuPor - Módulo de Treino (Rino e Módolo, 2004).

Este sistema implementa quatro métodos extractivos que foram desenvolvidos originalmente para a língua inglesa:

- classificador - utiliza um classificador bayseano para treinar o sistema de forma a reconhecer as características mais importantes, as características incluem o tamanho da frase, a frequência das palavras, a localização da frase ou paragrafo, a ocorrência de nomes próprios entre outras;

- cadeias lexicais - calcular a coesão textual entre as várias correlações entre as palavras considerando apenas os substantivos como as unidades básicas

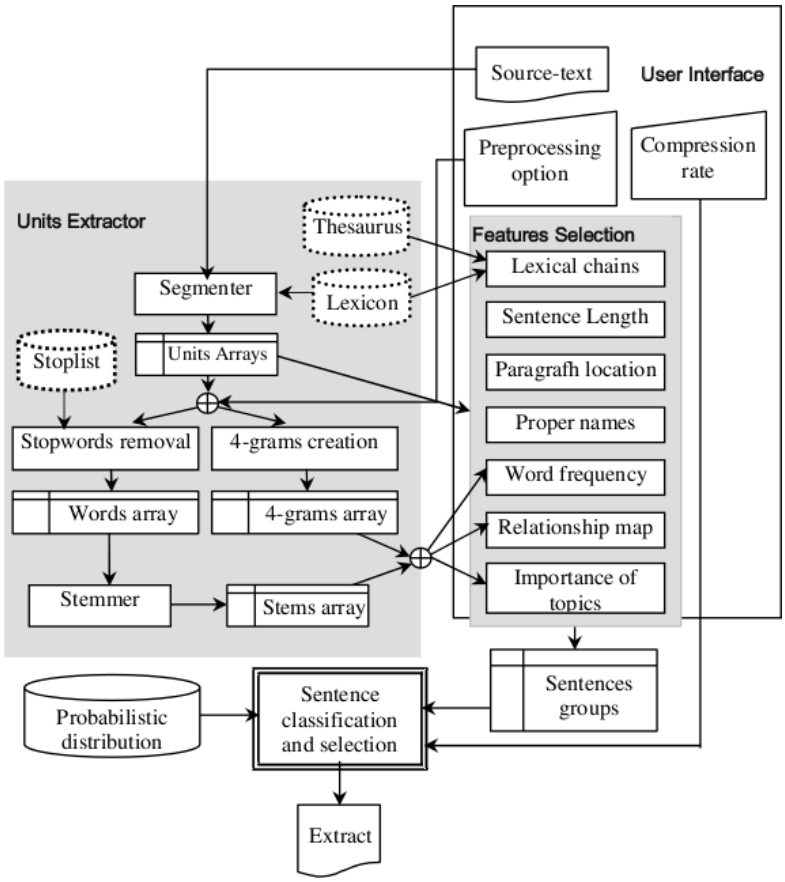

Figura 2.10: Arquitectura do sistema SuPor - Módulo de Extracção (Rino e Módolo, 2004).

de significação no texto. As relações lexicais mais fortes são aquelas que a relação semântica é mais expressiva;

- mapa de relações - define três métodos para interligar parágrafos construindo mapas obtendo caminhos densos, profundos ou segmentados. Os caminhos densos são os que têm mais ligações no mapa, os profundos focam-se em parágrafos que estão semanticamente relacionados;

- importância dos assuntos - é baseado na medida TF-ISF que identifica frases que transmitem informação mais relevante a incluir no extracto.

Para testar este sistema os autores utilizaram o corpus TeMário. Foi feita uma validação cruzada sendo definida uma taxa de compressão 30% para a geração dos

sumários. Foram calculadas a precisão, cobertura e *f-measure* para cada uma das execuções da validação cruzada. A média final obtida, para a comparação dos sumários automática com os extractos ideias do corpus usado, foi 44.9%, 40.8% e 42.8 para a precisão, cobertura e *f-measure* respectivamente.

2.5 Teoria da Estrutura Retórica e o Sistema AuTema-Dis

Um trabalho que marcou o desenvolvimento deste sistema foi o de (Leal, 2008), que apresenta uma arquitectura computacional para identificação da temática discursiva em textos em língua portuguesa. É ainda necessário fazer uma breve apresentação da teoria da estrutura retórica, visto que o trabalho de (Leal, 2008) fez uso desta teoria para a definição das regras que são utilizadas no processo de sumarização. Nesta secção é apresentada inicialmente a Teoria da Estrutura Retórica e por fim o Sistema AuTema-Dis.

2.5.1 Teoria da Estrutura Retórica

A teoria RST foi criada para suprir a carência de um formalismo teórico em relação aos estudos na área de geração automática de texto. Uma equipa no Instituto de Ciências de Informação (Universidade da Califórnia do Sul) que desenvolvia investigação na área da geração automática de texto, apercebeu-se da necessidade de uma teoria que definisse a estrutura de um discurso e que fosse suficientemente detalhada para a programação de um sistema de geração automática de texto.

Na busca para descrever e explicar os processos que organizam a informação de um discurso, os autores, William C. Mann e Sandra Thompson, propuseram a RST como uma teoria que define a organização de um texto. Identificam-no como uma unidade sem ausência de lacunas e de conjuntos aleatórios de frases. Nesse trabalho explicam a coerência dos textos e não os processos que levam à sua criação ou interpretação. Segundo os investigadores, para toda a parte de um texto coerente existe uma função (uma razão plausível para a sua presença) e ela deve ser evidente para o leitor, isto é, o leitor não deve ter a sensação que o texto está incompleto.

Este teoria é descritiva, a partir da organização dos discursos, demonstra as relações que se estabelecem entre as suas estruturas em termos funcionais. Com o objectivo de melhor explicar os procedimentos textuais e a organização discur-

siva, os autores apresentaram inicialmente 25 relações retóricas, no entanto, com a evolução da pesquisa o número destas relações foi ampliado para 32.

De acordo com (Mann e Thompson, 1988), a avaliação destas relações põe em evidência a base funcional da hierarquia textual, o que torna possível explicar a organização textual, bem como, a coerência de um texto. Para tal, a teoria RST define um conjunto de possibilidades que identificam e organizam em categorias padrões estruturais. As relações retóricas cumprem as formalidades textuais básicas, para que o objectivo proposto pelo autor seja reconhecido pelo leitor.

A ideia central na RST é a noção de relação retórica que se estabelece entre duas proposições discursivas (*spans*). Estas proposições desempenham diferentes papéis ou funções, ou seja, um é o núcleo (N) e outro o satélite (S). O núcleo representa o segmento mais significativo da relação, enquanto que o satélite representa um conteúdo adicional em relação ao núcleo. Normalmente, as relações retóricas são construídas entre os pares de proposições (*spans*) com um núcleo e um satélite, sendo neste caso denominadas de relações nucleares. Na figura 2.11 são apresentadas estruturas que mostram os núcleos e segmentos, podemos observar uma relação composta por apenas um núcleo.

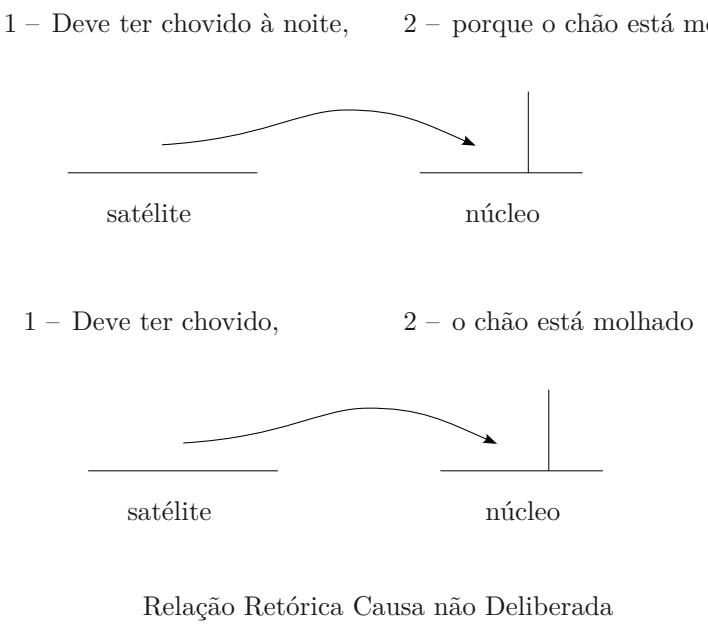

Figura 2.11: Exemplo de uma relação retórica nuclear (causa involuntária). Os arcos com setas apontam em direcção do núcleo e a linha vertical representa o ponto nuclear na proposição.

Podemos ainda identificar uma relação do tipo multi-nuclear, esta é construída entre *spans* do mesmo valor (núcleos). Pode ser observado na figura 2.12 a relação retórica multi-nuclear *sequence* (sequência).

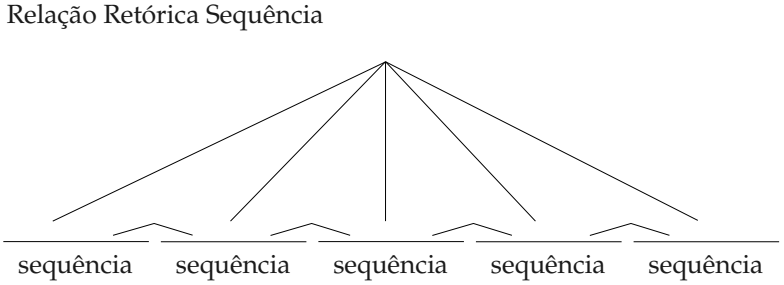

Figura 2.12: Relação retórica multi-nuclear (sequência).

A identificação dos pares de *spans*, bem como, a atribuição dos papéis que desempenham está condicionada ao valor significativo que ambos ocupam no segmento discursivo do qual fazem parte. A determinação dos papéis/funções está sujeita às características especificas da pessoa que executa a análise, ou seja, do analista.

Para além das proposições (*spans*) núcleos e satélite esta teoria envolve outros elementos que participam na identificação das relações, são eles: *writer* - escritor (W) - quem produz o texto; *reader* - reader (R) - quem recebe e interpreta o texto; e o *analyst* - analista - quem realiza a análise do texto. Dependendo do tipo de análise que se pretende realizar, o analista selecciona um dos elementos (*writer*, *reader*, *analyst*), a fim de direccionar o tipo de análise. Sendo o analista a figura responsável pelos julgamentos a respeito da composição da análise, é ele que a condiciona de acordo com as suas necessidades e prioridades em relação ao tipo de investigação que pretende realizar.

As relações retóricas são organizadas em dois grupos específicos de acordo com as suas características: *subject matter* - relações que apresentam parte do conteúdo do texto e *presentational* - relações que são utilizadas para auxiliar na apresentação. A classificação num ou no outro grupo está sempre subordinada à interpretação do analista ou do leitor e ao efeito que uma determinada relação lhe causou.

Apesar de ser possível utilizar as informações sobre a posição das proposições N e S num texto, com a finalidade de restringir a ocorrência de uma relação, (Mann e Thompson, 1988) relembram que podem ser produzidas múltiplas análises, este facto é conhecido como a ambiguidade da análise retórica. Esta ambiguidade

Elemento de definição	**Ponto de Vista do Observador**
Restrições no Núcleo N	Nenhuma
Restrições no Satélite S	S apresenta uma situação hipotética, futura ou não realizada (relativa ao contexto da situação).
Restrições na contribuição N + S	A realização da situação apresentada no N depende da realização do que foi apresentado em S.
O Efeito pretendido pelo leitor em usar a relação endereçada ao leitor L nunca é nulo.	O Leitor reconhece que a realização da situação apresentada no N depende da realização da situação apresentada no S.
Local do Efeito de onde o efeito é derivado.	N e S

Tabela 2.1: Exemplo de uma das relações apresentadas por (Mann e Thompson, 1988) com as restrições por área. Trata-se de um exemplo representativo da Relação Retórica de Condição.

retórica é anterior ao reconhecimento da posição de ocorrência das proposições e anterior à própria análise retórica, é uma ambiguidade que deriva da língua natural, que na sua concretização, é ambígua.

A RST apresenta-se como uma base descritiva para estudar as relações entre as frases de um texto em termos funcionais, utilizando a distinção entre as proposições núcleo e satélite e a sua hierarquia. É um sistema que apresenta uma combinação de características capaz de representar a estrutura hierárquica e o princípio central da organização de um texto. É de referir que o núcleo tem muito mais importância que o satélite, como se pode observar com observação empírica de algumas experiências simples: se for apagado do núcleo numa relação o conteúdo do satélite tornar-se-á muitas vezes incompreensível; se ao contrário, for removido um satélite é ainda possível identificar a informação contida na estrutura a partir

da informação presente no núcleo. Pode-se então afirmar que o satélite ganha a sua importância através da relação que estabelece com o núcleo.

Em termos práticos, uma análise utilizando a RST começa pela divisão do texto em unidades mínimas (segmentos). Cada proposição é constituída por segmentos, aos quais são atribuídos um papel de núcleo ou satélite, entre os quais pode ser atribuída uma relação retórica específica. O resultado da organização das relações reflecte uma estrutura hierárquica do texto original. A maioria das relações entre os segmentos são assimétricas, podendo ser entre um núcleo e um satélite ou entre diferentes núcleos.

2.5.2 Sistema AuTema-Dis

O projecto AuTema-Dis (Leal, 2008) define uma arquitectura que, implementada computacionalmente, realiza a análise textual, considerando as informações mais relevantes dispostas na superfície de um texto, bem como, as relações de significação que se estabelecem entre os elementos linguísticos que a compõe. O objectivo do trabalho foi desenvolver uma base metodológica, cuja sua sistematização fosse capaz de:

- reconhecer a informação principal num determinado discurso, considerando o resultado de uma análise sintáctica automática;
- reorganizar as estruturas relevantes ao tema em árvores de dependência dos segmentos (DTS[9]);
- atribuir automaticamente algumas relações retóricas entre os segmentos e subsegmentos organizados nas DTS;
- produzir automaticamente uma estrutura sintética em língua natural, a qual representa informação apresentada na estrutura discursiva, considerando os resultados das etapas anteriores do processamento.

Foi desenvolvida uma arquitectura modular que através da execução sequencial dos diferentes módulos é capaz de realizar a análise textual. Na elaboração da arquitectura foi prevista a realização de quatro etapas de análise distintas, mas relacionadas entre si. Cada módulo consiste numa das etapas da análise, sendo que o resultado da execução de cada uma gera informações para a execução da

[9]DTS: dependency tree segments, em português, árvore de dependência dos segmentos. O conceito foi desenvolvido (Leal, 2008).

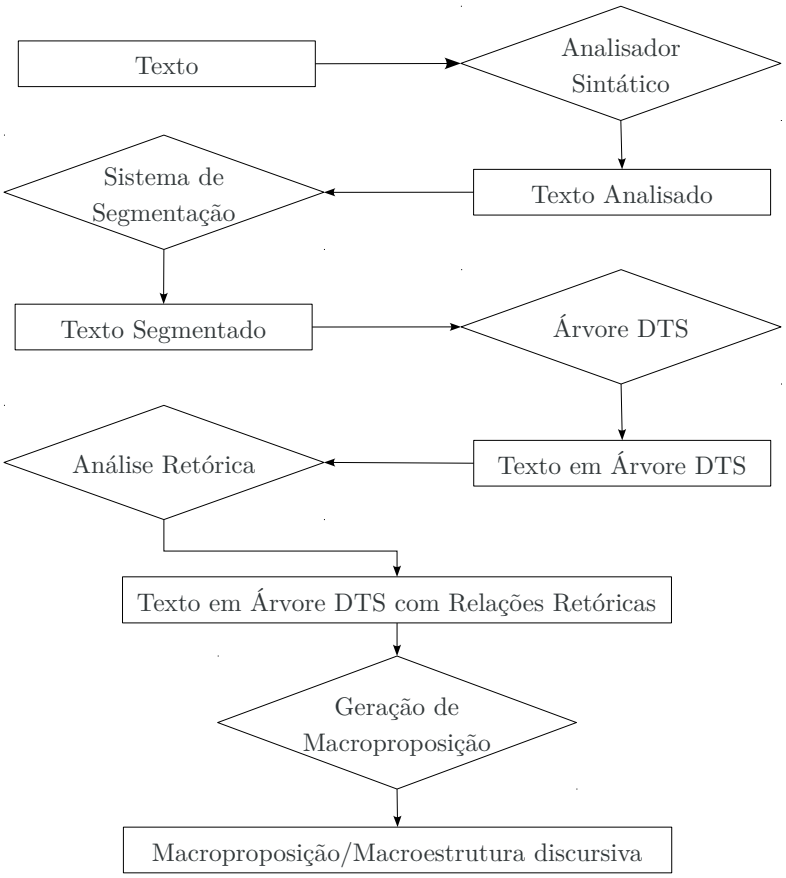

Figura 2.13: Arquitectura modular elaborada para a análise textual do sistema AuTema-Dis (Leal, 2008).

etapa seguinte. Foram determinados quatro módulos básicos para a realização da análise:

1. módulo para identificação, classificação e segmentação dos constituintes textuais;
2. módulo para organização arbórea dos constituintes textuais (DTS);
3. módulo para determinação das relações retóricas das DTS;
4. módulo para identificação da macroproposição e produção da macroestrutura discursiva.

De seguida serão apresentados os módulos que compõem o sistema. Na figura 2.13 é apresentado um esquema da arquitectura do sistema.

Módulo 1 - Identificação e Segmentação dos Constituintes Textuais

Esta primeira etapa tem como objectivo específico identificar as estruturas que constituem um texto e segmentá-las de acordo com a sua relação e importância com o tema do texto. Através da análise manual de um conjunto de textos do Jornal Público de 1994 e 1995 foram extraídas características que pudessem servir de base a um sistema automático para a análise textual. Após a identificação das regras foi utilizado um analisador automático para processar os textos e foram aplicadas as regras previamente definidas para a extracção dos segmentos.

Módulo 2 - Organização em Árvore (DTS)

Este módulo prevê a organização dos constituintes textuais, identificados no primeiro módulo, em árvores tipo DTS. As árvores de dependência de segmentos são utilizadas com o objectivo de demonstrar a hierarquia entre os segmentos que compõem a estrutura textual a partir da interacção de características estruturais, sintácticas e conceituais.

Módulo 3 - Identificação das Relações Retóricas

Neste módulo é utilizada uma metodologia para a identificação automática das relações retóricas entre segmentos e subsegmentos a partir da configuração em DTS. A constituição deste módulo conta com as informações advindas dos módulos iniciais. É estendida a organização arbórea realizada no segundo módulo sendo atribuídas algumas relações retóricas entre os segmentos (nós de primeiro nível) e subsegmentos (nós de segundo e terceiro níveis). Para a elaboração da metodologia que permitisse identificar as características presentes nos textos, que pudessem estar relacionadas directamente com a representação das relações retóricas, a autora recorreu aos textos dos corpora previamente analisados. Foi feita uma avaliação manual que possibilitou a atribuição de relações retóricas às várias estruturas encontradas nos textos. A autora baseou-se nos conjuntos de relações retóricas de (Mann e Thompson, 1988) e (Carlson e Marcu, 2001) não tento utilizado na integra todas as relações propostas. Optou pela utilização de algumas das relações retóricas de acordo com a necessidade de caracterizar as relações identificadas nos textos dos corpora. Foram ainda definidas novas relações retóricas em conformidade com a teoria RST.

Relação Retórica	Restrições e Efeitos	Exemplo
Apositiva de Nome Próprio	**Núcleo**: apresenta uma informação nominal pouco específica. **Satélite**: apresenta um Nome Próprio que especifica a informação descrita no núcleo. **Restrições N+S**: o S especifica através de uma expressão nominal própria, relacionada nominalmente ao que foi apresentado pelo N. **Efeito**: o leitor recebe do S a especificação através de um nome próprio daquilo que é mostrado no N.	(...) a posse dos presidentes do Banco do Brasil, **Paulo César Ximenes**, e da Caixa Económica Federal, **Sérgio Cutolo**. FSP950111-034

Tabela 2.2: Exemplo de uma das relações apresentadas por (Leal, 2008). Trata-se de um exemplo representativo da Relação Retórica Apositiva de Nome Próprio.

Módulo 4 - Representação Estrutural da Macroproposição Textual

Este módulo trata-se de um módulo resultado em que os constituintes textuais encontram-se organizados linearmente numa macroestrutura, representativa das macro-proposições que se configuram ao longo da estrutura analisada. Apresenta uma estrutura representativa do tema de um texto, para tal, é necessário considerar os dados e as características identificadas no nível da micro-estrutura, a fim de se chegar ao nível da macroestrutura/macroproposição, isto é, o texto na sua totalidade significativa.

2.6 Resumo do Capítulo

Neste capítulo foi apresentada a história e a evolução dos sistemas de sumarização automática desde a sua génese nos anos 50. Foram expostos vários sistemas de sumarização sendo enquadrados nos dois tipos de abordagens possíveis: superficial ou profunda. Nesta apresentação destacaram-se alguns projectos desta área para língua portuguesa, sendo todos eles focados em português do Brasil dado não existir praticamente nenhuma investigação na área do português Europeu. Finalmente foi discutida da teoria da estrutura retórica e o trabalho desenvolvido por

(Leal, 2008).

No próximo capítulo é descrito o sistema de sumarização implementado. São apresentados os vários módulos, desde a gramática até ao sistema de geração de extractos, sendo exibidas diversas figuras de forma a facilitar a compreensão do funcionamento do sistema. Inicialmente, são apresentados os conceitos e as teorias linguísticas nas quais este trabalho se baseia, na segunda parte do capítulo é feita uma breve descrição das linguagens e métodos de implementação utilizados.

Capítulo 3

Arquitectura

Neste capítulo é apresentada a arquitectura do sistema de sumarização e discutida brevemente a sua implementação. Na primeira secção são apresentados os conceitos e as teorias linguísticas nas quais este trabalho se baseia sendo descritos, em detalhe, os vários módulos que compõem o sistema. Na segunda secção deste capítulo é feita uma apresentação prática onde é discutida a implementação dos vários módulos, a sua ligação e formas de comunicação. São ainda apresentadas, em algum detalhe, as saídas dos vários módulos para um texto exemplo e a interface final da aplicação que foi desenvolvida.

3.1 Descrição Teórica

O projecto apresentado nesta tese segue a metodologia de (Leal, 2008) na organização dos módulos. Isto é, na elaboração da arquitectura foi prevista a realização de quatro etapas de análise distintas, mas relacionadas entre si. Cada módulo da arquitectura consiste numa das etapas de análise e o resultado da execução fornece dados para a execução da etapa seguinte até a conclusão de todo o processo. São definidos quatro módulos básicos para a realização da análise:

1. identificação e segmentação dos constituintes textuais;

2. organização em árvore constituintes textuais utilizando DTS;

3. identificação das relações retóricas entre os segmentos;

4. geração do sumário.

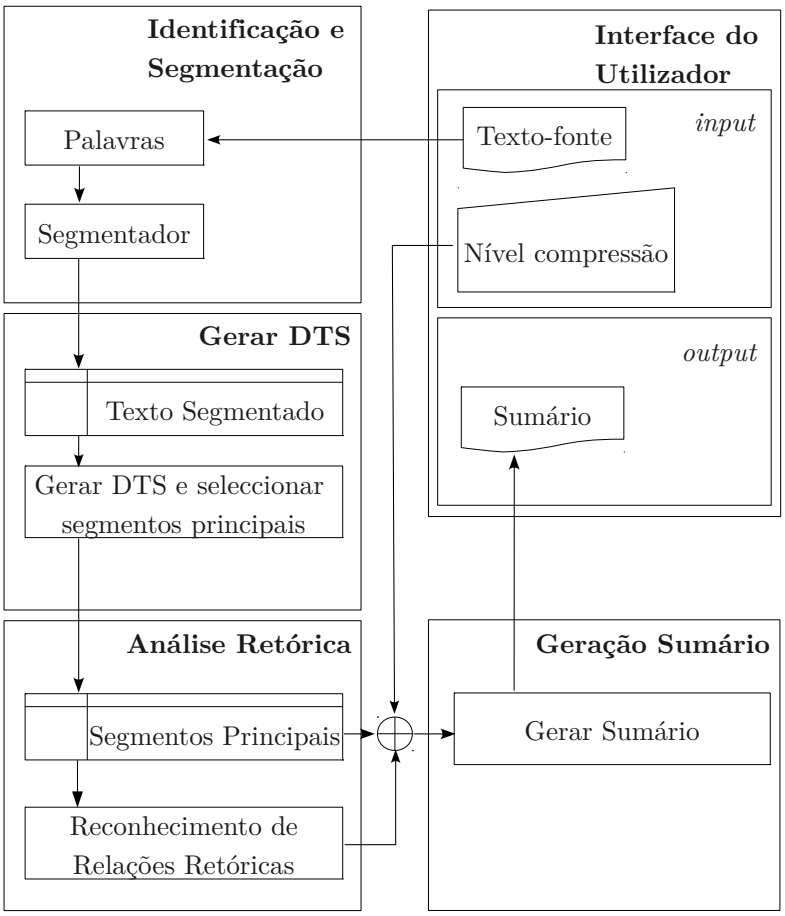

Figura 3.1: Arquitectura do sistema.

Na figura 3.1 é apresentada a arquitectura do sistema.

(Leal, 2008) refere que uma arquitectura modular é ideal para o tipo de análise que se pretende realizar, visto ser necessário garantir que cada um dos processos envolvidos possa ser realizado de forma autónoma. No entanto, apesar de se observar um determinado nível de independência entre os módulos, os resultados obtidos são necessariamente compartilhados nas sucessivas etapas. A metodologia prevê uma interacção entre todas as etapas de análise, não sendo possível chegar a um resultado satisfatório se ocorrer um problema na realização em qualquer uma das etapas.

Esta arquitectura modular tem a grande vantagem da fácil substituição de qualquer um dos módulos por uma nova versão desenhada para um caso específico (diferente do objectivo do desenho inicial do módulo). Por exemplo, se for con-

siderado que os problemas de performance advêm da análise sintáctica é possível substituir apenas este módulo por outro mais eficaz.

3.1.1 Identificação e Segmentação dos Constituintes Textuais

Nesta primeira fase, o objectivo é identificar as estruturas sintácticas que constituem o texto a analisar e dividi-las em segmentos. Sendo um dos objectivos deste projecto automatizar (tanto quanto possível) o processo de sumarização de um texto optou-se pela utilização de um analisador sintáctico automático. A utilização de um analisador automático, tem a vantagem de resolver o problema de existirem diferentes possibilidades no processo de identificação e determinação dos constituintes textuais para uma mesma estrutura.

Embora a construção de analisadores gramaticais automáticos seja, actualmente, uma área bem consolidada da linguística computacional, é uma tarefa complexa, dependente de restrições severas de robustez (para contemplar tarefas em tempo real) e abrangência (para servir a qualquer tipo de requisito). As principais razões da complexidade deste tipo de software advêm: da necessidade de se representar electronicamente grandes repositórios de informações linguísticas (léxicas, sintácticas e semânticas) e da incorporação de processos computacionais que consigam tratar, sistemática e eficientemente o processamento do texto. Considerando-se, em particular, o contexto do processamento automático do português, esta área reveste-se de maior importância, devido à escassez de ferramentas computacionais. Dos poucos analisadores que existem para a língua portuguesa o que apresenta melhores resultados é o Palavras, desenvolvido por (Bick, 2000), no âmbito do projecto VISL6, no Institute of Language and Communication da University of Southern Denmark.

Além de apresentar resultados satisfatórios, o analisador Palavras exibe o resultado do processamento em estruturas arbóreas devidamente etiquetadas com a identificação morfo-sintáctica e, em alguns casos, uma notação semântica. O resultado da análise é apresentado numa codificação específica, a partir de uma gramática própria, desenvolvida especificamente para análise automática de textos. Na figura 3.2 está patente a saída do Palavras na qual que podem ser observadas as características morfo-sintáctico-semânticas presentes na análise.

Como pode ser observado na figura 3.2, cada palavra é classificada na sua forma mais básica, havendo indicação de ordem morfológica associada à estrutura analisada. O analisador classifica ainda o género, o número, o tipo de verbo e a sua

```
STA:fcl
=SUBJ:np
==>N:art('o' <artd> F S)    A
==H:n('camara' F S) camara
=P:v-fin('nomear' PS 3S IND)  nomeou
=,
=ADVL:adv('entretanto' <kc>) entretanto
=,
=PRED:np
==>N:art('um' <arti> M S)   um
==H:n('grupo' <HH> M S) grupo
==N<:pp
===H:prp('de')  de
===P<:n('trabalho' <am> <act-d> M S) trabalho
=PRED:pp
==H:prp('com')  com
==P<:np
===>N:art('o' <artd> F S)   a
===H:n('finalidade' <am> F S) finalidade
 abastecimento
(restante conteúdo omitido)
```

Figura 3.2: Exemplo da saída parcial do Palavras para o texto Jornal Público-19950726-079.

conjugação, dados que são de extrema utilidade no processamento automático do texto. O analisador Palavras foi seleccionado como ferramenta a ser utilizada no primeiro módulo de processamento. É o primeiro passo (sub-módulo) do primeiro módulo do sistema, é responsável pela análise inicial do texto e será a partir desta análise que o primeiro módulo e módulos subsequentes irão trabalhar.

Foi seguido o trabalho de (Leal, 2008) para a definição das regras de segmentação. A autora estudou resultados e características da análise automática categorizando-os numa base de dados. Obteve um conjunto inicial de características, representativo das estruturas de dez textos em português europeu seleccionados a partir do corpus do Jornal Público. O conjunto das regras definido tem como objectivo identificar, e segmentar as estruturas presentes nos textos, no que toca

à categorização segmentos e subsegmentos, isto é, as unidades mínimas de significação. Após a definição a autora recorreu à implementação do sistema em Prolog de forma a validar as suas regras num conjunto mais abrangente de textos. Os dez textos iniciais foram submetidos ao sistema de segmentação implementado, apresentando um resultado satisfatório. Para tornar legítimo o conjunto das regras, que identificou a partir do corpus inicial, foi necessário testá-las noutros textos, assim, seleccionou quarenta novos textos. Os corpora utilizados são constituídos por dois corpus: um em português europeu e outro em português brasileiro.

(Leal, 2008) refere que o resultado da análise manual e análise do Palavras nos textos dos corpora, bem como, o tratamento/manipulação deste resultado revelou que os traços e as características linguísticas não ocorrem de forma aleatória, sendo possível colocar em categorias e quantificar padrões em forma de regras. Apresentou um conjunto de regras para reconhecimento e a classificação dos segmentos e subsegmentos, conforme as características identificadas na totalidade dos corpora, como pode ser identificado nas tabelas 3.1 e 3.2. Há regras distintas para identificação dos segmentos e dos subsegmentos. O número de regras para a identificação dos segmentos é mais reduzido do que as regras para os subsegmentos, segundo a autora essa diferença está relacionada à complementação verbal suportada na língua portuguesa.

Regras Segmento	Identificação da Regra
UTT:acl	Enunciado sem verbo - títulos, manchetes (jornal) e cabeçalhos
EXC:fcl	Estrutura Exclamativa
QUE:fcl	Estrutura Interrogativa
NPHR:prop	Enunciado sem verbo - com estrutura nominal própria (nome próprio)
NPHR:np	Enunciado nominal
STA:fcl	Enunciado com oração finita

Tabela 3.1: Regras para a identificação dos segmentos, bem como, a sua definição terminológica (Leal, 2008).

É de destacar que neste trabalho as *stopwords* não são removidas previamente. A remoção das *stopwords* do texto serve para diminuir o volume de processamento,

Regras subsegmentos	Identificação da Regra
N<:fcl	Informação Acessória /Complementar
Advl:pp	Circunstância Genérica
Advl:advp	Circunstância Genérica
Advl:fcl	Circunstância Genérica
Pred:pp	Circunstância Genérica
Advl:cu	Circunstância Genérica
App:prop	Circunstância Apositiva Nome Próprio
Advl:acl	Avaliação
Sta:icl	Acção
Co:conj-c ('mas')	Oposição / Antítese
Pred:np	Elaboração - Circunstância Genérica
App:np	Complementação Nominal Apositiva
Advl:adv - atemp	Quantificação Temporal
Advl:adv - aloc	Quantificação Locativa
Advl:np ou Advl:n	Circunstância de Tempo Decorrido

Tabela 3.2: Regras para identificação dos subsegmentos, bem como, a sua classificação terminológica (Leal, 2008).

mas (Riloff, Wiebe e Phillips, 2005) referem que estas palavras também são relevantes, porém não existem estudos comparativos que comprovem tal eficiência ou não. Os sinais de pontuação são considerados como indicativos de segmentação e estão incluídos no conjunto das regras que identificam e realizam a segmentação textual. É de salientar que os indicativos de pontuação não são os elementos que determinam os segmentos e os subsegmentos e não são eles os responsáveis pelo processo de segmentação dos constituintes textuais. Existem, no entanto, situações em que o analisador Palavras não consegue demarcar os constituintes, nestes casos recorre à utilização da pontuação para auxiliar o processo de segmentação das unidade textuais.

Como observamos na figura 3.2, o Palavras analisa e atribui uma classificação a todas as palavras do texto analisado. São também atribuídas características às estruturas (frases ou orações), sendo essas características utilizadas em conjunto

com regras propostas. A associação das características atribuídas às estruturas textuais pelo Palavras e as regras propostas para identificação dos constituintes determinam que ponto da estrutura textual oferece a possibilidade mais adequada para a segmentação, bem como, a determinação das suas fronteiras.

3.1.2 Organização de Segmentos em DTS

O segundo módulo organiza dos constituintes textuais, identificados no primeiro módulo, em árvores tipo DTS.

As características estruturais necessárias para a organização dos constituintes textuais em árvores, são identificadas tendo em conta as informações provenientes do primeiro módulo, isto é, identificação dos segmentos e subsegmentos. Conforme foi apresentado, as características que determinam as possibilidades de segmentação são constituídas a partir da análise sintáctica realizada pelo Palavras.

Além da codificação e da colocação de etiquetas nas estruturas, o resultado do Palavras contém o nível de profundidade em que cada um dos constituintes se encontra no interior das estruturas das quais fazem parte. A informação dos níveis de profundidade em que se encontram os segmentos, que compõem as estruturas ao longo do texto, é de extrema relevância para a determinação da composição e organização hierárquica das árvores DTS. Esta classificação dos níveis de profundidade apresentados na análise realizada pelo Palavras fornece dados que podem, em alguns casos, ser incorporados nas regras de segmentação. Os níveis de profundidade apresentados a partir do Palavras podem ser observados na figura 3.3.

A disposição dos constituintes textuais nas árvores DTS estão condicionadas ao nível de profundidade que ocupam na estrutura da qual fazem parte, bem como, a relação que se estabelece entre eles. É de referir que o segmento detém um papel subordinante em relação aos subsegmentos. As árvores DTS organizam os segmentos principais como nós de primeiro nível da árvore e os subsegmentos são identificados como nós de segundo e terceiro níveis, conforme figura 3.4. Os demais subsegmentos, ou seja, aqueles que se encontram em níveis de profundidade posterior ao terceiro nível, isto é, subsegmentos do quarto, quinto e demais níveis, não são organizados de maneira hierárquica em nós, são aglutinados com o nó de terceiro nível correspondente.

Dado que o objectivo deste módulo é a organização dos segmentos e subsegmentos em árvores do tipo DTS, é necessária a definição de regras que permitam utilizar os dados gerados pelo primeiro módulo e a informação complementar de profundidade em que se encontram cada um dos constituintes do texto. Mais

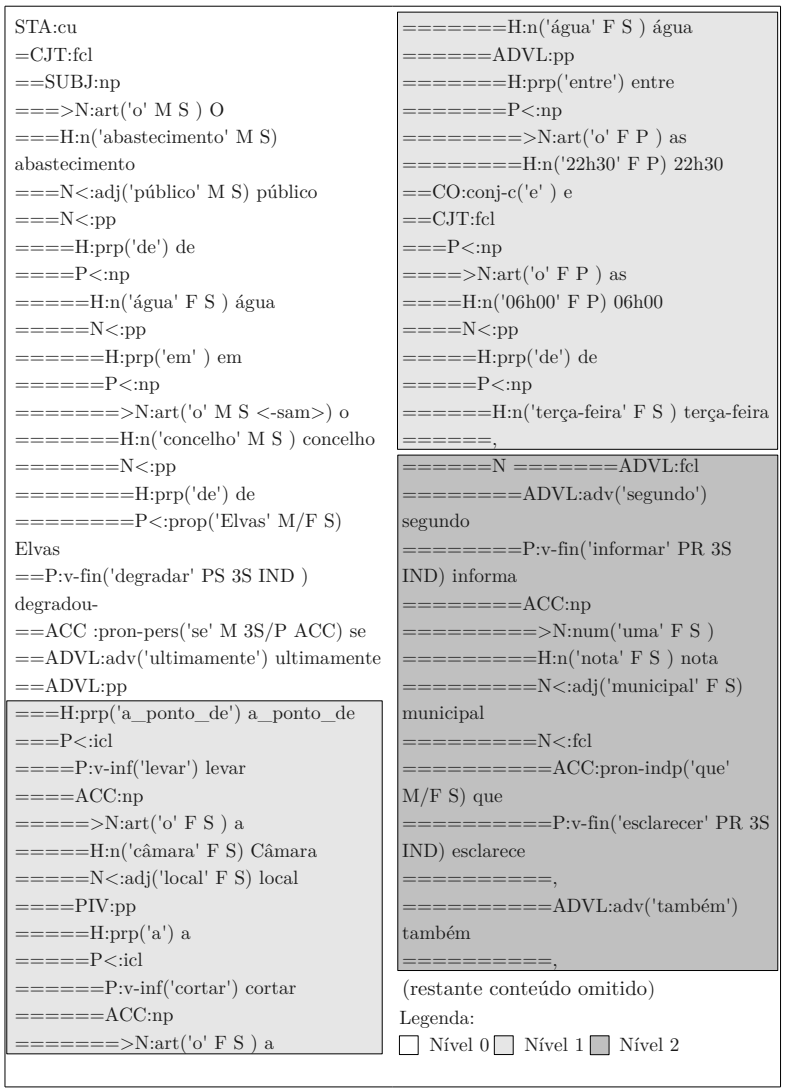

Figura 3.3: Análise automática do Palavras com a marcação dos níveis de profundidade em se se encontram os constituintes da estrutura para o texto Jornal Público-19950726-079.

uma vez será necessário recorrer ao trabalho de (Leal, 2008) de forma a balizar teoricamente a operação do módulo.

Numa primeira análise, a autora acreditou que as regras de segmentação poderiam ser suficientes para organização dos segmentos em árvores, no entanto, essa não foi a realidade observada na prática. As regras de segmentação são úteis para a identificação linear dos segmentos, mas não os distinguem entre si. Assim, era

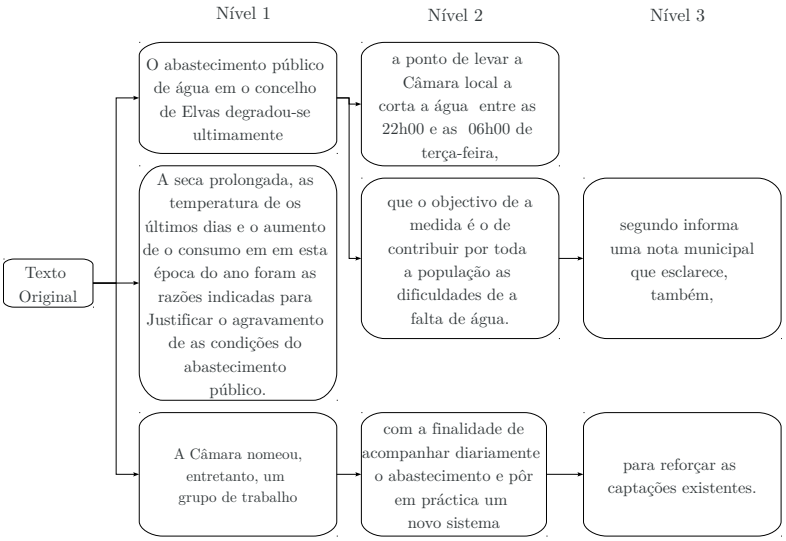

Figura 3.4: Representação da organização hierárquica de um texto em DTS com especificação dos níveis para o texto Jornal Público-19950726-079.

necessário encontrar alguma característica que associada à regra pudesse especificar os segmentos em função do seu papel na estrutura do texto. Recorreu à análise do Palavras para encontrar características que pudessem ser associadas às regras e que contemplassem essa diferença entre os segmentos.

Foi definida uma metodologia em que as regras de segmentação recebem a caracterização de nível de profundidade, isto é, cada regra que identifica um segmento ou um subsegmento tem agregada informação que determina a profundidade que cada segmento ou subsegmento pode ocupar na árvore DTS, conforme apresentado na tabela 3.3. Assim, à priori, os constituintes identificados como segmentos ocupam os nós de primeiro nível e os constituintes identificados como subsegmentos ocupam os nós de segundo e terceiro níveis. Mais um vez, é de notar que a análise realizada pelo Palavras determina níveis de profundidade além do terceiro mas como foi referido anteriormente esses nós são aglutinados com o nó de terceiro nível correspondente.

3.1.3 Análise Retórica

Este é o terceiro módulo do processamento. Utiliza os dados processados no primeiro módulo (que dividiu o texto nos seus constituintes - segmentos e subsegmentos) e a DTS gerada no segundo módulo para atribuir algumas relações

Níveis de Profundidade	Regras para Segmentos
Todos segmentos encontram-se no nível 1 de profundidade na estrutura	UTT : acl
	EXC : fcl
	QUE : fcl
	NPHR : prop
	NPHR : np
	STA : fcl
Níveis de Profundidade	Regras para subsegmentos
Todos subsegmentos encontram-se nos níveis 2 e 3 de profundidade na estrutura	N<:fcl
	Advl:pp
	Advl:advp
	Advl:fcl
	Pred:pp
	Advl:cu
	App:prop
	Advl:acl
	Sta:icl
	Co:conj-c ('mas')
	Pred:np
	App:np
	Advl:adv - atemp
	Advl:adv - aloc
	Advl:np ou Advl:n

Tabela 3.3: Regras de segmentação dos constituintes textuais e os níveis de profundidade que os segmentos e os subsegmentos podem ocupar em uma estrutura.

retóricas entre os segmentos (nós de primeiro nível) e subsegmentos (nós de segundo e terceiros níveis).

Como foi apresentado anteriormente, as árvores DTS apresentam o texto disposto numa estrutura esquemática na qual os segmentos e subsegmentos aparecem

hierarquicamente dispostos. Este tipo de organização é favorável à identificação das relações retóricas, dado que, as relações manifestam-se entre os constituintes respeitando a ordenação e a subordinação existente entre esses elementos.

Sendo este módulo também baseado no trabalho de (Leal, 2008) é importante referir que a autora realizou uma pesquisa dos padrões linguísticos nos vários textos dos corpora e simultâneamente efectuou a identificação manual de relações retóricas entre os constituintes dos vários textos. Como resultado desta actividade, foi possível verificar que a melhor forma de identificar as relações retóricas é considerar apenas informações linguísticas da superfície textual, assim, seriam associadas relações retóricas às regras de segmentação com marcação de nível, criadas no segundo módulo. Desta forma, foi possível excluir todo e qualquer tipo de intervenção humana.

Para a realização do processo de análise manual, a autora optou pela utilização de algumas das relações de (Mann e Thompson, 1988) e de (Carlson e Marcu, 2001) de acordo com a necessidade de caracterização das relações identificadas nos textos dos corpora. Das duas propostas, foram utilizadas as seguintes relações retóricas:

- Mann and Thompson: circunstância; avaliação; antítese, elaboração;
- Daniel Marcu: same-unit;

A autora identificou ainda um novo grupo de relações, desenvolvidas exclusivamente para suprir particularidades evidenciadas nos textos dos corpora. Foram agregadas mais cinco relações ao conjunto, sendo elas: apositiva de nome próprio; quantificação temporal, quantificação locativa, acção, circunstância de tempo decorrido. Encontra-se disponível mais informação relativamente a estas relações no apêndice B.

A autora determinou que a melhor opção para a automatização das relações seria associá-las a uma regra de identificação dos segmentos/subsegmentos. Verificou ainda que mais de uma regra para a identificação dos subsegmentos pode estar indexada um mesmo tipo de relação como é o caso das regras: ADVL:pp; ADVL:fcl; ADVL:advp; ADVL:cu e Pred:pp todas elas indexadas à relação retórica circunstância genérica. No conjunto das relações utilizadas na metodologia existem quinze regras para identificação dos subsegmentos implementadas, mas somente onze relações retóricas a elas indexadas. (Leal, 2008) optou por não indexar nenhuma relação retórica às seis regras de identificação de segmentos.

A tabela 3.4 apresenta as as relações retóricas já existentes conjuntamente com as que foram desenvolvidas pela autora indexadas às regras para identificação dos segmentos e dos subsegmentos respectivos.

Relações Retóricas	Regra para Segmento
Não definidas	UTT : acl
	EXC : fcl
	QUE : fcl
	NPHR : prop
	NPHR : np
	STA : fcl
Relações Retóricas	**Regra para subsegmento**
Same-Unit	N<:fcl
Circunstância Genérica	Advl:pp
	Advl:advp
	Advl:fcl
	Pred:pp
	Advl:cu
Circunstância Apositiva Nome Próprio	pp:prop
Avaliação	Advl:acl
Acção	Sta:icl
Oposição/Antítese	Co:conj-c ('mas')
Circunstância Apositiva Nome Próprio	Pred:np
Complementação Nominal Apositiva	App:np
Quantificação Temporal	Advl:adv - atemp
Quantificação Locativa	Advl:adv - aloc
Circunstância de Tempo Decorrido	Advl:np ou Advl:n

Tabela 3.4: A figura representa as relações retóricas indexadas às regras para identificação dos segmentos e subsegmentos (Leal, 2008).

3.1.4 Geração de Sumário

Este módulo segue a metodologia de (Leal, 2008) na medida em que faz uso da sua definição de selecção de frases para construir o sumário. No seu trabalho

Ana Luísa lança mão dos conceitos macroestrutura (representação abstracta da estrutura global de significado de um texto), macro-regras (regras gerais e convencionais que permitem elaborar o resumo e aceder à macroestrutura do texto) e macro-proposições (nível intermediário do texto entre o tema e os detalhes) (Dijk, 1972; Dijk, 1993; Dijk, 1992) com o objectivo de apresentar uma estrutura sintética representativa da macroestrutura do texto avaliado, isto é, a macroproposição.

Tal como referido por Van Dijk, a macroproposição pode ser considerado um sumário. Assim, este último módulo do processamento utiliza os dados dos três módulos precedentes para gerar o sumário. Trata-se de um módulo resultado que organiza linearmente os constituintes textuais mais importantes.

(Dijk, 1992 apud Leal, 2008) apresenta um conjunto de operações (macro-regras) que seleccionam, reduzem, generalizam e (re)constroem proposições noutras proposições menores. Com a aplicação destas macro-regras é possível suprimir informação que não é necessária para a compreensão do resto do discurso. A autora ressalva que as macro-regras de Dijk não foram utilizadas na sua essência no seu trabalho. Foram utilizadas como forma de orientação, ajudando a caracterizar as regras utilizadas para seleccionar os constituintes textuais que não apresentam conteúdo relacionado directamente com o assunto do texto tratado. (Leal, 2008) refere ainda que os níveis de profundidade relacionam-se com as macro-regras. Quanto mais interno estiver um subsegmento numa dada estrutura menor será a sua relação com o tema global e maior será a probabilidade deste elemento poder ser descartado. Desta forma alguns subsegmentos são eliminados pela aplicação da macro-regra de supressão.

O processo de selecção de segmentos e subsegmentos que constituem o sumário segue a metodologia de (Leal, 2008) a partir das regras de selecção propostas no segundo módulo acrescidas da informação sobre a posição/profundidade em que se encontra o constituinte conforme o princípio proposto por (Dijk, 1992 apud Leal, 2008). O sistema descrito nesta tese apenas diverge do de (Leal, 2008) dado que permite ao utilizador escolher quantos níveis pretende usar para o sumário final (esta selecção afectará a taxa de compressão do sumário).

3.2 Implementação

Nesta secção é apresentada a implementação dos vários módulos que constituem este sistema. Visto que cada módulo funciona de forma independente foi possível escolher um conjunto de ferramentas (por módulo) que melhor se adequam a resolução dos diversos problemas. As duas necessidades principais de todo o projecto

são a manipulação simbólica de texto (utilizada em todos os módulos de processamento) e o desenvolvimento de uma interface. As ferramentas escolhidas foram o Prolog para a manipulação simbólica de texto e o PHP para a interface.

A escolha da linguagem Prolog decorreu da ampla experiência existente na Universidade de Évora na utilização desta linguagem. É uma linguagem muito utilizada na área de inteligência artificial, em particular na linguística computacional (Bratko, 2000; Covington, 1993). O Prolog tem a sua origem na lógica de primeira ordem e na lógica formal. É uma linguagem é declarativa, isto é, a lógica do programa é expressa em termos de relações e representada por factos e regras. A computação é iniciada pela execução de uma pesquisa no conjunto de relações. Esta linguagem foi concebida por um grupo liderado por Alain Colmerauer em Marselha (França) no inicio da década de setenta. O primeiro sistema Prolog foi desenvolvido em 1972 por Alain Colmerauer e Philippe Roussel (Colmerauer e Roussel, 1993).

No módulo interface, um dos requisitos e, provavelmente o mais importante, era a simplicidade, qualquer utilizador devia poder utilizar facilmente o software. No actual panorama de uso de aplicações *online* revela-se que a interface mais simples e mais comum para um utilizador é um *browser web*. Fica assim justificada a escolha de uma interface deste tipo. Existem imensas ferramentas/linguagens para o desenvolvimento *web*. Era necessária uma linguagem simples e de desenvolvimento rápido, foi escolhido o PHP[1]. É uma linguagem de script de alto nível desenhada para o desenvolvimento *web*. O código PHP, geralmente, é escrito conjuntamente com o código HTML de uma dada página, sendo posteriormente interpretado pelo servidor através de um módulo de processamento que gera a página final. O PHP foi desenvolvido por Rasmus Lerdorf em 1995. Em Abril de 2007 mais de 20 milhões de domínios da Internet tinham serviços de PHP (SecuritySpace, 2007). O PHP é usado em mais de 75% de todos os servidores *web* (W3Techs, 2010).

3.2.1 Identificação e Segmentação dos Constituintes Textuais

O primeiro módulo de processamento tem como entrada um texto a sumarizar. Este é processado pelo Palavras de forma a obter a análise análise gramatical. A partir desta análise são construídos átomos Prolog os quais serão utilizados no processamento do segundo módulo. Este módulo de processamento recorre a dois

[1]Mais informação sobre o projecto disponível no site: http://www.php.net

comandos que processam todo o texto:

```
$ cat p19950726-079.txt | txt2visl > p19950726-079.visl
$ cat p19950726-079.visl | visl2pl > p19950726-079.pl
```

Figura 3.5: Conjunto de comandos para a execução primeiro módulo.

O resultado do processamento deste módulo é um ficheiro de átomos Prolog, que pode ser observado na figura 3.6. Esta saída serve como entrada para o módulo seguinte.

```
sentence(syn(sta(fcl,subj(np,n(art('o','<artd>','M','S'),'
    O'),h(n('abastecimento','M','S'),'abastecimento'),n(adj
    ('público','M','S'),'público'),n(pp,h(prp('de'),'de'),p
    (cu,cjt(np,h(n('água','<cm-liq>','F','S'),'água'),n(pp,
    h(prp('em','<sam->'),'em'),p(np,n(art('o','<artd>','<-
    sam>','M','S'),'o'),h(n('concelho','<HH>','M','S'),'
    concelho'),n(pp,h(prp('de'),'de'),p(prop('Elvas','<hum>
    ','M/F','S'),'Elvas'))))))),p(v_fin('degradar','<hyfen>
    ','PS','3S','IND'),'degradou-'),acc(pron_pers('se','M/F
    ','3S','ACC/DAT'),'se'),advl(adv('ultimamente'),'
    ultimamente'),advl(pp,h(prp('a_ponto_de'),'a_ponto_de')
    ,p(icl,p(v_inf('levar','0/1/3S'),'levar'),acc(np,n(art
    ('o','<artd>','F','S'),'a'),h(n('câmara','F','S'),'
    Câmara'),
(restante conteúdo omitido)
```

Figura 3.6: Resultado da execução do primeiro módulo para o texto Jornal Público-19950726-079

3.2.2 Organização de Segmentos em DTS

No segundo módulo de processamento os átomos provenientes do primeiro módulo são organizados em árvores DTS. Todo o processamento dos átomos Prolog é realizado recorrendo às regras de (Leal, 2008). Durante a conversão os segmentos

e subsegmentos em árvores são corrigidos alguns problemas de apresentação, por exemplo, a conversão da análise do Palavras para átomos Prolog origina alguns átomos sem conteúdo que são removidos nesta fase.

3.2.3 Análise Retórica

Neste módulo são atribuídas relações retóricas a alguns ramos das árvores DTS. Apesar da figura 3.7 não o evidenciar é possível que nem todos os ramos tenham relações atribuídas, isto deve-se à dificuldade de atribuição de relações retóricas a todas as relações.

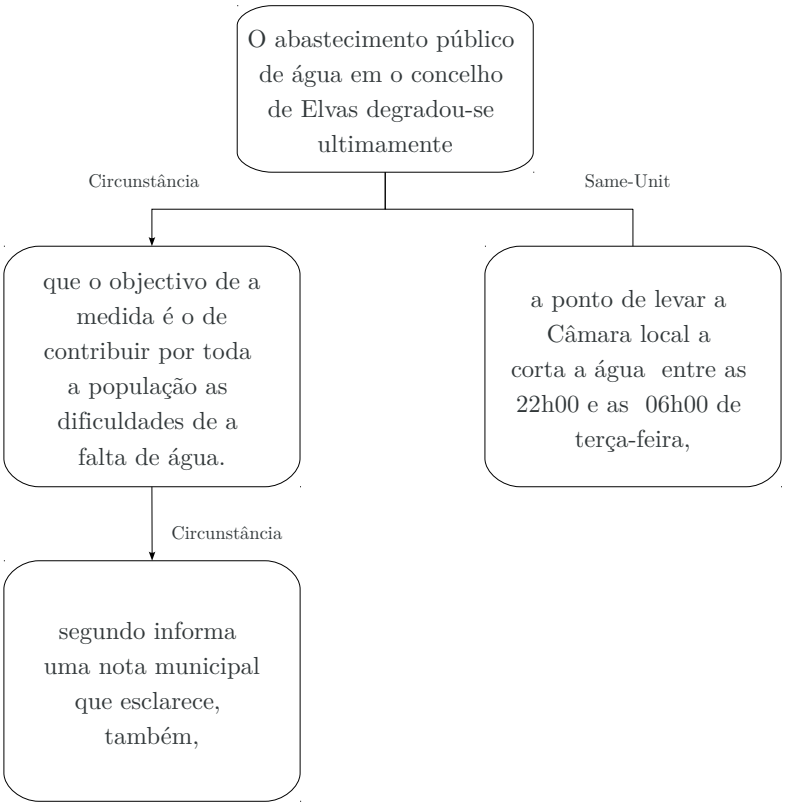

Figura 3.7: Relações retóricas entre segmentos e subsegmentos no texto Jornal Público-19950726-079

3.2.4 Geração de Sumário

Esta é a última fase do processamento. Neste último módulo são escolhidos quais os segmentos a manter e quais a serão retirados da árvore por não conterem informação suficiente. Após a selecção dos segmentos estes são aglutinados pela ordem que ocorriam no texto inicial para a geração do sumário. A figura 3.8 apresenta um exemplo de um sumário gerado pelo sistema para um texto do Jornal Público.

```
O abastecimento público de água em o concelho de Elvas
degradou-se ultimamente.
A seca prolongada, as temperaturas de os últimos dias e o
aumento de o consumo em esta época de o ano foram as
razões indicadas para justificar o agravamento de as
condições de o abastecimento público.
A câmara nomeou, entretanto, um grupo de trabalho.
```

Figura 3.8: Exemplo de um sumário para o texto Jornal Público-19950726-079.

3.2.5 Interface Web

O ultimo módulo a implementar trata da relação directa do sistema com o utilizador. Como pode ser observado na figura 3.1 a interface com o utilizador é composta por dois módulos distintos: o módulo de entrada ou *input* e o módulo de saída ou *output*. O módulo de *input* solicita ao utilizador a introdução do ficheiro a partir do qual pretende gerar o sumário e informação relativa ao tipo de taxa de compressão. Neste módulo é ainda possível configurar a exibição dos resultados intermédios do processamento do texto (*parser*, segmentação e relações retóricas). No módulo de *output* o utilizador tem acesso ao sumário gerado pelo sistema e qualquer um dos dados intermédios solicitados inicialmente. A interface web é apresentada nas figuras 3.9 e 3.10 para os módulos *input* e *output* respectivamente.

3.3 Resumo do Capítulo

Neste capítulo foi apresentada a arquitectura do sistema. Inicialmente foi feita uma contextualização teórica da implementação do sistema e os trabalhos ante-

Figura 3.9: Interface do sistema - entrada de dados.

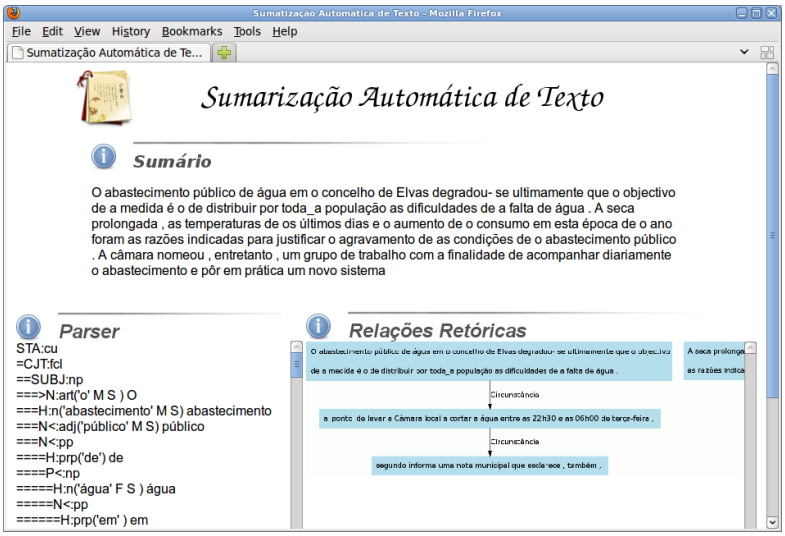

Figura 3.10: Interface do sistema - saída de dados.

riores que o influenciaram. Na parte final foi discutida a implementação real do sistema com uma descrição dos diversos módulos. Por fim é descrita a interface *web* que permite a interacção de um utilizador leigo com o sistema.

No próximo capítulo é discutida a avaliação do sistema. Inicialmente é feita uma contextualização teórica referindo os problemas relativos à avaliação de sis-

temas de sumarização automática e apresentados os vários tipos de avaliação possíveis. São referidos os resultados deste e de outros sistemas para a língua portuguesa.

Capítulo 4

Avaliação

A avaliação de um sumário não é uma tarefa fácil. É um procedimento bastante difícil dado que não existe um modelo de sumário ideal para um determinado documento ou conjunto de documentos. Observando os vários trabalhos apresentados ao longo desta tese, facilmente se percebe que o acordo entre "sumarizadores" humanos é muito baixo, tanto para avaliar, como para gerar sumários. Mesmo para sumários relativamente simples de artigos noticiosos, os especialistas apenas concordam cerca de 60% das vezes (Radev, Hovy e McKeown, 2002). (Mani et al., 1999) referem que "a sumarização de texto ainda é uma área emergente, ainda existem questões sérias acerca dos métodos e tipos de avaliação". Mais do que a forma do sumário, é difícil avaliar o seu conteúdo.

Um outro problema existente na avaliação é o uso generalizado de métricas díspares. A ausência de uma métrica padrão de avaliação humana ou automática torna difícil comparar os diferentes sistemas e estabelecer uma base. Segundo (Lin, 2004), a avaliação manual em larga escala de sumários como nas conferências DUC[1] exigiria mais de 3000 horas de trabalho, o que torna este tipo de avaliação muito dispendioso. Uma métrica com alta correlação com a avaliação humana seria o ideal para evitar o processo de avaliação manual.

Dado que a avaliação dos sistemas de geração automática de sumários é uma tarefa essencial para a validação de resultados e consequente evolução desta área, foram criadas uma série de conferências dedicadas a este tema. A TISPSTER SUMMAC[2] (Mani et al., 1999) foi a primeira a fazer uma avaliação, independente em larga escala, de sistemas de sumarização automática de texto. Surgiu

[1] *Document Understanding Conferences* (DUC) - mais informação em: http://duc.nist.gov/

[2] *TIPSTER Text Summarization Evaluation Conference* (SUMMAC) - mais informação em: http://www-nlpir.nist.gov/related_projects/tipster_summac

mais tarde a conferência DUC, que era inicialmente composta por uma série de workshops de avaliação organizados para incentivar a investigação na área do Processamento de Linguagem Natural e aplicações relacionados. Mais recentemente esta conferência tornou-se uma secção da TAC (*Text Analysis Conference*). Oferece uma grande colecção de testes e procedimentos de avaliação e um fórum para a partilha de resultados. A DUC/TAC está entre os eventos mais destacados, a nível mundial, dedicados à avaliação dos sistemas de sumarização automática (Rino e Pardo, 2003).

(Mani e Maybury, 1999 apud Rino e Pardo, 2003) e (Mani, 2001) referem algumas das principais dificuldades da avaliação dos sistemas de sumarização automática:

- A sumarização automática envolve a utilização de uma máquina para produzir uma saída que resulta em comunicação em linguagem natural. Nos casos em que a saída é uma resposta a uma pergunta, é relativamente simples perceber qual a resposta correcta, mas nos outros casos, é difícil chegar à noção do que é a saída correcta. Há sempre a possibilidade de um sistema gerar um bom sumário mas que é completamente diferente de qualquer sumário criado por um humano (que poderia ser usado como uma aproximação para a saída correcta);

- Como pode ser necessário recorrer a um ser humano para julgar a saída do sistema:

  - O custo da avaliação pode ser muito elevado;
  - Nem sempre há indivíduos com o perfil adequado disponíveis;
  - A identificação da forma de utilizar o julgamento humano nem sempre é trivial;
  - Se for pretendida uma avaliação robusta e abrangente certamente o julgamento será muito lento e complexo;
  - O alto grau de subjectividade do julgamento humano torna difícil o estabelecimento de conclusões definitivas.

  Seria preferível um processo de avaliação que fizesse uso de um programa de pontuação, em vez dos juízos humanos, uma vez que é facilmente reproduzível;

- Sumarização envolve a compressão, sendo importante a possibilidade de avaliar os sumários com taxas de compressão diferentes. Esta necessidade aumenta a escala e complexidade da avaliação. Em geral, quanto mais alta a taxa de compressão menos informativo será o sumário e vice-versa;

- Deve ser tido em conta que a sumarização envolve a apresentação de informação de uma maneira sensível às necessidades de um utilizador ou de uma aplicação. Esta característica dificulta o desenho do sistema de avaliação.

4.1 Tipos de avaliação

(Jones e Galliers, 1996 apud Rino e Pardo, 2003) foram os primeiros autores a definir as directrizes para a avaliação de sistemas de sumarização automática, as quais têm sido amplamente adoptadas. A primeira grande distinção refere a forma de avaliação: ela pode ser intrínseca ou extrínseca.

Uma avaliação intrínseca avalia o desempenho do sistema, pela verificação da quantidade e qualidade de informação dos sumários produzidos. A avaliação extrínseca verifica a adequação do sistema para o uso em tarefas específicas, distintas da sumarização automática, isto é, o sistema é avaliado em função de como este influencia a realização de outra tarefa que utiliza os sumários produzidos automaticamente (por exemplo a determinação da relevância de um texto para um dado assunto numa tarefa relacionada com a categorização de textos).

A validação de sumarizadores automáticos tem sido aplicada em sistemas de pergunta/resposta, de categorização automática de documentos e sistemas de recuperação de informação.

Quando a avaliação faz uso do julgamento humano diz-se que é *on-line*, caso contrário se são utilizados métodos automáticos diz-se *off-line*. Embora as avaliações *off-line* sejam preferíveis, ainda não existem métodos automáticos de avaliação que sejam tão satisfatórios quanto o julgamento humano (Rino e Pardo, 2003).

No caso em que apenas são avaliados os resultados finais do sistema, a avaliação é chamada *black-box*. O sistema é visto como uma "caixa-preta". Ou seja, os processos intermediários da sumarização não são avaliados. A comparação entre um sumário produzido automaticamente e seu texto-fonte é um exemplo deste tipo de avaliação. Se forem considerados resultados intermediários, a avaliação é chamada *glass-box*. Para um sistema do tipo apresentado na figura 2.2, uma avaliação *glass-box* avaliaria todas as fazes do sistema (análise, transformação e síntese).

Finalmente, se os resultados de um sistema de sumarização automática são comparados com os resultados de outro sistema, diz-se que a avaliação é comparativa. Caso contrário, diz-se que ela é autónoma. A avaliação comparativa é, normalmente, utilizada nas grandes conferências internacionais (SUMMAC, DUC, etc). Nessas conferências os sistemas participantes são pontuados pelo seu desempenho e comparada a sua pontuação final.

(Jones e Galliers, 1996 apud Rino e Pardo, 2003) referem que a característica mais importante na avaliação de um sistema de sumarização automática é a definição clara do que se pretende avaliar. Tendo esse facto em conta, é fácil determinar quais dos tipos de avaliação a aplicar, isto é, se ela será intrínseca ou extrínseca, *on-line* ou *off-line*, *black-box* ou *glass-box* e comparativa ou autónoma. É no entanto importante referir que estes tipos de avaliação não são exclusivos. Caso se pretenda proceder a uma avaliação intrínseca e a uma extrínseca, é totalmente possível e viável, depende somente dos objectivos da avaliação a realizar.

Seguem-se os métodos e métricas mais utilizados para a avaliação intrínseca, seguindo-se aqueles da avaliação extrínseca.

4.1.1 Avaliação Intrínseca

A avaliação intrínseca envolve as medidas de qualidade e quantidade de informação dos sumários produzidos automaticamente (Mani, 2001). Parâmetros linguísticos como a coerência e a coesão (Rino, 1996) são geralmente utilizados neste tipo de avaliação. É ainda possível comparar os sumários produzidos com sumários de referência, denominados sumários ideais. De seguida são destacados os principais aspectos da avaliação intrínseca de um sistema de sumarização automática.

Coerência e Qualidade do Sumário

Como se sabe um dos aspectos mais importantes de um sumário é a sua leitura. Os sumários automáticos, em geral, são produzidos recorrendo a extractos de um texto original, o que pode levar a problemas de coerência como por exemplo: anáfora pendurada e lacunas na estrutura retórica do sumário. Critérios de legibilidade são muitas vezes ligados a estes tipos de avaliação.

Os critérios para julgar a qualidade dos sumários dependem bastante do autor. Por exemplo, (Sylvaine et al., 1997) solicitaram que juízes humanos pontuassem os sumários observando os critérios: presença de referências anafóricas não resolvidas, não preservação da integridade de estruturas (por exemplo listas e tabelas), falta de coesão entre as frases do sumário. Já (Saggion e Lapalme, 2000), solicitaram

aos juízes que dessem notas aos sumários, observando ortografia e gramática, a indicação clara do assunto do texto-fonte, o estilo impessoal, a concisão, legibilidade e facilidade de compreensão do sumário. (Pardo, 2002a), também solicitou a juízes que atribuíssem notas a sumários de acordo com sua qualidade textual, isto é, sua coerência e coesão. O autor utilizou ainda outra sugestão de (Mani, 2001), avaliou a legibilidade dos sumários quando comparada com a legibilidade dos textos-fonte correspondentes.

É importante salientar que a legibilidade não é um critério decisivo, nem suficiente, para se poder afirmar que um sumário é bom. De facto, como discutido por (Mani, 2001), essa medida é muito "ingénua" dado que assume que o tamanho das palavras ou das frases é o único factor que pode influenciar a legibilidade de um texto. Como a avaliação da qualidade de sumários necessita de juízes humanos, os investigadores têm procurado formas automáticas de realizar tal avaliação.

A qualidade é um bom parâmetro para se centralizar as avaliações de sistemas extractivos, pois estes produzem, em geral, sumários em que a fluência do conteúdo não é muito boa. De qualquer forma, mesmo com a fluência prejudicada, ainda é possível obter sumários úteis. Devido à complexidade de modelação de tantos aspectos distintos sobre a qualidade e/ou utilidade de sumários automáticos, tornou-se comum fazer a avaliação somente pela verificação do conteúdo que é preservado, em relação a seus textos-fonte.

Conteúdo de Informação do Sumário

O conteúdo de informação do sumário, muitas vezes referido por informatividade, visa avaliar a quantidade de informação útil que o sumário contém. Quanto menor o sumário, menor é a quantidade de informação do texto-fonte que pode ser preservada. Portanto, uma medida da informatividade de um sumário é a quantidade de informação do texto-fonte que é preservada no sumário. Outra medida é a quantidade de informação do sumário de referência (sumário ideal) que é coberta pelo sumário gerado pelo sistema automático. Noutras palavras, como no caso da coerência, podem ser feitas comparações entre os sumários gerados por um sistema automático, o texto-fonte, sumários de referência e avaliações de outros sistemas de sumarização. A informatividade é uma métrica mais simples de utilizar em sistemas de avaliação automáticos do que a métrica coerência.

Um sumário de referência para um texto-fonte pode ser conseguido de várias formas:

- sumário autêntico: sumário produzido pelo próprio autor do texto-fonte;

- sumário profissional: sumário produzido a partir do texto-fonte por um especialista em técnicas de sumarização;
- extracto ideal: sumário composto somente por frases mais representativas do texto-fonte.

Há várias formas para a construção de sumários de referência. Os sumários de referência escritos por humanos reflectem toda a subjectividade do indivíduo e os elaborados por ferramentas automatizadas terão seu conteúdo influenciado pelas características da arquitectura do sistema que os gerou. É usual elaborar os extractos ideais a partir da medida do co-seno (Salton, 1989 apud Pardo, Rino e Nunes, 2003) procurando no texto-fonte as frases com maior grau de semelhança com frases do sumário autêntico. O uso de sumários autênticos e sumários profissionais como dados de referência pode dificultar a comparação entre os extractos e os sumários de referência, pois estes últimos geralmente não preservam as frases dos textos-fonte da forma que elas ocorrem.

O extracto ideal é o melhor tipo de sumário de referência para a avaliação de sistemas de sumarização automática (Rino e Pardo, 2003), pois, por conter somente frases do texto-fonte, pode ser comparado mais facilmente com um sumário automático. No caso da utilização de extractos, a comparação com o extracto ideal pode ser automatizada; no caso de sumários autênticos e profissionais, podem ser necessárias etapas de revisão humana após o processamento.

A comparação pode ser automatizada adoptando as métricas precisão e cobertura, amplamente utilizadas em tarefas de recuperação de informação. Podem ainda ser utilizadas métricas derivadas dessas duas como, por exemplo, a *f-measure*. Estas métricas são aplicáveis, preferencialmente, a extractos ideais como sumários de referência e a sistemas de sumarização automática extractivos. A precisão (P) e a cobertura (do inglês, *recall*) (R) são dadas pelas seguintes fórmulas:

$$P = \frac{\text{número de frases do sumário automático presentes no sumário de referência}}{\text{número de frases do sumário de automático}}$$

$$R = \frac{\text{número de frases do sumário automático presentes no sumário de referência}}{\text{número de frases do sumário de referência}}$$

A precisão indica o número de frases do sumário de referência que o sumário automático possui em relação a todas as frases que ele contém. A cobertura indica quantas frases do sumário de referência o sumário automático possui em relação a todas as frases que deveria conter. Outra medida, a *f-measure*, combina as medidas de precisão e cobertura, resultando numa medida única de eficiência do sistema: quanto mais próxima essa medida for de um, maior a capacidade do sistema em produzir sumários ideais. A fórmula da *f-measure* é a seguinte:

$$f - measure = \frac{2 \times P \times R}{P + R}$$

(Jing et al., 1998) demonstraram que os diferentes parâmetros das experiências podem influenciar profundamente a pontuação dos sistemas de sumarização. Alguns dos parâmetros investigados foram:

- concordância entre os juízes para a elaboração de sumários de referência;
- tamanho do sumário automático;
- a influência da formulação das métricas de precisão e cobertura;
- nível de dificuldade das perguntas para avaliações do tipo pergunta/resposta;
- características dos textos.

(Jing et al., 1998) observaram ainda que na avaliação baseada em sumários ideais, a validade da avaliação diminui na medida em que se aumenta o tamanho dos sumários. Os autores destacam que a comparação das medidas de precisão e cobertura entre diferentes sistemas pode não ser válida, dado que estas medidas dependem da estrutura dos textos utilizados e das diferentes estratégias utilizadas para calcular o tamanho dos sumários (em função da taxa de compressão especificada).

Quando são utilizados sumários de referência para a avaliação da informatividade de sumários automáticos, deve ser tido em conta se o sumário de referência é ou não adequado. Os sumários autênticos, por exemplo, podem conter informação não apresentada no texto-fonte ou mesmo ser pouco informativos. Nesses casos, a comparação fica prejudicada, já que não existem mecanismos de compreensão para concluir por um factor comum entre variações. É importante, portanto, seleccionar correctamente as fontes utilizadas para a avaliação.

4.1.2 Avaliação Extrínseca

A avaliação extrínseca tem como objectivo avaliar um sumarizador através da realização de tarefas específicas. (Mani, 2001) refere o tipo de tarefas que geralmente utilizam este tipo de avaliação:

- categorização de documentos: leitores humanos devem, após a leitura dos sumários, atribuir uma categoria ou classe aos textos. Na situação ideal, espera-se que a taxa de classificações correctas não se degrade e que o tempo necessário para a classificação diminua gradualmente;

- recuperação de informação: é realizada uma pesquisa numa base de textos. Dado um assunto, são retornados como resultado os textos cujo assunto coincida com o solicitado. Nessa avaliação, a pesquisa, que pode ser automática ou manual, é realizada utilizando-se os sumários em lugar das versões completas dos textos. O sucesso da pesquisa é analisado por juízes humanos verificando-se a taxa de textos correctamente seleccionados e o tempo da pesquisa. De forma semelhante à categorização, espera-se manter a taxa de textos correctamente seleccionados e reduzir o tempo de pesquisa;

- sistemas de perguntas e respostas: a informatividade dos sumários é avaliada. A partir de uma base de textos, são elaboradas perguntas de escolha múltipla para cada texto. De seguida, o sumarizador é utilizado para gerar os sumários correspondentes. Por fim os juízes humanos deverão responder as mesmas perguntas em três situações: sem a leitura dos textos originais nem dos sumários, lendo apenas os sumários e, finalmente, lendo os textos completos. Se os sumários forem suficientemente informativos, espera-se que os juízes sejam capazes de responder as perguntas lendo apenas estes;

Como acontece com a avaliação intrínseca, a avaliação extrínseca sofre também de um conjunto de problemas. Em geral estas avaliações são custosas, por dependerem de juízes humanos. É difícil utilizar textos longos dado que estes devem ser lidos pelos juízes em tempo útil. Este tipo de avaliação não fornece qualquer informação sobre que tipo de melhorias que podem ser introduzidas nos sistemas de sumarização automática, dado que são avaliados indirectamente, através de tarefas nas quais estão inseridos. Por vezes pode ser difícil criar tarefas extrínsecas que modelem adequadamente as situações do mundo real e, ao mesmo tempo, sejam passíveis de medição e possíveis de serem realizadas por juízes humanos.

4.2 Experiências e Resultados

Neste tipo de sistemas é usual a realização de avaliação comparativa da sua performance com os outros sistemas existentes no mercado. Com esse objectivo em mente, pretende-se um método totalmente automatizado para a avaliação do trabalho apresentado nesta tese.

À primeira vista a solução será utilizar a metodologia intrínseca com as métricas de co-selecção (Radev et al., 2003), mais especificamente precisão, cobertura e *f-measure*. Uma observação mais cuidada do sistema, revela que não é possível seguir esta abordagem dado que o sistema desenvolvido não é extractivo. Tal

como apresentado anteriormente este sistema comprime as várias frases (removendo informação supérflua). Dado que nenhum dos sistemas existentes para a língua portuguesa segue uma abordagem abordagem similar, não é possível fazer uma comparação directa entre estes e o sistema apresentado neste trabalho.

4.2.1 Outros Sistemas para a Língua Portuguesa

Com o intuito de apresentação completa dos sistemas para a língua portuguesa é, de seguida, discutida a sua avaliação. Alguns dos sistemas de sumarização automática referidos nesta tese não estão completamente disponíveis, revela-se impossível proceder a testes profundos. Devido a esta limitação, são utilizados os resultados de (Rino et al., 2004). Os autores produziram um trabalho comparativo dos sistemas: TF-ISF-Summ (Neto et al., 2000), GistSumm (Pardo, 2002c), NeuralSumm (Pardo, Rino e Nunes, 2003), ClassSumm (Neto, Freitas e Kaestner, 2002) e SuPor (Rino e Módolo, 2004). O sistema DMSumm (Pardo, 2002a) não foi analisado pelos autores é necessário recorrer à avaliação presente no trabalho que o apresenta.

No seu trabalho (Rino et al., 2004) procederam a uma avaliação do tipo *black-box* totalmente automatizada. Utilizaram o TeMário (Pardo e Rino, 2003)[3], um corpus de 100 textos jornalísticos (cerca de 613 palavras, ou de 1 a 2 páginas e meia), que foi construído propositadamente para avaliação de sistemas de sumarização automáticas. Os textos foram retirados de jornais brasileiros *online*, a Folha de São Paulo (60 textos) e o Jornal do Brasil (40 textos). Os textos são distribuídos equitativamente entre vários domínios, ou seja, são artigos dedicados aos mais diversos assuntos como: mundo, política e relações exteriores. Os sumários presentes no pacote são produzidos manualmente por um consultor especialista em português brasileiro. Estão ainda presentes extractos ideias, os quais são criados automaticamente por um gerador de extractos ideais[4]. Esta ferramenta é baseada num modelo vectorial e utiliza a medida de similaridade do co-seno (Salton, 1989 apud Pardo, Rino e Nunes, 2003). Os autores referem que não foi possível comparar os extractos automáticos com os resumos manuais presentes no TeMário, porque eles são produzidos manualmente não permitindo uma avaliação automática viável. Assim, foram utilizados os extractos ideais correspondentes.

Para evitar enviesamento nos sistemas que necessitavam de treino os autores fizeram uma validação cruzada 10 vezes (cada volta composta por 10 textos). A

[3]disponível em http://www.linguateca.pt/Repositorio/TeMario
[4]disponível em http://www.nilc.icmc.usp.br/thiago

taxa de compressão escolhida foi de 30%. Foi seleccionada por estar de acordo com os tamanhos de ambos, os sumários manuais e os extractos ideias (o comprimento destes varia de 25% a 30%). Todos os sistemas foram executados de forma independente. A tabela 4.1 mostra os valores médios para a precisão, a cobertura e a *f-measure* registados nas várias experiências.

Sistema	Precisão	Cobertura	*F-measure*
SuPor	44.9	40.8	42.8
ClassSumm	45.6	39.7	42.4
TF-ISF-Summ	39.6	34.3	36.8
GistSumm	49.9	25.6	33.8
NeuralSumm	36.0	29.5	32.4

Tabela 4.1: Performance média dos sistemas extractivos (em percentagem).

Relativamente ao sistema DMSumm foi necessário recorrer à avaliação realizada pelo autor. (Pardo, 2002a) executou várias experiências para avaliar o DMSumm, focando-se, principalmente, nas premissas básicas do sistema, isto é, a satisfação do objectivo comunicativo e a preservação da proposição central. O autor realizou ainda uma outra avaliação utilizado as métricas de co-selecção. Nas suas experiências o autor utilizou o Theses Corpus (Pardo, 2002b). Este corpus é composto por 10 introduções de teses e dissertações da área das ciências da computação, contendo, em média, 530 palavras por cada introdução. O autor refere que o corpus foi escolhido pelo facto dos textos apresentarem a estrutura problema-solução e serem acompanhados por sumários autênticos, isto é, produzidos pelos próprios autores dos textos. A tabela 4.2 apresenta os dados da avaliação.

Sistema	Precisão	Cobertura	*F-measure*
DMSumm	44	54	48

Tabela 4.2: Performance média do DMSumm (em percentagem).

Como refere (Jing et al., 1998) o a comparação das medidas de precisão e cobertura entre diferentes sistemas pode não ser válida, em função da estrutura dos

textos utilizados e das diferentes estratégias utilizadas para calcular o tamanho dos sumários e em função da taxa de compressão especificada. A comparação directa não pode ser realizada entre o sistema DMSumm e os outros sistemas visto que utilizaram corpus diferentes.

4.2.2 Sistema Descrito

Tal como foi referido anteriormente, devido a abordagem inovadora do sistema apresentado nesta tese, não é possível utilizar as métricas co-selecção. Todavia, com o objectivo de avaliar a qualidade dos sumários temos de optar por uma avaliação intrínseca recorrendo a avaliadores humanos. O corpus utilizado na avaliação é composto por dez textos escritos em português (cinco em português europeu e cinco em português do Brasil). Estes textos são artigos jornalísticos retirados das edições de 2004 e 2005 do Jornal Público e Folha de São Paulo. A escolha de textos jornalísticos justifica-se com a correcção e qualidade da escrita em português neste tipo de literatura.

Para cada texto são gerados dois sumários, o primeiro (denominado sumário A) apenas utilizando os nós de primeiro nível das árvores DTS, o segundo (denominado sumário B) utilizando a agregação dos nós de primeiro e segundo nível. Na figura 4.1 são apresentados os dois sumários gerados para o texto Jornal Público-19950726-079. No apêndice C podem ser encontrados os vários textos que compõem o corpus de teste e respectivos sumários.

O grupo de dezasseis avaliadores é composto por estudantes da Universidade de Évora, de diferentes áreas do conhecimento, todos falantes nativos de português europeu. Seguindo alguns dos critérios de (Saggion e Lapalme, 2000; Pardo, 2002a) foi solicitado aos avaliadores que atribuíssem notas observando:

- a legibilidade - observar a ortografia e a correcta escrita das várias palavras;
- qualidade textual - verificar se as frases estão relacionadas entre si, sem mudanças bruscas de assunto perfazendo um conjunto coerente;
- identificação adequada do assunto - verificar se através a leitura do resumo é possível perceber qual é o assunto/assunto do texto.

Os resultados da avaliação estão explanados nas tabelas 4.3 , 4.4 e 4.5 para a legibilidade, qualidade textual e identificação adequada do assunto respectivamente. Os valores apresentados são referentes ao número de avaliadores que escolheram uma dada avaliação.

Sumário A

O abastecimento público de água em o concelho de Elvas degradou-se ultimamente.
A seca prolongada, as temperaturas de os últimos dias e o aumento de o consumo em esta época de o ano foram as razões indicadas para justificar o agravamento de as condições de o abastecimento público.
A câmara nomeou, entretanto, um grupo de trabalho.

Sumário B

O abastecimento público de água em o concelho de Elvas degradou-se ultimamente a ponto de levar a Câmara local a cortar a água entre as 22h30 e as 06h00 de terça-feira, que o objectivo de a medida é o de distribuir por toda a população as dificuldades de a falta de água. A seca prolongada, as temperaturas de os últimos dias e o aumento de o consumo em esta época de o ano foram as razões indicadas para justificar o agravamento de as condições de o abastecimento público. A câmara nomeou, entretanto, um grupo de trabalho com a finalidade de acompanhar diariamente o abastecimento e pôr em prática um novo sistema.

Figura 4.1: Exemplo dos sumários A e B para o texto Jornal Público-19950726-079.

Como se pode observar, à primeira vista, os avaliadores têm um elevando grau de concordância em alguns textos mas, no entanto noutros têm um grau muito baixo. Numa observação cuidada dos textos em que a concordância é baixa observa-se que nestes ocorreram alguns de processamento. Alguns juízes não interpretaram correctamente os critérios de avaliação como se pode observar na tabela 4.3, neste critério era solicitado a avaliação da legibilidade (por legibilidade entendia-se a correcta formação das palavras).

Para retirar maiores ilações deste dados é necessário aferir o grau de concordância entre os juízes para os diversos critérios. Na tabela 4.6 são apresentados os resultado para o calculo da medida Kappa[5](Fleiss, 1971).

[5]Fleiss kappa (criada por Joseph L. Fleiss) é uma medida estatística que avalia a concordância entre um número fixo de avaliadores ao atribuir classificações a uma série de itens.

Texto	Sumário A				Sumário B			
	mau	razoável	bom	excelente	mau	razoável	bom	excelente
p19940101-007	0	1	3	12	0	0	2	14
p19950726-079	0	0	1	15	0	0	1	15
p19950725-025	0	2	2	12	0	0	3	13
p19950422-141	0	4	1	11	0	0	1	15
p19950912-022	0	2	2	12	0	0	2	14
fs940101-132	0	1	4	11	0	0	3	13
fs950101-011	1	2	1	12	0	0	3	13
fs940101-085	0	0	2	14	0	0	1	15
fs940101-074	0	0	0	16	0	0	0	16
fs950111-014	0	0	0	15	0	0	0	16

Tabela 4.3: Avaliação da Legibilidade dos Sumários.

Texto	Sumário A				Sumário B			
	mau	razoável	bom	excelente	mau	razoável	bom	excelente
p19940101-007	0	7	7	2	0	2	7	7
p19950726-079	0	7	7	2	0	1	8	7
p19950725-025	9	4	3	0	1	3	7	5
p19950422-141	5	4	5	2	0	0	7	9
p19950912-022	4	5	3	4	0	1	8	7
fs940101-132	3	10	3	0	1	0	6	9
fs950101-011	6	5	3	2	0	2	8	6
fs940101-085	0	3	9	4	0	0	6	10
fs940101-074	0	2	9	5	0	0	6	10
fs950111-014	0	2	10	4	0	0	4	12

Tabela 4.4: Avaliação da qualidade textual.

Texto	Sumário A				Sumário B			
	mau	razoável	bom	excelente	mau	razoável	bom	excelente
p19940101-007	0	2	11	13	0	0	4	12
p19950726-079	0	3	7	6	0	0	0	16
p19950725-025	5	3	2	6	0	2	2	12
p19950422-141	2	4	4	6	0	0	3	13
p19950912-022	6	5	4	1	0	0	1	15
fs940101-132	5	3	3	5	0	1	4	11
fs950101-011	3	4	5	4	0	0	7	9
fs940101-085	0	1	4	11	0	0	1	15
fs940101-074	0	1	5	10	0	0	2	14
fs950111-014	0	1	8	7	0	0	1	15

Tabela 4.5: Avaliação da identificação adequada do assunto.

	Legibilidade	Qualidade Textual	Identificação assunto
Sumário A	0.693	0.339	0.335
Sumário B	0.818	0.433	0.723

Tabela 4.6: A medida Kappa representa o nível de concordância entre os avaliadores.

Os juízes têm um grau de concordância maior relativamente ao sumário B, devido provavelmente à sua proximidade ao texto original. Foi observada uma grande discordância entre os avaliadores em alguns textos, possivelmente devido aos juízes serem oriundos de diferentes áreas do conhecimento. Este facto origina que as suas aspirações, conhecimento e modo de interpretar os textos seja muito diferentes entre si. Por exemplo, no sumário A do texto Jornal Público-19950912-022, relativamente à identificação adequada do assunto seis avaliadores atribuíram mau, cinco razoável, quatro bom e um excelente. Do grupo de avaliadores apenas um juiz era especialista em língua portuguesa, para o caso em questão atribuiu mau.

Seria interessante repetir a avaliação apenas com juízes especialista em linguística, certamente que o nível de concordância seria muito maior. Apenas por motivos de completude e não querendo generalizar a partir de um único juízo apresenta-se, na tabela 4.7, a avaliação realizada pelo juiz especialista em linguística.

Texto	Legibilidade		Qualidade Textual		Identificação assunto	
	Sum. A	Sum. B	Sum. A	Sum. B	Sum. A	Sum. B
p19940101-007	excelente	excelente	excelente	excelente	bom	excelente
p19950726-079	excelente	excelente	bom	bom	excelente	excelente
p19950725-025	excelente	excelente	razoável	excelente	excelente	excelente
p19950422-141	excelente	excelente	bom	excelente	excelente	excelente
p19950912-022	excelente	excelente	excelente	bom	mau	excelente
fs940101-132	excelente	excelente	razoável	excelente	excelente	excelente
fs950101-011	excelente	excelente	mau	razoável	mau	excelente
fs940101-085	excelente	excelente	razoável	bom	excelente	excelente
fs940101-074	excelente	excelente	excelente	bom	excelente	excelente
fs950111-014	excelente	excelente	bom	excelente	razoável	excelente

Tabela 4.7: Avaliação atribuída por um especialista em língua portuguesa.

Uma componente que também merece ser avaliada, é a taxa de compressão que o sistema consegue obter. Na tabela 4.8 são apresentadas as taxas de compressão (face ao texto original) para os sumários gerados automaticamente.

É de notar, que devido à métrica utilizada na compressão ser o número de palavras, no sumário B, por vezes, ocorre expansão em vez de compressão do texto. Este efeito deve-se à separação que o Palavras faz do artigo e da proposição durante o seu processamento. Por exemplo, no texto Jornal Público-19940101-007 temos "Ao longo de a sua carreira" em vez de "Ao longo da sua carreira". É válido assumir que a taxa de compressão seria ligeiramente maior se fosse realizado um pré-processamento ao sumário, que realizasse esta operação de contracção.

O sumário B é bastante similar ao texto original, dado que as regras de segmentação e selecção do texto (Leal, 2008) utilizadas para gerar as DTS privilegiam a colocação dos segmentos nos primeiros e segundos níveis.

Texto	Original	Sumário A	Sumário B	Compressão A	Compressão B
p19940101-007	89	41	95	54%	-6%
p19950726-079	115	56	111	52%	4%
p19950725-025	57	19	51	67%	11%
p19950422-141	78	35	83	56%	-6%
p19950912-022	96	57	82	41%	15%
fs940101-132	54	24	51	56%	6%
fs950101-011	126	51	123	60%	3%
fs940101-085	128	89	136	31%	-6%
fs940101-074	134	95	138	30%	-2%
fs950111-014	82	62	87	25%	-6%

Tabela 4.8: A tabela apresenta o número de palavras no documento original, respectivos sumários e a taxa de compressão.

Observando as tabelas de resultados é visível que os juízes tiveram um grau de concordância maior nos textos Jornal Público-19940101-007, Jornal Público-19950726-079, Jornal Folha-940101-085, Jornal Folha-940101-074 e Jornal Folha-950111-014. Estes foram também os textos em que ocorreram menos problemas de processamento.

É importante referir que alguns do erros (não foi quantificado o número mas pela análise empírica parece-nos mais de metade), que ocorrem na geração de sumários se devem a problemas na análise gramátical, isto é, no Palavras. Por exemplo, no texto Jornal Público-19950725-025 a primeira frase do sumário é: "Disseram os operadores." Este erro deve-se a uma falha de classificação do nível do segmento, proveniente do Palavras, como pode ser observado na figura 4.2. O palavras apresenta problemas com o processamento de textos quando é utilizada a forma passiva, como por exemplo na frase "No caso dos transportes, por exemplo, um dos setores mais visados no relatório, o sobrepreço médio na construção de estradas, segundo a CEI, é de 40%" do texto Jornal Folha-950101-011.

O sistema apresentado nesta tese, apesar de ser baseado numa abordagem completamente diferente do que é habitual nesta área, conseguiu resultados satisfatórios quando comparado com outros sistemas para língua portuguesa. Apenas em três textos os avaliadores consideraram que foi difícil a aferição do assunto

```
(restante conteúdo omitido)
==ADVL:pp
====H:prp('em' <sam->)    em
====P<:np
======>N:art('o' <artd> <-sam> F P)        as
====H:n('sessão' <occ> F P)      sessões
====N<:adj('anterior' M/F P)     anteriores
==,
=P:v-fin('dizer' PS/MQP 3P IND) disseram <--- ERRO
=SUBJ:np
==>N:art('o' <artd> M P)         os
==H:n('operador' <Hprof> M P)    operadores
=.
```

Figura 4.2: Exemplo de um erro de processamento do Palavras no texto Jornal Público-19950725-025, neste caso falha na identificação do nível.

através da leitura do sumário e, também apenas em três a qualidade textual do sumário era reduzida. Como era de esperar, o sumário B obteve classificações melhores do que o sumário A respectivo, visto que foi retirada menos informação. As taxas de compressão conseguidas no sumário A, que se cifram no intervalo 30% a 67% (com média de 47%) são boas para os resultados de legibilidade, qualidade textual e identificação adequada do assunto. Nos resultados, é visível que o sistema obtém melhor performance nos textos escritos em português do Brasil, este facto deve-se ao sistema Palavras ter sido desenhado utilizando esta variante do português.

4.3 Resumo do Capítulo

Neste capítulo foi discutida a avaliação do sistema. Inicialmente foi apresentada uma contextualização teórica referindo os problemas relativos à avaliação de sistemas de sumarização automática e definidos os vários tipos de avaliação existentes. Foram referidos os resultados deste e de outros sistemas para a língua portuguesa. Apresentou-se a comparação de performance dos vários sistemas desta área também focados na língua portuguesa. Por fim foi discutido o método de avaliação do sistema implementado nesta tese e apresentados os resultados.

No próximo capitulo será revisto todo o conteúdo desta tese, discutidas as

vantagens do sistema aqui apresentado e possíveis caminhos futuros para o desenvolvimento deste projecto.

Capítulo 5

Conclusão

O processamento da língua natural é um problema antigo e complexo da área de inteligência artificial. Sendo a sumarização automática uma das áreas mais investigadas desde a década de 1960 no processamento de língua natural. Os estudos de sumarização automática descrevem o desenvolvimento e avaliação de sistemas destinados à geração automática de resumos de textos, os sumários.

A taxa de crescimento da informação devido à *World Wide Web* criou a necessidade de desenvolver sistemas de sumarização eficientes e precisos. Embora a investigação na área tenha começado à mais de 50 anos, há ainda um longo caminho a percorrer neste domínio. Ao longo do tempo, o foco da sumarização mudou do resumo de artigos científicos para notícias, mensagens de correio electrónico e *blogs*.

A recente popularidade de sistemas de resumos de notícias confirma que os sistemas discutidos nesta tese estão cada vez mais em voga. Embora a abordagem fundamental constitua uma proposta mais interessante, a programação dos componentes para a geração automática ainda representa um grande desafio para os investigadores, tornando mais viável a exploração dos métodos extractivos. É importante referir que qualquer estratégia de sumarização automática deve levar em consideração a finalidade dos sumários.

Esta tese apresentou as várias abordagens para a sumarização automática descrevendo a teoria e vários sistemas reais, quer extractivos quer profundos. Foi ainda feita uma apresentação de seis sistemas para a língua portuguesa, cinco superficiais: TF-ISF-Summ, GistSumm, NeuralSumm, ClassSumm, SuPor e um profundo: DMSumm. Foi ainda contextualizada toda a teoria utilizada na elaboração deste trabalho, que se inspirou em duas grandes fontes teóricas: a teoria RST e o projecto AuTema-Dis desenvolvido por Ana Luísa Leal.

A RST apresenta-se como uma base descritiva para estudar as relações entre as frases de um texto em termos funcionais, utilizando a distinção entre as proposições núcleo e satélite e a sua hierarquia. Este sistema é capaz de representar a estrutura hierárquica e o princípio central da organização de um texto.

O projecto AuTema-Dis define uma arquitectura que, implementada computacionalmente, realiza a análise textual, considerando as informações mais relevantes dispostas na superfície de um texto, bem como, as relações de significação que se estabelecem entre os elementos linguísticos que a compõe. O objectivo do trabalho foi desenvolver uma base metodológica, cuja sua sistematização fosse capaz de reconhecer a informação principal num determinado discurso, reorganizar as estruturas relevantes ao tema numa estrutura genérica de fácil processamento, atribuir automaticamente algumas relações retóricas entre os segmentos e subsegmentos e finalmente, utilizando toda a informação gerada pelos processos iniciais, produzir automaticamente uma estrutura sintética em língua natural do texto inicial.

Foi dedicado um capítulo à arquitectura do sistema e sua respectiva implementação. O projecto segue a metodologia do sistema AuTema-Dis contendo quatro etapas de análise distintas, mas relacionadas entre si. Cada módulo consiste numa das etapas de análise e o resultado da execução fornece dados para a execução da etapa seguinte, até a conclusão de todo o processo. São definidos quatro módulos básicos para a realização da análise: identificação e segmentação dos constituintes textuais, organização em árvore constituintes textuais utilizando DTS, identificação das relações retóricas entre os segmentos e finalmente a geração do sumário. Este sistema utiliza a abordagem fundamental, para cada frase é processado o seu conteúdo e retirada toda a informação supérflua. O sumário final é construído pela agregação de todas as frases comprimidas do texto original. Há que referir que esta abordagem é totalmente inovadora, não existe nenhum sistema para a língua portuguesa que siga uma abordagem similar.

A avaliação é um tema de grande importância para a sumarização automática. Só recorrendo à avaliação se consegue verificar o estado da arte e definir novas técnicas de sumarização ou melhorias para as existentes. Conferências internacionais como a TIPSTER SUMMAC e a TAC demonstram o grande interesse na investigação dedicada à avaliação da sumarização automática. A avaliação pode ser do tipo intrínseco ou extrínseco. A avaliação intrínseca foca-se na qualidade e informatividade dos sumários. Na avaliação extrínseca, o sumarizador é avaliado mediante a realização de outras tarefas, como categorização de documentos, recuperação de informação e perguntas e respostas. Tal como ocorre na produção dos sumários, o processo de avaliação deve levar em consideração as necessidades

e características dos utilizadores ou tarefas para os quais o sumário foi produzido. Nem sempre um sumário com baixa qualidade no seu fluxo textual deve ser tomado como insucesso. Em alguns casos pode ser necessário apenas que os sumários preservem as informações essenciais do texto-fonte.

Na secção de avaliação foram discutidas algumas tendências na avaliação de sistemas de sumarização automática. Claramente se percebe, que o futuro desta área de investigação depende fortemente da capacidade de desenvolvimento de formas eficazes de avaliação automática destes sistemas e, da definição de métricas suficientemente objectivas para a comunidade científica as aceitar. O sistema apresentado não é extractivo e utiliza uma abordagem totalmente diferente de todos os outros existentes para a língua portuguesa, assim não foi possível efectuar uma comparação directa com os resultados destes. Foi necessário recorrer a juízes humanos (todos eles falantes nativos da língua portuguesa) para procederem a uma avaliação qualitativa dos resultados deste sistema de sumarização.

Foram apresentados resultados para diversos textos, quer em português europeu quer em português do Brasil. Os resultados foram bastante interessantes apesar de ainda ser possível remover alguma informação supérflua. Apenas em três textos os juízes consideraram que foi difícil a aferição do assunto através da leitura do sumário e, também apenas em três a qualidade textual do sumário era reduzida. As taxas de compressão conseguidas no sumário A, que se situam no intervalo 30% a 67% (com média de 47%) são bastante boas para os resultados de legibilidade, qualidade textual e identificação adequada do assunto. Nos resultados, é visível que o sistema obtém melhor performance nos textos escritos em português do Brasil, este facto deve-se ao sistema Palavras ter sido desenhado utilizando esta variante do português.

Uma grande contribuição deste projecto é o sistema de segmentação, este faz principalmente uso de marcadores semânticos auxiliados por alguns marcadores sintácticos para a realização da segmentação, conseguido resultados bastante atractivos neste domínio.

Um possível trabalho futuro para este projecto, que terá grande impacto na performance, será a utilização da abordagem estatística para executar um pré-processamento do texto-fonte de forma a identificar as frases (e respectivos segmentos e subsegmentos) que realmente contêm informação relevante. Finalmente a estas frases deverá ser aplicado o sistema actual. Desta forma, estaríamos a utilizar o "melhor de dois mundos". No sumário final estaria apenas a informação relevante das frases mais importantes do texto-fonte. Esta extensão é relativamente fácil de implementar e tudo indica que teria resultados bastante atractivos.

É de notar que este sistema é modular, facilmente adaptável para outras línguas, bastando para tal a alteração do módulo de processamento sintáctico e a adaptação das regras de segmentação/atribuição de relações retóricas.

Bibliografia

Barzilay, Regina e Michael Elhadad (1997). «Using lexical chains for text summarization». Em: *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–17. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.6428.

Baxendale, P. B. (out. de 1958). «Machine-made index for technical literature: an experiment». Em: *IBM J. Res. Dev.* 2 (4), pp. 354–361. ISSN: 0018-8646. DOI: http://dx.doi.org/10.1147/rd.24.0354. URL: http://dx.doi.org/10.1147/rd.24.0354.

Bick, Eckhard (2000). *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bratko, Ivan (set. de 2000). *Prolog Programming for Artificial Intelligence*. 3rd. Addison Wesley. ISBN: 0201403757. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0201403757.

Burges, Chris et al. (2005). «Learning to rank using gradient descent». Em: *Proceedings of the 22nd international conference on Machine learning*. ICML '05. Bonn, Germany: ACM, pp. 89–96. ISBN: 1-59593-180-5. DOI: http://doi.acm.org/10.1145/1102351.1102363. URL: http://doi.acm.org/10.1145/1102351.1102363.

Carlson, L. e Daniel Marcu (2001). *Discourse Tagging Reference Manual*. Rel. téc. ISITR545. ISI Technical Report.

Colmerauer, Alain e Philippe Roussel (mar. de 1993). «The birth of Prolog». Em: *SIGPLAN Not.* 28 (3), pp. 37–52. ISSN: 0362-1340. DOI: http://doi.acm.org/10.1145/155360.155362. URL: http://doi.acm.org/10.1145/155360.155362.

Covington, Michael A. (1993). *Natural Language Processing for PROLOG Programmers*. 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR. ISBN: 0136292135.

Dijk, T. A. Van (1972). *Some Aspects of Text Grammars*. The Hague:Mouton.

Dijk, T. A. Van (1992). *Cognição, discurso e interação*. São Paulo: Contexto.

— (1993). *Texto y contexto. Semántica y pragmática del discurso*. Cátedra : Madrid.

Earl, Lois L. (1970). «Experiments in automatic extracting and indexing». Em: *Information Storage and Retrieval* 6.4, pp. 313 –330. ISSN: 0020-0271. DOI: DOI:10.1016/0020-0271(70)90025-2. URL: http://www.sciencedirect.com/science/article/B6X2J-465CW7T-2T/2/4516bfa1857ae5d2b774ad17da4324a6.

Edmundson, H. P. (abr. de 1969). «New Methods in Automatic Extracting». Em: *J. ACM* 16 (2), pp. 264–285. ISSN: 0004-5411. DOI: http://doi.acm.org/10.1145/321510.321519. URL: http://doi.acm.org/10.1145/321510.321519.

Endres-Niggemeyer, Brigitte (1990). «A procedual model of abstracting and some ideias for its implementation». Em: *TKE'90 - Second International Congress on Terminology and Knowledge Engineering* 1 (1), pp. 230–243.

Fleiss, Joseph L. (nov. de 1971). «Measuring nominal scale agreement among many raters.» Em: *Psychological Bulletin* 76 (5), pp. 378–382. DOI: http://psycnet.apa.org/doi/10.1037/h0031619. URL: http://psycnet.apa.org/doi/10.1037/h0031619.

Hutchins, John (1987). «Summarization: Some Problems and Methods». Em: *Meaning: The Frontier of Informatics*. Aslib, pp. 151–173.

Jing, Hongyan et al. (1998). «Summarization Evaluation Methods: Experiments and Analysis». Em: *In AAAI Symposium on Intelligent Summarization*, pp. 60–68.

Jones, Karen Sparck e Julia R. Galliers (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 3540613099.

Kupiec, Julian, Jan Pedersen e Francine Chen (1995). «A trainable document summarizer». Em: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '95. Seattle, Washington, United States: ACM, pp. 68–73. ISBN: 0-89791-714-6. DOI: http://doi.acm.org/10.1145/215206.215333. URL: http://doi.acm.org/10.1145/215206.215333.

Leal, Ana Luísa Varani (set. de 2008). «AuTema-Dis: uma arquitetura computacional para identificação da temática discursiva em textos em Língua Portuguesa». Tese de doutoramento. Universidade de Évora.

Lin, Chin yew (2004). «Rouge: a package for automatic evaluation of summaries». Em: pp. 25–26.

Luhn, H. P. (abr. de 1958). «The automatic creation of literature abstracts». Em: *IBM J. Res. Dev.* 2 (2), pp. 159–165. ISSN: 0018-8646. DOI: http://dx.doi.org/10.1147/rd.22.0159. URL: http://dx.doi.org/10.1147/rd.22.0159.

Lyman, P. e H.R. Varian (out. de 2003). *How Much Information?* URL: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Mani, Inderjeet (2001). *Summarization Evaluation: An Overview*.

Mani, Inderjeet e Mark T. Maybury (1999). *Advances in Automatic Text Summarization*. MIT Press. ISBN: 0-262-13359-8.

Mani, Inderjeet et al. (1999). «The TIPSTER SUMMAC Text Summarization Evaluation». Em: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Bergen, Norway: Association for Computational Linguistics, pp. 77–85. DOI: http://dx.doi.org/10.3115/977035.977047.

Mann, William C. e Sandra A. Thompson (1988). «Rhetorical structure theory: Toward a functional theory of text organization». Em: *Text* 8.3, pp. 243–281.

Marcu, Daniel (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Rel. téc. University of Toronto.

— (1998). «Improving Summarization Through Rhetorical Parsing Tuning». Em: *Proceedings of the Sixth Workshop on Very Large Corpora*. Montreal, Canada, pp. 206–215.

McKeown, Kathleen e Dragomir R. Radev (1995). «Generating summaries of multiple news articles». Em: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '95. Seattle, Washington, United States: ACM, pp. 74–82. ISBN: 0-89791-714-6. DOI: http://doi.acm.org/10.1145/215206.215334. URL: http://doi.acm.org/10.1145/215206.215334.

Miller, George A. *WordNet. a large lexical database for English*. URL: http://wordnet.princeton.edu/.

Neto, Joel Larocca, Alex Alves Freitas e Celso A. A. Kaestner (2002). «Automatic Text Summarization Using a Machine Learning Approach». Em: *SBIA '02: Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*. London, UK: Springer-Verlag, pp. 205–215. ISBN: 3-540-00124-7.

Neto, Joel Larocca et al. (2000). *Document Clustering and Text Summarization*.

Ono, Kenji, Kazuo Sumita e Seiji Miike (1994). «Abstract generation based on rhetorical structure extraction». Em: *Proceedings of the 15th conference on Computational linguistics - Volume 1*. Kyoto, Japan: Association for Computational

Linguistics, pp. 344–348. DOI: http://dx.doi.org/10.3115/991886.991946. URL: http://dx.doi.org/10.3115/991886.991946.

Paice, C. D. (1981). «The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases». Em: *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*. Cambridge, England: Butterworth & Co., pp. 172–191. ISBN: 0-408-10775-8.

Pardo, Thiago e Lucia Rino (out. de 2003). *TeMário: Um Corpus para Sumarização Automática de Textos*. Rel. téc. Universidade de São Carlos.

Pardo, Thiago Alexandre Salgueiro (mar. de 2002a). «Descrição do DMSumm: Um Sumarizador Automático Baseado em um Modelo Discursivo». Em: *Série de Relatórios do Núcleo de Interinstitucional de Lingüística Computacional*.

— (2002b). «DMSumm: Um Gerador Automático de Sumários.» Tese de mestrado. Universidade Federal de São Carlos. São Carlos - SP.

— (set. de 2002c). «GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos». Em: *Série de Relatórios do Núcleo de Interinstitucional de Lingüística Computacional*.

Pardo, Thiago Alexandre Salgueiro, Luciana Helena Machado Rino e Maria das Graças Volpe Nunes (2003). «NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos». Em: *Anais do IV Encontro Nacional de Inteligência Artificial*.

Quinlan, J. Ross (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1-55860-238-0.

Radev, Dragomir R., Eduard Hovy e Kathleen McKeown (2002). «Introduction to the Special Issue on Summarization». Em: *Computational Linguistics* 28.4, pp. 399–408. DOI: 10.1162/089120102762671927. eprint: http://www.mitpressjournals.org/doi/pdf/10.1162/089120102762671927. URL: http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671927.

Radev, Dragomir R. et al. (2003). «Evaluation challenges in large-scale document summarization». Em: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL '03. Sapporo, Japan: Association for Computational Linguistics, pp. 375–382. DOI: http://dx.doi.org/10.3115/1075096.1075144. URL: http://dx.doi.org/10.3115/1075096.1075144.

Rau, Lisa F. e Ron Brandow (1994). «Domain-independent summarization of news». Em: *Summarizing Text for Intelligent Communication*.

Riloff, Ellen, Janyce Wiebe e William Phillips (2005). «Exploiting subjectivity classification to improve information extraction». Em: *Proceedings of the 20th*

national conference on Artificial intelligence - Volume 3. Pittsburgh, Pennsylvania: AAAI Press, pp. 1106–1111. ISBN: 1-57735-236-x. URL: http://portal.acm.org/citation.cfm?id=1619499.1619511.

Rino, Lucia Helena Machado (1996). «Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos». Tese de doutoramento. IFSC-USP. São Carlos.

Rino, Lucia Helena Machado e Marcelo Módolo (2004). «SuPor: An Environment for AS of Texts in Brazilian Portuguese». Em: *Advances in Natural Language Processing*. Ed. por José Luis Vicedo et al. Vol. 3230. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 419–430. URL: http://dx.doi.org/10.1007/978-3-540-30228-5_37.

Rino, Lucia Helena Machado e Thiago Alexandre Salgueiro Pardo (2003). «A Sumarização Automática de Textos: Principais Características e Metodologias». Em: *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. Vol. VIII: III Jornada de Mini-cursos de Inteligência Artificial. Campinas-SP, pp. 203–245.

Rino, Lucia Helena Machado et al. (2004). «A Comparison of Automatic Summarizers of Texts in Brazilian Portuguese». Em: *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA*, pp. 235–244.

Saggion, Horacio e Guy Lapalme (2000). «Concept Identification and Presentation in the Context of Technical Text Summatization». Em: *NAACL-ANLP Workshop on Automatic Summarization*, pp. 1–10.

Salton, Gerard (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0-201-12227-8.

SecuritySpace (jul. de 2007). *PHP: PHP Usage Stats*. URL: http://www.php.net/usage.php.

Skorochod'ko, E. F. (1972). «Adaptive method of automatic abstracting and indexing». Em: *IFIP Congress*. Vol. 71, pp. 1179–1182.

Sparck Jones, K. (1995). «Discourse modelling for automatic summarising». Em: *Travaux du Cercle Linguistique de Prague (Prague Linguistic Circle Papers)* 1 (1), pp. 201–227.

Svore, Krysta M. (2007). *Enhancing Single-document Summarization by Combining RankNet and Third-party Sources*.

Sylvaine, Jean-Luc Minel et al. (1997). «How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN». Em: *Examples of FAN and MLUCE Protocols and*

their Results on SERAPHIN. Intelligent Scalable Text Summarization, Proc. of a Workshop, ACL, pp. 25–30.

W3Techs (out. de 2010). *Usage of server-side programming languages for websites.* URL: `http://w3techs.com/technologies/overview/programming_language/all`.

Apêndice A

Relações Retóricas - RST (Mann e Thompson, 1988)

Definições das relações de apresentação

Relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Antítese	em N: A tem atitude positiva face a N	N e S estão em contraste (cf. a relação de Contraste); devido à incompatibilidade suscitada pelo contraste, não é possível ter uma atitude positiva perante ambas as situações; a inclusão de S e da incompatibilidade entre as situações aumenta a atitude positiva de L por N	A atitude positiva do L face a N aumenta

Continua na próxima página...

Relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Concessão	em N: A possui atitude positiva face a N em S: A não afirma que S não está certo	A reconhece uma potencial ou aparente incompatibilidade entre N e S; reconhecer a compatibilidade entre N e S aumenta a atitude positiva de L face a N	A atitude positiva de L face a N aumenta
Elaboração	em N: apresenta uma acção de L (incluindo a aceitação de uma oferta), não realizada face ao contexto de N	A compreensão de S por L aumenta a capacidade potencial de L para executar a acção em N	A potencial capacidade de L para executar a acção em N aumenta
Evidência	em N: L pode não acreditar em N a um nível considerado por A como sendo satisfatório em S: L acredita em S ou considera-o credível	A compreensão de S por L aumenta a crença de L em N	A crença de L em N aumenta
Fundo	em N: L não compreende integralmente N antes de ler o texto de S	S aumenta a capacidade de L compreender um elemento em N	A capacidade de L para compreender N aumenta
Justificação	nenhuma	A compreensão de S por L aumenta a sua tendência para aceitar que A apresente N	A tendência de L para aceitar o direito de A a apresentar N aumenta

Continua na próxima página...

Relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Motivação	em N: N é uma acção em que L é o actor (incluindo a aceitação de uma oferta), não realizada face ao contexto de N	A compreensão de S aumenta a vontade de L para executar a acção em N	A vontade de L para executar a acção em N aumenta
Preparação	nenhuma	S precede N no texto; S tende a fazer com que L esteja mais preparado, interessado ou orientado para ler N	L está mais preparado, interessado ou orientado para ler N
Reformulação	nenhuma	em N + S: S reformula N, onde S e N possuem um peso semelhante; N é mais central para alcançar os objectivos de A do que S	L reconhece S como reformulação
Resumo	em N: N deve ser mais do que uma unidade	S apresenta uma reformulação do conteúdo de N, com um peso inferior	L reconhece S como uma reformulação mais abreviada de N

Definições das relações de conteúdo

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Alternativa (anti-condicional)	em N: N representa uma situação não realizada em S: S representa uma situação não realizada	realização de N impede a realização de S	L reconhece a relação de dependência de impedimento que se estabelece entre a realização de N e a realização de S
Avaliação	nenhuma	em N + S: S relaciona N com um grau de atitude positiva de A face a N	L reconhece que S confirma N e reconhece o valor que lhe foi atribuído
Causa involuntária	em N: N não representa uma acção voluntária	S, por outras razões que não uma acção voluntária, deu origem a N; sem a apresentação de S, L poderia não conseguir determinar a causa específica da situação; a apresentação de N é mais importante para cumprir os objectivos de A, ao criar a combinação N-S, do que a apresentação de S	L reconhece S como causa de N

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Causa voluntária	em N: N constitui uma acção voluntária ou mesmo uma situação possivelmente resultante de uma acção voluntária	S poderia ter levado o agente da acção voluntária em N a realizar essa acção; sem a apresentação de S, L poderia não perceber que a acção fui suscitada por razões específicas ou mesmo quais foram essas razões; N é mais importante do que S para cumprir os objectivos de A, na criação da combinação N-S	L reconhece S como a causa da acção voluntária em N
Circunstância	em S: S não se encontra não realizado	S define um contexto no assunto, no âmbito do qual se pressupõe que L interprete N	L reconhece que S fornece o contexto para interpretar N
Condição	em S: S apresenta uma situação hipotética, futura, ou não realizada (relativamente ao contexto da situação de S)	Realização de N depende da realização de S	L reconhece de que forma a realização de N depende da realização de S
Condição inversa	nenhuma	S afecta a realização de N; N realiza-se desde que S não se realize	L reconhece que N se realiza desde que S não se realize

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Elaboração	nenhuma	S apresenta dados adicionais sobre a situação ou alguns elementos do assunto apresentados em N ou passíveis de serem inferidos de N, de uma ou várias formas, conforme descrito abaixo. Nesta lista, se N apresentar o primeiro membro de qualquer par, então S inclui o segundo: conjunto :: membro abstracção :: exemplo todo :: parte processo :: passo objecto :: atributo generalização :: especificação	L reconhece que S proporciona informações adicionais a N. L identifica o elemento do conteúdo relativamente ao qual se fornece pormenores
Incondicional	em S: S poderia afectar a realização de N	N não depende de S	L reconhece que N não depende de S
Interpretação	nenhum	em N + S: S relaciona N com várias ideias que não se encontram directamente relacionadas com N, e que não estão relacionadas com a atitude positiva de A	L reconhece que S relaciona N com várias ideias que não se encontram relacionadas com o conhecimento apresentado em N

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Método	em N: uma actividade	S apresenta um método ou instrumento que tende a aumentar as probabilidades de realização de N	L reconhece que o método ou instrumento de S tende a aumentar as probabilidades de realização de N
Propósito	em N: N é uma actividade; em S: S é uma situação que não se encontra realizada	S será realizado através da actividade de N	L reconhece que a actividade em N se inicia para realizar S
Resultado involuntário	em S: S não representa uma acção voluntária	N causou S; a apresentação de N é mais importante para cumprir os objectivos de A, ao criar a combinação N-S, do que a apresentação de S	L reconhece que N poderia ter causado a situação em S
Resultado voluntário	em S: S constitui uma situação ou acção voluntária possivelmente resultante de uma acção voluntária	N pode ter causado S; a apresentação de N é mais importante para cumprir os objectivos de A do que a apresentação de S	L reconhece que N pode ser uma causa da acção ou situação em S
Solução	em S: S apresenta um problema	N constitui uma solução para o problema apresentado em S	L reconhece N como uma solução para o problema apresentado em S

Definições das relações multi-nucleares

Nome da relação	Condições em cada par de N	Intenção de A
Conjunção	Os elementos unem-se para formar uma unidade onde cada um dos elementos desempenha um papel semelhante	L reconhece que os elementos interrelacionados se encontram em conjunto
Contraste	Nunca mais de dois núcleos; as situações nestes dois núcleos são (a) compreendidas como sendo as mesmas em vários aspectos (b) compreendidas como sendo diferentes em alguns aspectos, e (c) comparadas em termos de uma ou mais destas diferenças	L reconhece a possibilidade de comparação e a(s) diferença(s) suscitadas pela comparação realizada
Disjunção	Um dos elementos apresenta uma alternativa (não necessariamente exclusiva) a(s) outra(s)	L reconhece que os elementos inter-relacionados constituem alternativas
Junção	nenhuma	nenhuma
Lista	Um elemento comparável a outros e ligado a outro N através de uma relação de Lista	L reconhece a possibilidade de comparação dos elementos relacionados
Reformulação multi-nuclear	Um elemento constitui, em primeiro lugar, a repetição de outro, com o qual se encontra relacionado; os elementos são de importância semelhante aos objectivos de A	L reconhece a repetição através dos elementos relacionados
Sequência	Existe uma relação de sucessão entre as situações apresentadas nos núcleos	L reconhece as relações de sucessão entre os núcleos

Apêndice B

Relações Retóricas - (Leal, 2008)

Relações definidas por Leal, 2008

Relação	Regras de Segmentação	Restrições e Efeitos	Exemplo
Apositiva de Nome Próprio	App:prop	**Núcleo**: apresenta uma informação nominal pouco específica. **Satélite**: apresenta um Nome Próprio que especifica a informação descrita no núcleo. **Restrições N+S**: o S especifica através de uma expressão nominal própria, relacionada nominalmente ao que foi apresentado pelo N. **Efeito**: o leitor recebe do S a especificação através de um nome próprio daquilo que é mostrado no N.	(...) a posse dos presidentes do Banco do Brasil, **Paulo César Ximenes**, e da Caixa Económica Federal, **Sérgio Cutolo**. FSP950111-034

Continua na próxima página...

Relação	Regras de Segmentação	Restrições e Efeitos	Exemplo
Circunstância de Tempo ou Quantificadora	Advl:adv atemp	**Núcleo**: não há. **Satélite**: apresenta um elemento temporal. **Restrições N+S**: o S apresenta uma característica temporal para a situação descrita no N. **Efeito**: o leitor reconhece quando o fato apresentado pelo N foi realizado.	A inauguração do Mercado Abastecedor de Coimbra (MAC) foi **ontem**, ao fim da tarde, interrompida. Público19950705-167
Circunstância de Lugar ou Quantificadora	Advl:adv aloc	**Núcleo**: não há. **Satélite**: apresenta um elemento lugar. **Restrições N+S**: o S apresenta uma característica locativa para a situação descrita no N. **Efeito**: o leitor reconhece onde o fato apresentado pelo N foi realizado.	(...) Certamente passarei **aqui** a maior parte de Janeiro, (...) FSP950101-084

Continua na próxima página...

Relação	Regras de Segmentação	Restrições e Efeitos	Exemplo
Acção	Sta:icl	**Núcleo**: apresenta uma situação. **Satélite**: demonstra uma acção a ser realizada a partir do que é apresentado pelo N. **Restrições N+S**: a situação apresentada pelo N condiciona a acção descrita pelo S. **Efeito**: o leitor percebe que o S apresenta uma acção realizada ou a ser realizada condicionada pelo que é descrito no N.	(...) a Alemanha deu um passeio, **vencendo fácil com a charmosa dupla Steffi Graf e Michael Stich**. FSP940101-095
Temporal/ Tempo decorrido	Advl:np ou Advl:n	**Núcleo**: não há. **Satélite**: caracteriza uma situação temporal concluída ou em andamento com base no que é apresentado no N. **Restrições N+S**: S situa no tempo (concluído ou não) a informação apresentada no N. **Efeito**: o leitor reconhece a temporalidade do que é descrito no N.	**A certa altura**, uma mulher das Caxinas, que já tinha boné, não estava disposta a deixar Gomes, (...) Publico19950924-121

Apêndice C

Textos e Sumários

C.1 publico-19940101-007.txt

Texto Original

Dominic «Sonny» Constanzo, que acompanhou, com o seu trombone, cantores como Ella Fitzgerald e Tony Bennett, morreu quinta-feira, em New Haven, no estado americano do Connecticut, aos 61 anos, depois de um transplante cardíaco. Ao longo da sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Durante muito tempo acompanhou a vocalista Rosemary Clooney, à frente da sua grande orquestra. Mas apenas em 1992 conseguiu fazer a primeira gravação para uma grande etiqueta, no caso a Stash.

Sumário A

Dominic Sonny Constanzo, morreu. Ao longo de a sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Acompanhou a vocalista Rosemary Clooney. Conseguiu fazer a primeira gravação a Stash.

Sumário B

Dominic Sonny Constanzo, que acompanhou, cantores como Ella Fitzgerald e Tony Bennett, morreu quinta-feira, em New Haven, em o estado americano de o Connecticut, a os 61 anos, depois de um transplante cardíaco. Ao longo de a sua carreira

tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Durante muito tempo acompanhou a vocalista Rosemary Clooney a a frente de a sua grande orquestra . Mas apenas em 1992 conseguiu fazer a primeira gravação para uma grande etiqueta , em o caso a Stash .

C.2 publico-19950726-079.txt

Texto Original

O abastecimento público de água no concelho de Elvas degradou-se ultimamente a ponto de levar a Câmara local a cortar a água entre as 22h30 e as 06h00 de terça-feira, segundo informa uma nota municipal que esclarece, também, que o objectivo da medida é o de «distribuir por toda a população as dificuldades da falta de água». A seca prolongada, as temperaturas dos últimos dias e o aumento do consumo nesta época do ano foram as razões indicadas para justificar o agravamento das condições do abastecimento público. A câmara nomeou, entretanto, um grupo de trabalho com a finalidade de acompanhar diariamente o abastecimento e pôr em prática um novo sistema para reforçar as captações existentes.

Sumário A

O abastecimento público de água em o concelho de Elvas degradou-se ultimamente. A seca prolongada, as temperaturas de os últimos dias e o aumento de o consumo em esta época de o ano foram as razões indicadas para justificar o agravamento de as condições de o abastecimento público. A câmara nomeou, entretanto, um grupo de trabalho

Sumário B

O abastecimento público de água em o concelho de Elvas degradou-se ultimamente a ponto de levar a Câmara local a cortar a água entre as 22h30 e as 06h00 de terça-feira, que o objectivo de a medida é o de distribuir por toda a população as dificuldades de a falta de água. A seca prolongada, as temperaturas de os últimos dias e o aumento de o consumo em esta época de o ano foram as razões indicadas para justificar o agravamento de as condições de o abastecimento público. A

câmara nomeou, entretanto, um grupo de trabalho com a finalidade de acompanhar diariamente o abastecimento e pôr em prática um novo sistema.

C.3 publico-19950725-025.txt

Texto Original

As acções cotadas na praça australiana terminaram em baixa, após a tomada de mais valias em virtude das subidas verificadas nas sessões anteriores, disseram os operadores. O índice da Bolsa de Sidney, o All Ordinaries, perdeu 2,8 pontos durante a sessão e fechou nos 2108,7 pontos, menos 0,13 por cento face ao valor de fecho de sexta-feira.

Sumário A

Disseram os operadores. O índice de a Bolsa de Sidney, perdeu 2,8 pontos e fechou em os 2108,7 pontos.

Sumário B

As acções cotadas em a praça australiana terminaram em baixa, em as sessões anteriores, disseram os operadores. O índice de a Bolsa de Sidney, o All Ordinaries, perdeu 2,8 pontos durante a sessão e fechou em os 2108,7 pontos, menos 0,13 por cento face a o valor de fecho de sexta-feira.

C.4 publico-19950422-141.txt

Texto Original

Acções desvalorizadas. Apesar de ter iniciado a semana passada em alta, a Bolsa de Madrid viria a encerrar este período com uma queda generalizada nas cotações. Terça-feira o mercado abriu um ciclo de perdas, que só viria a encerrar na sexta-feira quando os investidores entenderam que os descontos já tinham sido dados e as cotações voltaram a subir. Nesta sessão o índice Geral fechou nos 276,06 pontos, menos 0,15 por cento face ao último valor da semana anterior.

Sumário A

Acções desvalorizadas. A Bolsa de Madrid viria a encerrar este período em as cotações. O mercado abriu um ciclo de perdas, e as cotações voltaram a subir. O índice Geral fechou em os 276,06 pontos.

Sumário B

Acções desvalorizadas. Apesar de ter iniciado a semana passada em alta, a Bolsa de Madrid viria a encerrar este período com uma queda generalizada em as cotações. Terça-feira o mercado abriu um ciclo de perdas, que só viria a encerrar em a sexta-feira quando os investidores entenderam que os descontos tinham sido dados e as cotações voltaram a subir Em esta sessão o índice Geral fechou em os 276,06 pontos, menos 0,15 por cento face a o último valor de a semana anterior.

C.5 publico-19950912-022.txt

Texto Original

Há mais de dois mil educadores de infância no desemprego. Quem o afirma é a Fenprof num comunicado onde chama a atenção para a degradação da situação destes profissionais num sector onde, de há sete anos a esta parte, existe um congelamento na criação de novos jardins de infância. Segundo a federação, há mais de mil lugares que não vêm a concurso, porque o Ministério da Educação, alheando-se do problema, deixou a colocação dos educadores ao critério das Câmaras municipais. Estas, por sua vez, limitam-se a reconduzir nos lugares os educadores que já lá se encontravam.

Sumário A

Há mais de dois mil educadores de infância. Quem o afirma é a Fenprof em um comunicado. Segundo a federação, há mais de mil lugares porque o Ministério da Educação, deixou a colocação de os educadores a o critério de as Câmaras municipais . Estas, por sua vez, limitam- se a reconduzir em os lugares os educadores.

Sumário B

Há mais de dois mil educadores de infância em o desemprego. Quem o afirma é a Fenprof em um comunicado chama a atenção para a degradação de a situação de estes profissionais em um sector. Segundo a federação, há mais de mil lugares que não vêm a concurso, porque o Ministério da Educação, alheando-se deixou a colocação de os educadores a o critério de as Câmaras municipais. Estas, por sua vez, limitam-se a reconduzir em os lugares os educadores que se encontravam.

C.6 FSP940101-132.txt

Texto Original

O Filho da Pantera Cor de Rosa, com direção de Blake Edwards, estréia hoje na cidade. O filme mostra as atrapalhadas aventuras do filho ilegítimo do inspetor Closeau, na investigação do rapto de uma princesa. No papel principal o ator italiano Roberto Benigni (foto). Em cartaz nos cines Gemini 1, Belas Artes e circuito.

Sumário A

O Filho da Pantera Cor de Rosa, estréia. O filme mostra as atrapalhadas aventuras de o filho ilegítimo de o inspetor Closeau. E circuito.

Sumário B

O Filho da Pantera Cor de Rosa, com direção de Blake Edwards, estréia hoje em a cidade. O filme mostra as atrapalhadas aventuras de o filho ilegítimo de o inspetor Closeau, em a investigação de o rapto de uma princesa. Em cartaz em os cines Gemini 1, Belas Artes e circuito.

C.7 FSP950101-011.txt

Texto Original

A Folha, em editorial na quarta-feira sob o título “Chega de roubalheira”, comenta a apresentação do relatório final da Comissão Especial de Investigação (CEI) sobre suspeitas de irregularidades no Executivo. O editorial afirma que é razoável

imaginar que o relatório não revele mais que a "ponta do iceberg", mas que oferece um mapa inicial da corrupção na administração pública. "No caso dos transportes, por exemplo, um dos setores mais visados no relatório, o sobrepreço médio na construção de estradas, segundo a CEI, é de 40% ", diz o editorial. "O próximo governo será então posto à prova desde o seu início, com o desafio de mostrar, com celeridade e ações concretas, se vai ou não compactuar com o binômio corrupção-impunidade que há tanto sangra o país".

Sumário A

A Folha, comenta a apresentação de o relatório final de a Comissão Especial de Investigação CEI. O editorial afirma que é razoável imaginar que o relatório não revele que a ponta de o iceberg. Diz o editorial. O próximo governo será posto a a prova não compactuar com o binômio corrupção-impunidade.

Sumário B

A Folha, em editorial em a quarta-feira sob o título Chega de roubalheira, comenta a apresentação de o relatório final de a Comissão Especial de Investigação CEI sobre suspeitas de irregularidades em o Executivo. O editorial afirma que é razoável imaginar que o relatório não revele mais que a ponta de o iceberg, mas que oferece um mapa inicial de a corrupção em a administração pública . Por exemplo, um de os setores mais visados em o relatório, é de 40%, diz o editorial . O próximo governo será então posto a a prova desde o seu início, com o desafio de mostrar, com celeridade e ações concretas, se vai ou não compactuar com o binômio corrupção-impunidade que há tanto sangra o país.

C.8 FSP940101-085.txt

Texto Original

Para este ano a maior novidade no setor dos transportes no município deve ser a introdução do sistema de catracas eletrônicas, que poderá gerar demissões nas empresas de transporte, já que os cobradores não serão mais necessários. O sindicato dos condutores é contra a medida e, no ano passado, ameaçou fazer greves em protesto. Com a catraca eletrônica, a compra de bilhetes poderia ser antecipada, como no Metrô, e também permitiria integração gratuita entre uma linha e outra.

O programa de corredores, outra promessa para este ano, teve as primeiras licitações lançadas no final do ano passado. Vai permitir a integração mais rápida dos bairros através da circulação dos ônibus em faixas exclusivas. Se tudo der certo, os corredores devem começar a ser implantados no final do ano.

Sumário A

A maior novidade em o setor de os transportes em o município deve ser a introdução de o sistema de catracas eletrônicas. O sindicato de os condutores é contra a medida e ameaçou fazer greves em protesto. Com a catraca eletrônica, a compra de bilhetes poderia ser antecipada. O programa de corredores, outra promessa teve as primeiras licitações lançadas em o final de o ano passado . Vai permitir a integração mais rápida de os bairros. Os corredores devem começar a ser implantados em o final de o ano.

Sumário B

Para este ano a maior novidade em o setor de os transportes em o município deve ser a introdução de o sistema de catracas eletrônicas, que poderá gerar demissões em as empresas de transporte, já que os cobradores não serão necessários. O sindicato de os condutores é contra a medida e em o ano passado, ameaçou fazer greves em protesto. Com a catraca eletrônica, a compra de bilhetes poderia ser antecipada permitiria integração gratuita entre uma linha e outra. O programa de corredores, outra promessa para este ano, teve as primeiras licitações lançadas em o final de o ano passado. Vai permitir a integração mais rápida de os bairros através de a circulação de os ônibus em faixas exclusivas. Se tudo der certo, os corredores devem começar a ser implantados em o final de o ano.

C.9 FSP940101-074.txt

Texto Original

O shopping Center Norte vai sortear uma viagem ao Caribe. Para concorrer, é preciso trocar notas fiscais recebidas durante as compras em lojas do shopping por cupons. Cada CR$ 5.000,00 em notas vale um cupom, que ficará depositado na urna do shopping até as 18h do dia 30 de janeiro, quando acontece o sorteio. O prêmio é um cruzeiro pelo Caribe, com direito a um acompanhante e todas as

despesas pagas. O slogan da promoção é “Que tal catar coquinho no Caribe?”. Essa é a última viagem sorteada pelo shopping. Todos os meses, desde junho, o Center Norte dá como prêmio uma viagem internacional a seus frequentadores. A promoção vale para as compras feitas a partir de segunda-feira. Os cupons serão trocados no posto do shopping apenas até o dia 29, véspera do sorteio.

Sumário A

O shopping Center Norte vai sortear uma viagem a o Caribe. Para concorrer, é preciso trocar notas fiscais recebidas durante as compras por cupons . Cada Cr$ 5.000,00 em notas vale um cupom. O prêmio é um cruzeiro por o Caribe. O slogan de a promoção é Que tal catar coquinho no Caribe? Essa é a última viagem sorteada por o shopping. O Center Norte dá como prêmio uma viagem internacional a seus frequentadores. A promoção vale para as compras feitas a partir de segunda-feira. Os cupons serão trocados em o posto de o shopping.

Sumário B

O shopping Center Norte vai sortear uma viagem a o Caribe. Para concorrer , é preciso trocar notas fiscais recebidas durante as compras em lojas de o shopping por cupons. Cada Cr$ 5.000,00 em notas vale um cupom, que ficará depositado em a urna de o shopping quando acontece o sorteio. O prêmio é um cruzeiro por o Caribe, com direito a um acompanhante e todas as despesas pagas. O slogan de a promoção é Que tal catar coquinho no Caribe? Essa é a última viagem sorteada por o shopping. Todos os meses, desde junho, o Center Norte dá como prêmio uma viagem internacional a seus frequentadores . A promoção vale para as compras feitas a partir de segunda-feira. Os cupons serão trocados em o posto de o shopping apenas até o dia 29, véspera de o sorteio.

C.10 FSP950111-014.txt

Texto Original

Cerca de 200 policiais procuram no norte de Minas Gerais os fazendeiros Darly Alves e seu filho Darci Alves Pereira. Os dois foram condenados a 19 anos de prisão cada um, em dezembro de 90, pelo assassinato do líder seringueiro Chico Mendes. O crime ocorreu em dezembro de 88 no município de Xapuri (AC). A

polícia de Minas iniciou as buscas em dezembro, após uma denúncia anônima. Segundo informações recebidas pela polícia, os dois estariam escondidos em uma fazenda de difícil acesso.

Sumário A

Cerca de 200 policiais procuram em o norte de Minas Gerais os fazendeiros Darly Alves e seu filho Darci Alves Pereira. Os dois foram condenados a 19 anos de prisão cada um. O crime ocorreu em dezembro de 88 AC. A polícia de Minas iniciou as buscas em dezembro. Por a polícia, os dois estariam escondidos em uma fazenda de difícil acesso.

Sumário B

Cerca de 200 policiais procuram em o norte de Minas Gerais os fazendeiros Darly Alves e seu filho Darci Alves Pereira. Os dois foram condenados a 19 anos de prisão cada um, em dezembro de 90, por o assassinato de o líder seringueiro Chico Mendes. O crime ocorreu em dezembro de 88 em o município de Xapuri AC. A polícia de Minas iniciou as buscas em dezembro, após uma denúncia anônima. Segundo informações recebidas por a polícia, os dois estariam escondidos em uma fazenda de difícil acesso.