UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Sumarização Automática de Texto

Ângelo Filipe da Silva dos Santos

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor João Paulo Cordeiro

Covilhã, Junho de 2012

Agradecimentos

Gostaria em primeiro lugar de agradecer ao Professor Doutor João Paulo Cordeiro pela completa disponibilidade e imensa paciência demonstradas na orientação desta dissertação, pelos ensinamentos transmitidos, pelo apoio e dedicação manifestados.
A todos os amigos e colegas que de alguma forma influenciaram o meu percurso ao longo desta empreitada o meu muito-obrigado.
Um especial agradecimento à minha família, por serem os exemplos de dedicação, força e espírito sacrifício que sempre tentei seguir, por todo o apoio e incentivo, e pela forma como sempre acreditaram em mim.
À minha namorada Sónia, melhor amiga e companheira de aventuras, sempre presente nos melhores e piores momentos, grande amparo em momentos de maior desânimo e maior entusiasta na partilha de momentos de alegria, símbolo de amor, carinho e dedicação, o meu mais sentido agradecimento.

Resumo

O acto de sumarizar ou resumir, isto é, tornar mais sucinta a descrição de uma ideia ou conceito, é uma actividade bastante trivial. As pessoas produzem constantemente, este tipo de representações sucintas para algo que pretendam descrever ou comunicar, sendo que, uma forma muito comum de síntese são os sumários escritos. Tradicionalmente este tipo de sumários são manualmente produzidos por pessoas que analisam textos e tentam identificar os principais conceitos presentes nos mesmos. A chamada "sobrecarga de informação", em muito potenciada pela explosão da Internet, tem instigado a disponibilidade de um cada vez maior volume de informação, que torna esse trabalho manual bastante difícil, senão mesmo impossível. Vários têm sido os esforços realizados na tentativa de resolução deste problema, procurando desenvolver técnicas que possibilitem obter o conteúdo mais relevante de documentos, de maneira condensada, sem alterar o seu significado original, e com a mínima intervenção humana.
O trabalho desenvolvido no âmbito desta dissertação visou explorar diversas abordagens de sumarização extractiva de texto através da implementação de métodos computacionais baseados em estatísticas textuais e teoria de grafos. Foi ainda implementado um método baseado na fusão das abordagens anteriores com outras características como a procura de palavras-chave e a posição das frases no texto, o que resultou na denominação de método híbrido. A sumarização realizada é puramente extractiva, ou seja, a composição do sumário gerado é baseada na classificação das frases do texto original e posterior selecção do subconjunto das frases mais informativas, por forma a satisfazer determinada taxa de compressão.
Numa abordagem puramente estatística, foi desenvolvido um método que pretende avaliar a relevância de termos do texto com base nos valores das suas frequências, no texto fonte e num corpus. A abordagem baseada em teoria de grafos foi utilizada para levar a cabo duas tarefas distintas, a classificação de frases através da avaliação da sua centralidade, e a extracção de palavras-chave. A abordagem híbrida utiliza as várias características descritas numa combinação linear, mediada por um conjunto de pesos associados às diversas componentes.
O desempenho das diferentes abordagens exploradas é avaliado utilizando colecções de textos noticiosos. Estes dados são provenientes das *Document Understanding Conferences* (DUC). Para avaliar a qualidade dos sumários produzidos, foi utilizada a ferramenta ROUGE. Os diversos métodos propostos foram, então, comparados entre si avaliando-se intrínseca e automaticamente o nível de informação dos extractos produzidos. Os resultados obtidos evidenciam que o método híbrido é o que apresenta melhor desempenho aquando da comparação da sua pontuação ROUGE com os demais, ficando esta tendência a dever-se essencialmente à utilização de uma heurística posicional que atribui maior importância a frases que ocupem uma posição cimeira no texto, sendo que este modelo se adequa especialmente bem à estrutura textual de artigos noticiosos.

Palavras-chave

Sumarização Automática de Texto; Sumarização Automática Extractiva; Teoria de Grafos; Estatísticas Textuais; Métodos Híbridos; Relevância de Termos; TextRank; ROUGE.

Abstract

The act of summarizing, i.e., making the depiction of an idea or concept briefer, is a fairly trivial activity. People constantly produce this kind of concise representations for something they wish to describe or communicate, with a very common form of synthesis being written summaries. Traditionally this type of summary is manually produced by people who analyze texts and try to identify the key concepts in them. The so-called "information overload", much boosted by the explosion of the Internet, has instigated the availability of an increasing amount of information that makes this manual work very difficult, if not impossible. There have been made numerous efforts in trying to solve this problem, by developing techniques that allow the most relevant content of documents to be extracted, in a condensed manner, without altering its original meaning, and with minimal human intervention.

The work developed in this thesis aimed to explore several approaches of extractive text summarization through the implementation of computational methods based on textual statistics and graph theory. A hybrid method was also implemented based on the combination of the previous approaches with other features like keyword extraction and sentence position in the text. The summarization done is purely extractive, i.e., the summary is generated based on classification and extraction of sentences from the original text by selecting the most informative subset of the sentences, in order to satisfy certain compression rate.

In a purely statistical approach, we developed a method to evaluate the relevance of the terms of a text based on the values of their frequencies in the source text and in a corpus. The approach based on graph theory was used to carry out two distinct tasks, the classification of sentences by assessing its centrality, and the extraction of keywords. The hybrid approach uses the various features described before in a linear combination mediated by a set of weights associated with the various components.

The performance of the different approaches explored is evaluated using collections of news texts. This data is from the *Document Understanding Conferences* (DUC). To assess the quality of the produced summaries ROUGE package was used. The various methods proposed were compared with each other in an intrinsic and automatic fashion, to evaluate the level of informative content of the extracts produced. The results showed that the hybrid method is the one with better performance when comparing its ROUGE score with the other methods scores, this tendency being mainly due to the use of a positional heuristic that assigns greater importance to sentences that have leading positions in the text, and this model is especially suited to the text structure of news articles.

Keywords

Automatic Text Summarization; Extractive Automatic Summarization; Graph Theory; Textual Statistics; Hybrid Methods; Term Relevance; TextRank; ROUGE.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AA	Aprendizagem Automática
ACL	*Association for Computational Linguistics*
BLUE	*Bilingual Evaluation Understudy*
DUC	*Document Understanding Conference*
EI	Extracção de Informação
HMM	*Hidden Markov Model*
IA	Inteligência Artificial
IBM	*International Business Machines*
LSA	*Latent Semantic Analysis*
MMR	*Maximal Marginal Relevance*
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informação
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
SVD	*Singular Value Decomposition*
TAC	*Text Analysis Conference*
TF-IDF	*Term Frequency - Inverse Document Frequency*
WWW	*World Wild Web*
XML	*Extensible Markup Language*

Capítulo 1

Introdução

O ritmo acelerado a que hoje em dia a informação, e sobretudo a informação sob a forma textual, é disponibilizada e se acumula, num fenómeno apelidado de "sobrecarga de informação", pode transformar a pesquisa e consulta de informação numa tarefa árdua e penosa. Este excesso de informação torna difícil, senão impossível, a seres humanos consultar manualmente toda a informação disponível e conseguir filtrar o conteúdo que lhes é relevante, em tempo útil e de forma satisfatória, mesmo que esta procura se resuma a um tópico muito restrito.
Este problema tem sido potenciado em grande parte pelo crescimento explosivo, a que temos assistido, da *World Wild Web* (WWW) ao longo de mais de duas décadas. Este crescimento tem fomentado não apenas o crescimento da quantidade de informação disponível, mas também a diversificação dos formatos em que esta informação é disponibilizada, existindo actualmente uma enorme abundância de conteúdo multimédia disponível e de fácil acesso ao comum utilizador.
A resolução, ou pelo menos atenuação, deste problema parece passar então por encontrar soluções de redução do volume de informação disponível. Trata-se assim de um problema de condensação da informação disponível [JS08]. Quando se fala de condensação de informação em formato textual, em particular, fala-se da produção de um resumo do conteúdo do(s) texto(s) original(ais). Esta prática é de certa forma comum em muitos domínios, podendo ser identificada nas seguintes áreas: sinopse de um filme, síntese de uma notícia, resumo de uma publicação científica, tendo sido a sua aplicação nesta última tarefa referida de particular interesse à comunidade científica na génese desta área de pesquisa.
A *Sumarização Automática de Texto* é o campo do Processamento de Linguagem Natural (PLN) que tem por objectivo produzir computacionalmente uma versão abreviada de um dado texto original mantendo o seu teor informativo. Embora não exista uma definição elaborada e universalmente aceite de *resumo*, um conjunto de características são unanimemente aceites como devendo ser parte de qualquer resumo. Por exemplo, um resumo pode ser resultado da síntese de um ou mais documentos, um resumo deve preservar a informação relevante e um resumo deve ser de tamanho reduzido. Estes são elementos identificados da definição de resumo dada por Radev *et al.* em [RHM02], frequentemente citada na literatura [DM07, SUN11].
A dimensão desta condensação de conteúdos é habitualmente ditada por uma determinada *taxa de compressão*, que é a medida indicadora da proporção entre o tamanho de um texto original e o tamanho do resumo correspondente.
Na Figura 1.1, pode-se observar um exemplo simples de um pequeno texto e duas possíveis versões condensadas, a primeira produzida por um sumarizador humano e a segunda por extracção automática de texto.
Embora o resumo da informação textual continue sendo uma preocupação deveras pertinente, está longe de ser uma preocupação recente. A tarefa de produzir resumos de publicações científicas tem sido, desde sempre, levada a cabo maioritariamente por intervenientes humanos. Já em finais da década de 50 existia a preocupação com a crescente proliferação de textos técnicos e científicos, que requerem a elaboração de resumos, com o propósito de poupar tempo e esforço ao potencial leitor, quando este procura informação relevante num determinado do-

cumento [Luh58].

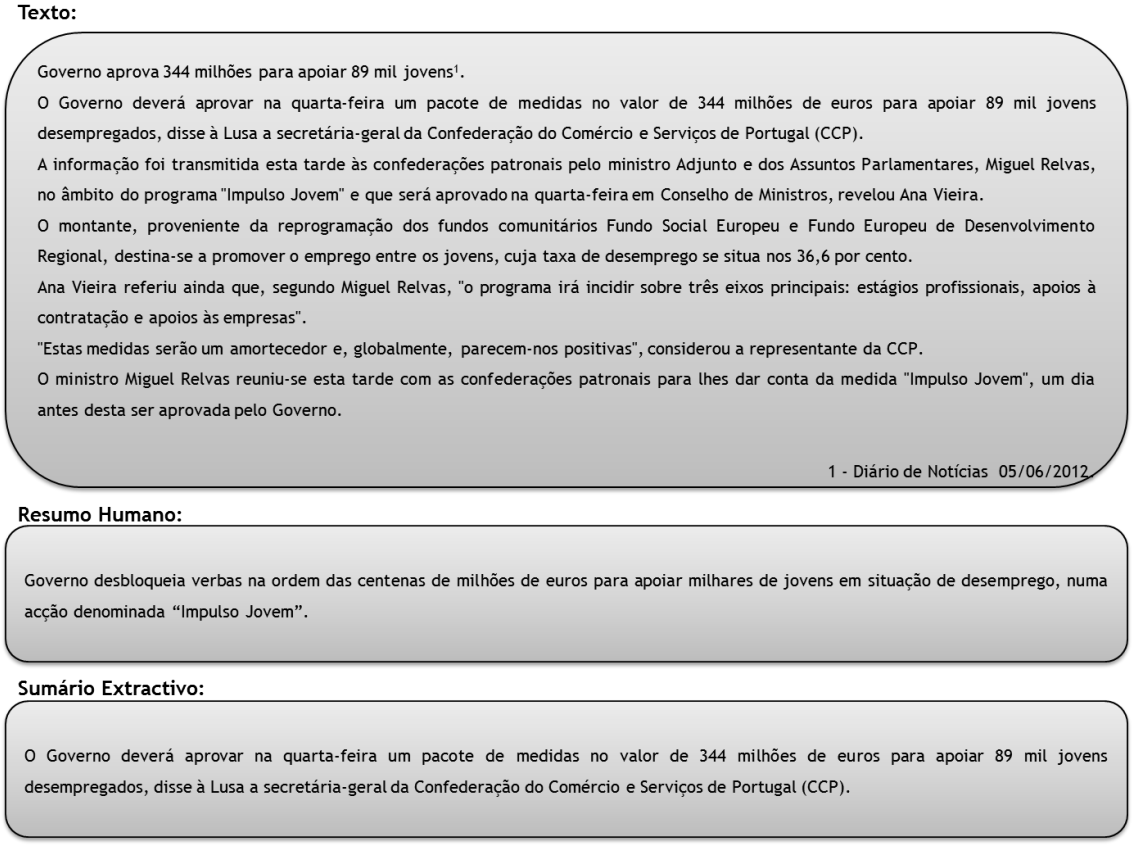

Figura 1.1: Exemplo de sumarização de uma notícia.

Luhn, elabora ainda sobre o esforço intelectual de se produzir um resumo, que requer familiaridade com o assunto, técnica e experiência. É ainda dito que os resumos produzidos por humanos tendem a ser influenciados pela formação, postura e disposição do produtor do resumo, ou seja, as opiniões ou interesses pessoais do produtor do resumo podem por vezes influenciar a sua interpretação das ideias do autor do texto original. Neste sentido, resumos do mesmo texto original produzidos por pessoas diferentes ou pela mesma pessoa em alturas diferentes, poderão variar qualitativamente.

Assim, em [Luh58] é sugerido, pela primeira vez, a implementação de um sistema automático de sumarização de texto, que pretendia eliminar o esforço e comportamentos tendenciosos de sumarizadores humanos. O sistema proposto tratava-se essencialmente de uma abordagem estatística para extrair frases do texto original, classificando-as com base na frequência de palavras ou expressões.

Outra abordagem, introduzida por Baxendale [Bax58], sugeria a classificação de frases através da avaliação de características estruturais, como a sua posição no documento e o número de palavras do título, cabeçalhos e palavras de sinalização que continham.

Luhn [Luh58], ditou que o desempenho de um sistema de sumarização automática de texto estaria, na altura, dependente da capacidade da comunidade científica de fornecer conteúdo textual em formato inteligível pela máquina. Esta capacidade era no entanto muito limitada uma vez que, texto previamente impresso teria de ser passado para formatos inteligíveis pela máquina manualmente, por não haver na altura processos automatizados para o fazer. Esta área de investigação viria mesmo a assistir a um período de dormência sendo que o próximo trabalho

de relevo viria a ser apresentado apenas uma década depois dos esforços iniciais [Edm69].
Desde então, assistiu-se a um moderado desenvolvimento e a um crescente interesse nesta área de investigação por parte da comunidade científica, mas foi só na década de 90 que verdadeiramente se deu o seu renascimento. Uma utilização mais ampla de métodos de inteligência artificial [KPC95], em conjunto com a combinação de métodos clássicos com novas abordagens em sistemas híbridos, aliada à melhoria generalizada das tecnologias de PLN, como analisadores textuais rápidos e robustos e etiquetadores morfossintáticos, deram novo folgo à área, por virem facilitar a construção de plataformas experimentais para explorar a aplicação da sumarização automática de texto a novas tarefas, através de estratégias adequadas para o fazer [SJ07].
Outro factor que ajudou a fomentar este renovado interesse na área foi o já referido enorme crescimento do volume de material em formato digital e em particular documentos completos em formato textual [SJ07]. Este fenómeno veio reforçar a já evidente necessidade de sistemas capazes de produzir versões reduzidas dos conteúdos textuais.
Embora o crescimento do volume de material digital, se tenha tornado evidente tanto no domínio dos conteúdos públicos como no domínio dos conteúdos privados, a dificuldade de acesso (publicações de assinatura paga), a especificidade e opacidade técnica do material (publicações da área das ciências exactas) ou mesmo heterogeneidade (dados de empresas) dos conteúdos de domínio privado, fizeram desviar o foco de atenção da comunidade científica que no início se centrava fundamentalmente neste tipo de conteúdos (principalmente publicações científicas), de forma a que esta se tenha passado a centrar maioritariamente na análise/utilização de conteúdos disponíveis livremente e em particular conteúdo textual proveniente dos diversos meios noticiosos [SJ07].
Os avanços nesta área de investigação têm em muito sido promovidos também pela organização de seminários e conferências (DUCs [DUC07], TACs [TAC12], ACL *Workshops* [ACL12], etc.) por parte da comunidade científica, que têm em vista reunir esforços de desenvolvimento, bem como encontrar soluções de avaliação do desempenho dos diversos métodos propostos. Esta é uma questão particularmente difícil e embora várias abordagens tenham já sido propostas, um único consenso foi alcançado, o de que a tarefa de avaliar a utilidade e qualidade de resumos gerados automaticamente é extremamente complexa e difícil, continuando a não existir por isso uma estratégia universal de avaliação [DM07].

1.1 Objectivos

Com o panorama descrito em mente, o trabalho apresentado nesta dissertação pretende mostrar a viabilidade de resolver ou de pelo menos atenuar os efeitos do problema de sobrecarga de informação que vivemos actualmente, através da construção experimental de uma ferramenta de sumarização automática de texto.
Para que este objectivo central fosse alcançado, alguns objectivos específicos foram definidos:

- Estabelecer a capacidade da ferramenta interpretar e trabalhar material textual composto por textos noticiosos, organizados em ficheiros de aglomerados de notícias, recolhidos automaticamente de diversos sítios da Internet;
- Investigar a implementação de diferentes paradigmas e métodos de sumarização descritos na literatura e de um modelo híbrido de sumarização capaz de conjugar diferentes características avaliativas de diferentes métodos;

- Por último e por forma a determinar a viabilidade e desempenho do sistema, é também objectivo deste trabalho abordar os desafios enfrentados na avaliação deste tipo de sistemas.

1.2 Estrutura da Dissertação

Esta dissertação é composta por cinco capítulos. Neste primeiro capítulo são introduzidas as principais temáticas e é fornecida uma breve contextualização numa perspectiva histórica da evolução e motivações da área de investigação em que esta dissertação se insere. São ainda apresentados os objectivos que motivaram o trabalho desenvolvido nesta dissertação.
No Capítulo 2 é feita a revisão bibliográfica sobre os aspectos teóricos considerados necessários para a realização do trabalho proposto, é ainda apresentado o actual estado da arte desta área de investigação. É inicialmente apresentado o modelo conceptual seguido por qualquer sistema de sumarização automática de texto. Sendo introduzida de seguida a taxonomia dos principais tipos de resumos presentes na literatura. Também neste capítulo é feita a descrição dos métodos mais salientes e relevantes desenvolvidos no âmbito da sumarização automática de texto até aos dias de hoje, fazendo notar, sempre que pertinente, as suas vantagens e desvantagens.
No Capítulo 3 é feita a caracterização geral do sistema, dos seus diversos componentes e características. Neste capítulo são ainda justificadas algumas decisões tomadas, descritos os métodos e as soluções implementadas.
No Capítulo 4 são apresentados em detalhe os métodos de avaliação utilizados para avaliar o sistema de sumarização automática de texto implementado. São ainda apresentados os conjuntos de dados usados na avaliação, e por fim são apresentados e discutidos os resultados obtidos da avaliação do sistema implementado.
Por último, no Capítulo 5 são apresentadas as principais conclusões deste trabalho e são discutidas algumas perspectivas de trabalho futuro.

Capítulo 2

Estado da Arte

O acto de sumarizar ou resumir, isto é, tornar mais sucinta a descrição de uma ideia, conceito, ou de uma notícia, é uma actividade bastante trivial. Quando um evento é narrado por alguém, o interlocutor faz habitualmente o resumo do que aconteceu e não a narração detalhada de todo o acontecimento. Os seres humanos, produzem constantemente e de forma inconsciente, este tipo de representações sucintas para algo que pretendam descrever ou comunicar. Outra forma muito comum de síntese são os sumários escritos. Exemplos disso são: notícias de jornais, artigos de revistas, resumo de artigos científicos, prefácios de livros, ou mesmo a sinopse de filmes [MPER01].
Apesar de frequente e aparentemente trivial, esta é uma actividade que envolve processos de raciocínio e memória complexos, inerentes à condição humana mas de difícil modelação computacional. Assim, de modo a tentar automatizar esta tarefa de forma satisfatória, é necessário tentar compreender os processos envolvidos na sumarização humana para os melhor poder simular em algum paradigma computacional.

2.1 Sumarização Humana de Texto

De forma geral, a sumarização de conteúdos em formato textual é uma actividade bastante comum na vida de qualquer pessoa e em particular para pessoas com um nível de escolaridade médio ou superior. A informação disponibilizada em formato textual é assim, um instrumento importante de comunicação e actualização em áreas como o meio social, académico e profissional. Por exemplo, estudantes recorrem por norma, a versões resumidas dos seus conteúdos curriculares em vez de consultar toda a bibliografia recomendada, e membros da comunidade científica fazem a selecção de material literário que consideram relevante com base nos seus títulos e resumos [RP03].
O foco da sumarização textual é distinguir no texto original o conteúdo relevante daquilo que poderá ser descartado, para a composição de uma versão resumida do texto fonte. Quando uma pessoa resume um texto, tenta primeiro identificar a sua essência, e em seguida adiciona gradualmente informação retirada do texto original para complementar o sumário [Man01b].
No entanto, esta não é uma análise simples e está sujeita a diversos factores influenciadores. A importância atribuída a determinada unidade textual ou excerto de um texto fonte pode depender das características do autor do sumário, das características dos potenciais leitores do sumário ou mesmo da importância subjectiva que tanto o autor como o leitor possam atribuir ao teor do texto, fazendo com que não só o conteúdo, mas também a estrutura do sumário possa estar sujeita à forma na qual se pretende representar o conteúdo do texto fonte.
Como já referido, quando se fala de sumarização humana é fácil reconhecer diversos tipos de sumários, como: o resumo de notícias jornalísticas, a síntese dos movimentos das bolsas de valores, sumários de obras literárias, resumos de trabalhos científicos ou mesmo apanhados de previsões meteorológicas. Cada um destes tipos de síntese envolve pressupostos e características específicas, bem como grande diversidade em termos da relação entre o teor da síntese

produzida e o conteúdo da(s) fonte(s) da informação original(ais) [MPER01].
Tendo em conta estes factores, é fácil perceber que, autores humanos diferentes produzem geralmente resumos diferentes, mesmo que a partir das mesmas fontes. Assim, resumos fruto de sumarização humana, variam habitualmente em termos de: conteúdo informativo, multiplicidade frásica ou estrutural e intenção comunicativa do autor [SR05].
Na óptica da sumarização automática, tantos condicionalismos levam-nos ao problema de como modelar tal diversidade de modo adequado, para que os resultados automáticos reflictam a diversidade de sumários sem que estes percam sua interdependência com os textos fonte correspondentes.
Em [CLW07], os autores sugerem que, uma vez que, o processo de sumarização é definido por competências e concepções humanas, abordagens cognitivas ajudam a determinar a forma como o processo de sumarização é estruturado, que características do texto fonte influenciam o sumário resultante ou como a utilização a que este se destina pode moldar o sumário. Os autores seguem dizendo que uma influência importante na sumarização automática de textos tem sido o estudo psicológico da sumarização humana em laboratório. Estas experiências [KV78] revelaram que os seres humanos criam uma organização hierárquica do discurso que permite fornecer pistas para recuperação de informação da memória. Outra forte influência na sumarização automática de textos prende-se com a análise do trabalho de sumarizadores profissionais. Segundo Endres-Niggemeyer, *et al*. [ENMS95, EN98] sumarizadores profissionais adoptam uma estratégia de cima para baixo (*top down*) na exploração da estrutura discursiva de um texto.
Ainda num esforço para tentar perceber as principais características do processo humano de sumarização de textos, no sentido de melhor modelar a acção de sumarizar, Jing [Jin01], analisa algumas directrizes para sumarização textual presentes na literatura [ENMS95, EN98, Thu24, Fid86, ANS97, Cre82], mas conclui que estas orientações estão longe de ser consensuais e que de uma forma geral são demasiado gerais ou de muito alto nível.
Mostrando-se infrutífera a procura por informação útil na literatura, no que diz respeito há produção manual de sumários, Jing [Jin01] realizou a análise de um conjunto de textos resumidos por sumarizadores profissionais no sentido de tentar identificar técnicas usadas que pudessem ser aplicadas também na sumarização automática de textos. Foi concluído que sumarizadores profissionais de um modo geral reutilizam frequentemente texto do documento original na produção do seu sumário. No entanto, não se verifica a prática de extracção simples do texto original, em vez disso, sumarizadores humanos por norma editam as frases extraídas. Isto corrobora a versão dada em [ENMS95, EN98], sobre a utilização de fragmentos do texto original na composição do sumário, por sumarizadores humanos.
Assim, foi identificado um conjunto de seis operações usadas de forma recorrente para a transformação de frases na construção de sumários humanos (redução de frases, combinação de frases, transformação sintáctica, parafraseando léxico, generalização e especificação, e reordenação). O autor [Jin01] denominou estas transformações de *operações de revisão*. De notar que existem transformações que não foram consideradas por serem muito pouco frequentes e muitas vezes o resultado de sumários muito curtos. É ainda observado que existem obviamente frases que não são de todo retiradas do texto original, sendo estas criadas de raiz para inclusão no sumário.
De um modo geral, para determinar o conteúdo relevante a incluir no sumário, de determinado texto fonte, para além das características superficiais (estruturais e linguísticas) que o autor confere ao texto com o seu estilo de escrita, devem também ser considerados, pelo sumarizador humano, outros factores dos quais se destacam essencialmente, o domínio do tema específico que o sumarizador detém para entender e abstrair ou generalizar a informação capturada do

texto fonte, e o conhecimento empírico prévio que ele possa ter nesse domínio [RP03].
Ainda que considerações sobre o conteúdo textual sejam altamente relevantes para a compreensão do processo humano de sumarização, o seu alto nível de subjectividade e o facto de exigirem representações muito complexas do conhecimento do domínio, faz com que a sua modelação computacional seja difícil. Em termos gerais, os modelos investigados para a sumarização automática adoptam essencialmente duas estratégias, podendo basear-se numa *análise superficial*, em que apenas são tidas em conta as características estruturais do texto fonte, ou numa *abordagem profunda*, quando a analise se baseia no conteúdo explícito dos textos fonte, bem como nas suas características estruturais.
De uma forma geral o modelo conceptual do processo de síntese seguido por sumarizadores humanos para produção dos sumários pode dividir-se em três passos: **interpretação** do texto fonte, o que envolve a leitura e compreensão do texto; **identificação** das informações mais relevantes; **condensação** do conteúdo e produção do texto do sumário, podendo isto envolver uma estrutura e expressão linguística nova e diferente. Estes passos coincidem com as fases de **interpretação**, **transformação** e **geração** que compõem a arquitectura conceptual de um sistema de sumarização automática proposta por Spärck Jones [SJ07, Jon98], como será visto na próxima secção.

2.2 Sumarização Automática de Texto

Sumarização automática de texto é uma técnica para tratar a sumarização de forma automática, onde uma máquina resume e produz uma versão condensada do texto original, segundo determinados requisitos [Man01b].
Esta é uma área de investigação cada vez mais importante, inserindo-se no campo do PLN, por sua vez subcampo das Ciências da Computação, que se ocupa da interacção dos computadores com as linguagens naturais humanas. Esta é geralmente uma área de investigação também associada às áreas da Linguística e Inteligência Artificial (IA).
A investigação nesta área foi inicialmente motivada pela necessidade de conseguir indexar o crescente número de publicações científicas, uma vez que na altura os recursos tecnológicos não permitiam o armazenamento em grande escala de documentos em formato digital. Armazenar apenas pequenos sumários dos documentos permitiria uma pesquisa e selecção mais eficiente. No entanto, a dificuldade em obter documentos nos formatos legíveis por computadores e a baixa qualidade dos sumários produzidos pelos métodos clássicos [Luh58, Bax58, Edm69], levou a um estagnar da investigação nesta área. Este interesse viria a renascer com a chegada da Internet e o consequente aumento considerável de documentos disponíveis *online*. Este movimento gerou a necessidade de pesquisar, seleccionar e assimilar informação em larga escala de forma eficiente [MPER01].
Com a enorme evolução das ferramentas de Recuperação de Informação (RI) passou a ser fácil o rápido acesso a quantidades enormes de informação, no entanto, o processamento de tal quantidade de informação requer meios automáticos e eficazes de condensação de informação, pelo que de outra forma é humanamente impossível filtrar conteúdo relevante de forma eficiente. Agora mais do que nunca a construção de sistemas de sumarização automática está na ordem do dia.
De uma forma geral, os sistemas de sumarização automática são tradicionalmente descritos [Jon98, Hov05, MM99] como seguindo um modelo tripartido [Llo08]. A Figura 2.1, representa o modelo conceptual de um sumarizador automático de texto, como descrito por Spärck Jones

[Jon98].

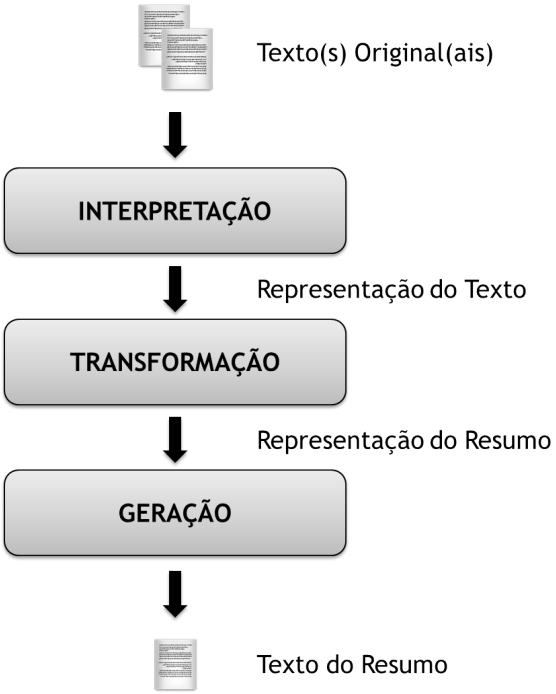

Figura 2.1: Modelo conceptual de um sistema de sumarização automática de texto.

No modelo representado, a fase **interpretação** corresponde à análise e interpretação do(s) texto(s) original(ais), por forma a construir uma representação computável. Na fase **transformação** a representação obtida na etapa anterior é trabalhada no sentido de formular uma representação daquilo que deverá ser o resumo, ainda num formato computável. A fase **geração** trata da tradução da estrutura representativa do resumo para o seu formato textual final.

O modelo conceptual descrito é genérico o suficiente para delinear os contornos gerais que devem definir um sistema de sumarização automática, no entanto, modelar o processo de sumarização humana descrito anteriormente, é uma tarefa extremamente complexa.

Spärck Jones [SJ93], atribui a ausência de progresso significativo nesta área, até meados da década de 90, à dificuldade de se modelar adequadamente o processo de sumarização, o que impossibilita a obtenção de sumários automáticos de qualidade. A autora observa no entanto que, para aplicações restritas, é possível explorar até certo nível características estruturais ou de composição, com base nos aspectos linguísticos do texto fonte [MPER01].

Esta incapacidade de definir formalmente o processo de sumarização promoveu um desenvolvimento da investigação em direcções muito diversas, quer estas direcções seguidas sejam motivadas por objectivos específicos dos investigadores ou meramente por uma atitude de tentativa e erro, existem algumas formas de classificar as diferentes abordagens ao problema da sumarização automática de texto.

Mani e Maybury [MM99] sugerem que uma forma útil de fazer esta classificação será analisar o nível de processamento. Assim, métodos de sumarização podem ser classificados como fazendo uma abordagem ao nível **superficial**, das **entidades** ou do **discurso**.

A abordagem superficial caracteriza-se pela representação da informação em termos de características superficiais que são selectivamente combinadas no sentido de obter uma função de saliência que possa ser usada para extrair informação relevante ou saliente. Estas características incluem:

- **Características temáticas** : métodos baseados nestas características fundamentam a análise do texto na ocorrência de termos estatisticamente relevantes. Termos que aparecem mais frequentemente num texto consideram-se mais relevantes e consequentemente frases contendo estes termos têm maior probabilidade de serem escolhidas para fazer parte do resumo;

- **Localização**: refere-se à avaliação da informação relevante em função da sua posição no texto, parágrafo, ou qualquer outra secção em particular. Exemplos frequentemente usados deste tipo de métodos são a extracção de frases do início de um texto ou de secções particulares como títulos ou cabeçalhos;

- **Características de fundo ou contexto**: nesta óptica assume-se que a relevância das unidades textuais a extrair é determinada pela presença de termos contidos no título, nos cabeçalhos, na parte inicial do texto ou numa consulta do utilizador;

- **Termos ou expressões sinalizadoras**: recurso a este tipo de características avalia a presença de termos ou expressões sinalizadores que possam fornecer indícios que permitam ajudar a determinar a presença de conteúdo relevante. Estes elementos incluem por exemplo indicações de síntese como: "em conclusão" ou "nossa investigação"; termos que realçam importância como: "importante" ou "em particular"; ou ainda tópicos específicos de determinado domínio como "bónus" ou "estigma".

A abordagem ao nível das entidades constrói uma representação interna do texto, modelando entidades textuais e as relações entre elas. Representam-se assim padrões de conectividade presentes no texto que podem ajudar a determinar o conteúdo saliente. Estas relações podem incluir:

- **Similaridade**: pode ser calculada pela sobreposição de vocabulário ou por técnicas linguísticas e poderá verificar-se a diferentes níveis. Por exemplo duas palavras que partilham a mesma forma canónica, ou frases, expressões ou parágrafos que partilhem as mesmas palavras;

- **Proximidade**: ditada pela distância entre as unidades textuais, é um factor determinante para se estabelecerem relações entre unidades textuais;

- **Co-ocorrência**: termos estão relacionados se ocorrem em contextos comuns;

- **Semelhança léxica**: relações entre as palavras determinadas com auxílio de dicionários de sinónimos. Algumas destas relações incluem sinonímia, hiponímia ou metonímia;

- **Co-referência**: relações deste tipo dizem respeito à identificação de ligações entre expressões de referência (por exemplo sintagmas nominais) que se co-referenciem, estas ligações de co-referência são usadas para construir cadeias de co-referência;

- **Lógica**: contempla relações de concordância, contradição, vinculação, e consistência lógica;

- **Sintaxe**: relações baseadas em árvores de análise sintáctica (*parsing trees*);

- **Representação de significado**: relações semânticas entre entidades textuais que podem ser do tipo estabelecido entre o predicado e o seu argumento.

A abordagem ao nível do discurso modela a estrutura global do texto, e sua relação com os objectivos comunicativos. Duas importantes características da estrutura discursiva de um texto fonte, que são habitualmente tidas em conta, são a coesão e coerência. Métodos que sigam esta abordagem analisam informação em termos de:

- **Formato**: do documento definido por marcação com hipertexto, contornos do documento ou disposição do conteúdo (secções, capítulos, etc.);
- **Segmentos de tópicos (*Threads of topics*)**: linhas de ideia sobre determinado tópico vão sendo relevadas ao longo do texto;
- **Estrutura retórica**: representa a argumentação ou estrutura narrativa, através da construção da estrutura de coerência do texto, de modo que a centralidade das unidades textuais nesta estrutura reflicta a sua importância. Estas relações são habitualmente representadas por estruturas em árvore.

As abordagens descritas são exemplos de técnicas puras aplicadas no sentido de desenvolver sistemas de sumarização automática. No entanto, a tendência predominante em sistemas actuais é adoptar abordagens híbridas que combinam e integram algumas das técnicas descritas [MM99, Llo08].
Mais adiante é feita a descrição dos métodos mais relevantes que adoptam estas abordagens, métodos que usam abordagens híbridas, e as suas áreas de foco.
Como já mencionado a actividade de sumarizar é bastante frequente e está presente nas mais variadas áreas [Man01b, HM00]. Apesar de a sumarização tradicionalmente se concentrar em conteúdo textual, não está restrita apenas a este tipo de fonte, podendo também ser considerado como entrada conteúdo multimédia, como imagens [FGL+08], vídeo [HSGG99, LPE97, MMM97], áudio [ZW00], gráficos ou tabelas.
Esta diversidade de conteúdos sumarizáveis é igualada pelo número de áreas onde a aplicação da sumarização automática seria de extrema utilidade [Man01b, Jin01]. Algumas das áreas onde a sua utilização é mais relevante actualmente são por exemplo:

- **Dispositivos móveis**: dada a crescente utilização de plataformas móveis em que o tamanho do ecrã é reduzido, é importante conseguir a condensação do conteúdo a apresentar para que este se adeqúe à área de visualização disponível [YW03, BGMP01, GTGJ01];
- **Internet**: como já referido, com a avassaladora quantidade de informação actualmente acessível online, procurar e recuperar informação relevante de forma eficiente torna-se extremamente importante. A apresentação de informação indicativa de conteúdo de forma condensada, poderá ajudar os utilizadores a avaliar de forma rápida a relevância de documentos ou páginas, para que não seja necessário o acesso aos mesmos, caso estes se venham a revelar de nenhum interesse para os utilizadores. Exemplo disto são os pequenos sumários que alguns motores de busca apresentam actualmente para cada ligação resultante de uma pesquisa, no entanto estes sumários são ainda muito rudimentares, e por isso, na maior parte das vezes absolutamente inúteis;
- **Bibliotecas digitais**: à medida que cada vez mais documentos vão ficando disponíveis em formato digital, a construção de bibliotecas digitais tenta dar resposta à necessidade de armazenar e indexar eficientemente estes documentos. Produzir sumários de documentos em bibliotecas digitais para que estes possam ser eficientemente organizados e recuperados, requer a utilização de técnicas automáticas de sumarização, uma vez que, o número

de documentos a incluir em bibliotecas digitais é enorme e a produção manual destes sumários seria impraticável em tempo útil [Mly06, OKG07];

- **Notícias multimédia**: desconstrução e rearranjo de conteúdo noticioso multimédia podem possibilitar a apresentação de um sumário das notícias mais relevantes sobre determinado tópico, emitidas durante determinado espaço de tempo, poupando o utilizador da consulta individual de cada notícia sobre o tema escolhido [HWC95, Mer97, LFZ11].

2.2.1 Taxonomias de Sumário

No sentido de melhor definir o objectivo a ser concretizado pela sumarização automática de texto, também será importante fornecer um entendimento daquilo que é afinal um sumário. Nesta perspectiva foram, ao longo do tempo, surgindo na literatura diversas definições de sumário [JS08, SJ07, RHM02, Cre82, Jon98, Hov05, For08, SJ01, SJ01, BB75]. No âmbito da sumarização automática de texto, foco deste trabalho, considera-se um sumário, como sendo o texto produzido a partir de um ou mais textos fonte, que contém parte significativa da informação do(s) texto(s) original(ais), e que não é mais longo do que metade do texto original [Hov05], e por norma substancialmente menos que isso. Mesmo a definição de texto [Hov05] inclui documentos multimédia, documentos *online*, hipertextos, etc. [SSZ+05]. Mais se acrescenta que, no sentido do que deve ser a automação da produção de sumários, o texto produzido deverá ser o resultado de um processo automático da computação de determinadas características do(s) texto(s) original(ais), por métodos pré-determinados, com o mínimo de intervenção humana possível.
Os sumários produzidos por um sistema de sumarização automática podem ser classificados em função dos diversos factores que influenciam a sua criação. Em [Jon98] a autora sugere que estes factores contextuais influenciadores da produção automática de sumários devem ser tidos em conta, por forma a desenvolver melhores sistemas de sumarização automática. Assim, a autora considera que estes factores podem ser categorizados essencialmente em três classes: **factores de entrada**, de **finalidade** e de **saída** [Jon98].

- **Factores de entrada**: as características do(s) texto(s) fonte a resumir pode determinar de modo crucial a forma como um resumo pode ser obtido. Estes factores podem dividir-se em três grupos: **formato de texto** (estrutura do documento); **tipo de tópico** (comum, específico ou restrito) e **número de fontes** (apenas um ou múltiplos documentos de entrada);

- **Factores de finalidade**: este é o conjunto mais importante de factores. Estes distribuem-se essencialmente por três categorias: **situação** (contexto para o qual se prevê o uso do sumário); **audiência** (tipo de leitor a que se destina o sumário) e **uso** (finalidade que o sumário pretende cumprir);

- **Factores de saída**: a última classe de factores relevantes prende-se com factores de saída. Esta classe encerra as categorias: **abrangência** (amplitude da cobertura do conteúdo informativo da fonte); **formato** (texto corrido ou texto estruturado) e **estilo** (informativo, indicativo, crítico e agregador).

A classificação descrita permite a caracterização de sumários produzidos por sistemas de sumarização automática por uma vasta gama de propriedades.
Não existindo uma classificação universalmente válida para textos resumidos de forma automática, uma vez que estes sumários podem ser classificados com base numa grande diversidade de

critérios [Jon98, HL99], será útil dar uma visão exaustiva das diferentes classificações, mais recorrentes na literatura, que tentam classificar um sumário de acordo com os diferentes critérios de classificação mais frequentemente discutidos. A Figura 2.2, ilustra o diagrama de classes de sumários, que visa satisfazer este objectivo.

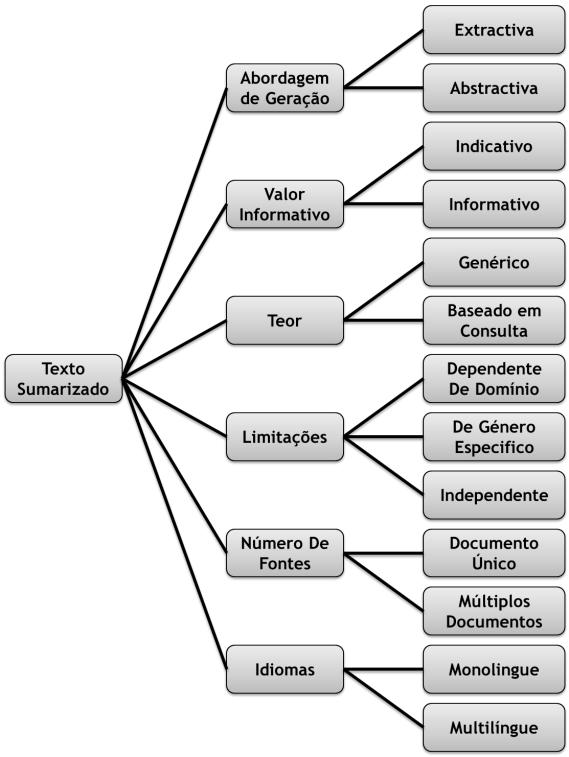

Figura 2.2: Tipos de sumários (adaptado de [GSG09]).

No diagrama (Figura 2.2), é possível observar a classificação de sumários produzidos por sistemas de sumarização automática de texto em função das suas características. De seguida é feita a descrição destas classificações.
No que diz respeito à metodologia de geração do sumário, um sistema de sumarização automática de texto poderá produzir:

- **Sumário extractivo**: composto através da selecção e cópia de unidades textuais retiradas de forma literal do texto original, concatenando-as para compor o sumário. A extensão das unidades textuais a retirar pode variar, sendo que habitualmente as unidades extraídas são frases. Este tipo de sumário está habitualmente associado a sistemas de sumarização que adoptem uma abordagem superficial;
- **Sumário abstractivo**: o texto resumido é uma interpretação do texto original. O processo de produção envolve reescrever o texto original numa versão mais curta, substituindo conceitos extensos por outros análogos mas mais curtos. Um sistema que produza sumários abstractivos analisa o texto fonte e tenta apresentar a sua compreensão do texto de uma forma humanamente inteligível, ou seja, em linguagem natural clara. A produção deste tipo de sumários envolve utilização de métodos linguísticos para analisar e interpretar o texto, no sentido de encontrar novos conceitos e expressões que lhe permitam gerar texto novo, mais curto, mas que transmita a informação mais importante do texto original.

Tradicionalmente [Hut87] quando se fala do valor informativo ou propósito que o sumário produzido pretende servir, este poderá ser considerado:

- **Sumário indicativo**: contém apenas os tópicos essenciais do texto fonte, pretendendo informar sobre o teor do texto fonte sem transmitir conteúdo específico como detalhes de resultados, argumentações ou conclusões. Tem geralmente por objectivo ajudar o utilizador a decidir se o documento original vale a pena ler. Podem ser usados, por exemplo, em sistemas de recuperação de informação, para indexação mais eficiente e para apresentação de resumos indicativos de tamanho muito reduzido;
- **Sumário informativo**: pretende ser substituto do texto original, devendo por isso conter todos os seus principais aspectos. Assim, se o texto original for organizado em função de dados, métodos ou hipóteses, um sumário informativo deverá conter os detalhes quantitativos e qualitativos relevantes de cada um desses tópicos. O sumário deve explicar determinado conceito com o maior nível de detalhe possível, para determinada taxa de compressão;
- **Sumário crítico ou avaliativo**: captura o ponto de vista do autor do sumário sobre um determinado conteúdo. Estes sumários vão além da descrição objectiva dos tópicos, métodos e dados, avaliando a relevância do texto original no contexto de outros trabalhos no mesmo campo. Assim, sumários avaliativos servem de revisões críticas. Estão de certa forma fora da mira dos sumarizadores automáticos actuais.

Em termos de teor, um sumário poderá ser:

- **Sumário genérico**: todos os tópicos do texto original são igualmente importantes e deveram ser incluídos no sumário. Destinam-se geralmente a comunidades amplas de leitores;
- **Sumário baseado em consulta**: conteúdo do sumário é formatado para conceder maior detalhe a um tópico ou consulta específica.

O sumário a produzir por um sistema de sumarização automática de texto poderá estar sujeito a diversas limitações no que diz respeito ao domínio do seu conteúdo ou ao género de sua escrita:

- **Sumário dependente do domínio**: produto de sistemas de sumarização sensíveis a determinado domínio. Estes sistemas apenas são capazes de gerar sumários de textos que pertençam a um domínio pré-determinado, podendo existir diferentes graus de portabilidade. A limitação a um domínio específico permite abordagens mais profundas do texto, resultando regra geral em sumários de maior qualidade;
- **Sumário de género específico**: gerado por sistemas de sumarização que tiram partido de características típicas de determinado género de escrita do texto. Exemplos disso são a organização em pirâmide invertida da informação em textos noticiosos ou o desenvolvimento argumentativo de um artigo científico;
- **Sumário independente de género e domínio**: sumários gerados por sistemas de uso geral, não dependentes da informação sobre os domínios e que não exploram características do género dos textos. Fazem geralmente uso de uma abordagem mais superficial para a análise dos documentos de entrada. No entanto, alguns sistemas de uso geral estão preparados para explorar informações específicas de domínio.

Sistemas de sumarização automática de texto poderão diferenciar-se em termos do número de fontes consideradas para a geração do sumário:

- **Sumário com documento fonte único**: é resultado da síntese de um único documento;

- **Sumário com múltiplos documentos fonte**: é o resultado da condensação do conteúdo mais relevante de um conjunto de documentos. Este tipo de sumarização enfrenta alguns desafios específicos importantes, dos quais se destacam, reconhecer e lidar com redundância, identificar as diferenças importantes entre os documentos e garantir a coerência do sumário, mesmo quando o material provém de documentos de origens diferentes.

Determinados sistemas de sumarização automática de texto (em especial para sumarização de múltiplos documentos) poderão suportar a sumarização de fontes em diferentes idiomas:

- **Sumário monolingue**: sumários cujo conjunto de textos fonte se encontra todo no mesmo idioma. A arquitectura do sistema pode ser determinada pelas características de uma linguagem concreta, fazendo do sistema dependente do idioma. Isto significa que algumas adaptações devem ser realizadas para que o sistema possa lidar com línguas diferentes;

- **Sumário multilíngue**: produzido por sistemas independentes de língua, que permitem a exploração das características dos documentos fonte que possuem transversalidade linguística. Neste caso o sumário produzido é monolingue.

De notar que estas classificações não são mutuamente exclusivas e que se sobreporão em função das características do sistema que produz o sumário.

2.3 Abordagens de Sumarização Automática

Como já referido, a sumarização humana de documentos é um processo complexo de abstracção de conceitos difícil de modelar matemática, lógica e computacionalmente, como o são, muitos outros tipos de processamento de informação realizados por seres humanos [JM99]. Os sumários diferem de pessoa para pessoa e variam normalmente em estilo de linguagem e nível de detalhe. Das classificações de sistemas sumarização anteriormente descritas, uma em particular é especialmente útil e frequentemente usada para distinguir estes sistemas. Esta classificação prende-se com o tipo de geração do sumário usado, que como já referido poderá ser *extractivo* ou *abstractivo*.

2.3.1 Abordagens Abstractivas

Na última década, têm vindo a ser desenvolvidos alguns sistemas para gerar sumários abstractivos através do uso das mais recentes ferramentas de PLN. Estes sistemas extraem expressões e cadeias léxicas do(s) documento(s) fonte e tentam fundi-las através de ferramentas geradoras de linguagem natural para produzir um sumário abstractivo.
No entanto, abordagens abstractivas requerem Processamento de Linguagem Natural profundo, como representação semântica, inferência e geração de linguagem natural, que actualmente não atingiram ainda uma fase maturidade satisfatória [YCKQ07].
Alguns exemplos deste tipo de sistema são:

- ***SumUM*** [SL02] é um sistema de sumarização de documento único e de domínio restrito, no caso artigos científicos. O sistema baseia a sua avaliação na análise sintáctica superficial e semântica, identificação de conceitos e extracção de informação relevante, fazendo depois a construção da representação do resumo e regeneração de texto;

- ***GISTexter*** [HL02] produz sumários extractivos e abstractivos a partir de documentos individuais ou conjuntos documentos, através de técnicas de Extracção de Informação (EI) baseadas em modelos (*template-driven*). O sistema funciona de formas diferentes dependendo de se se pretende a sumarização de um único documento ou de um conjunto de documentos. Para um único documento, as frases mais relevantes são extraídas e comprimidas por regras aprendidas a partir de um corpus de resumos escrito por humanos. Numa fase final, é realizada a redução do sumário para um comprimento máximo de cem palavras. Quando se trata da sumarização de um conjunto de documentos, o sistema com base em técnicas de extracção de informação, usa modelos de extracção de informação de um conjunto prévio (se o tópico for bem conhecido) ou gera novos modelos específicos (se o tópico for desconhecido). Os modelos gerados pelo sistema de EI ***CICERO*** [SH02] são depois mapeados para trechos de texto dos documentos, por forma a resolver expressões anafóricas. Esses trechos de texto podem depois ser usados para gerar resumos informativos coerentes dos múltiplos documentos;

- ***Cut & Paste*** [Jin01] usa técnicas de redução e combinação de frases. As frases mais relevantes são identificadas por um algoritmo de extracção de frases que engloba vários métodos (coerência léxica, importância estatística, expressões de sinalização e posição no texto). O resumo final não é conseguido a partir da geração de texto como nos sistemas anteriores, em vez disso, o sistema usa um conjunto de seis transformações (redução de frases, combinação de frases, transformação sintáctica, parafraseando léxico, generalização e especificação, e reordenação), denominadas operações de revisão, para editar as frases mais relevantes extraídas.

Estes são apenas alguns exemplos de sistemas de sumarização abstractiva, existindo muitos outros descritos na literatura [SJ07, Llo08, ACC+03, LP12].
Outra abordagem, que é comparativamente menos complexa, é a construção de sumários extractivos, nos quais frases do(s) documento(s) original(ais) são seleccionadas e justapostas para formar o sumário.

2.3.2 Abordagens Extractivas

Métodos Clássicos
Luhn, num trabalho pioneiro nesta área [Luh58], sugere que a frequência de uma determinada palavra num texto fornece uma medida útil da sua importância no texto. Para o método proposto, numa fase inicial as palavras são reduzidas à sua forma canónica e palavras insignificantes são removidas. No passo seguinte as palavras significativas são ordenadas numa lista, por ordem decrescente das suas frequências, passando o índice de cada palavra na lista a representar o seu valor de significância. A importância de cada frase é depois calculada em função do número de palavras significativas contidas na frase e distância entre elas, calculada pelo número de palavras não significativas entre cada ocorrência de uma palavra significativa. Numa fase final as frases são ordenadas por ordem decrescente das suas importâncias, e as frases melhor classificadas são escolhidas para formar o sumário.
Outro método surgido nos primórdios desta área de investigação defendia que informação saliente pode ser encontrada quando considerada a sua posição no documento [Bax58]. Neste sentido, foi analisado um conjunto de 200 parágrafos e foi concluído que em 85 % dos casos a frase mais representativa do tópico era a primeira e em 7 % dos casos a última. Apesar da aparentemente insignificante percentagem associada à importância representativa do tópico,

atribuída à última frase, o autor considerou que esta também é importante para o sumário, uma vez que representa geralmente a ligação de um parágrafo com o seguinte, ou seja, com a primeira frase do parágrafo seguinte, que segundo o estudo, está fortemente relacionada com o contexto textual. Assim, no sentido de preservar a coesão textual, o autor sugeriu que tanto a primeira como a última frase de cada parágrafo deveriam ser incluídas no sumário.
Mais de uma década depois da apresentação dos dois métodos descritos anteriormente, Edmundson [Edm69], sugere um novo método que para além de integrar as duas características usadas pelos métodos anteriores explora também duas novas características, a presença de palavras sinalizadoras (palavras como "significativo", "dificilmente", "conclusão", etc.) e a estrutura do documento (valorizando a presença de palavras contidas em elementos como títulos ou cabeçalhos). A pontuação final de uma frase é dada pela Equação 2.1:

$$F_i = S_i \times p_1 + E_i \times p_2 + T_i \times p_3 + L_i \times p_4 \quad (2.1)$$

Onde F_i é a pontuação final da fase i, e S_i, E_i, T_i e L_i são, respectivamente, a pontuação da frase i em função das palavras sinalizadoras que esta contém, das palavras estatisticamente relevantes que esta contém, das palavras presentes em elementos estruturalmente relevantes (títulos, cabeçalhos, etc.) que esta contém e da sua posição. Os pesos p_1 a p_4 definem a distribuição linear de importância atribuída às quatro pontuações parciais.
No âmbito do trabalho desenvolvido, foi produzido o resumo manual de um conjunto de 400 documentos técnicos e em fase de avaliação verificou-se que cerca de 44 % dos sumários extractivos automáticos produzidos coincidiam com os resumos manuais [DM07].

Métodos de Aprendizagem Automática

A partir de meados da década de 90, a área da sumarização automática de texto ganha novo folgo com a introdução de técnicas de Aprendizagem Automática (AA) ao processamento de linguagem natural. Surgem diversas abordagens que aplicam técnicas estatísticas para geração de sumários extractivos.
Um primeiro trabalho baseado em AA foi apresentado por Kupiec *et al.* [KPC95] e descrevia o uso de um classificador *naive-Bayes*, que com base em determinada função de classificação, ditava a probabilidade de cada frase dever ser incluída no sumário. Foram consideradas as mesmas características tidas em conta por Edmundson, mas para além destas eram ainda consideradas a presença de palavras em maiúsculas e o comprimento das frases. O modelo assumia a independência das características. Depois das classificações atribuídas, as frases melhor classificadas eram escolhidas até ser preenchido o tamanho de sumário pretendido. O classificador foi treinado num corpus de 188 pares de documentos completos e respectivos sumários.
A avaliação do sistema foi feita usando um corpus de documentos técnicos com resumos manuais. Para o efeito, os autores analisaram manualmente cada documento procurando no texto uma frase correspondente a casa frase do resumo manual desse documento. Os sumários automáticos eram depois avaliados de acordo com esse mapeamento entre frases do resumo manual e frases extraíveis. Esta avaliação revelou que o melhor desempenho era obtido quando consideradas apenas características de posição e sinalização e comprimento das frases.
Outro método, exemplo do uso deste tipo de classificador é o desenvolvido por Aone *et al.* [AOGL99]. Este método tenta evidenciar palavras indicativas de conceitos fundamentais de um documento, através do uso da métrica $tf - idf$. Um termo terá um valor $tf - idf$ elevado se for muito frequente em determinado documento e rara no corpus. O valor idf é calculado a

partir de um corpus de grandes dimensões do mesmo domínio do documento onde o termo, para o qual o valor $tf - idf$ está a ser calculado, se insere. Outra característica usada foi a identificação de entidades que poderiam ser representadas por palavras únicas ou pares de substantivos. Era ainda empregue a análise superficial de discurso para identificar referências a uma mesma entidade por nomes diferentes, por forma a manter a coesão. Também foi feita a identificação de relações de sinonímia com recurso à ferramenta *Wordnet* [Mil95].
Numa abordagem que pretendia ter em conta as interdependências locais entre frases, Conroy and O'leary propõem um novo método [CO01], que em contraste com os anteriormente descritos adopta a estrutura sequencial de um *Hidden Markov Model* (HMM), considerando apenas três características: (1) a posição da frase no documento, (2) número de termos na frase e (3) a probabilidade dos termos da frase dados os termos do documento. A posição das frases no documento é implicitamente representada pela estrutura sequencial de estados do modelo constituído por $2s + 1$ estados, que se dividem em s estados de sumário e $s + 1$ estados de não pertença ao sumário. A probabilidade de a frase seguinte ser incluída no sumário irá variar, dependendo de se a frase actual é uma frase de sumário ou não. O sistema é treinado com base num corpus, onde as frases de um sumário gerado por humanos são mapeadas ao documento original, no sentido de, com base nas ligações estabelecidas, tentar determinar as probabilidades de transição entre os estados do modelo.
Motivados pelo objectivo de superar o sistema de sumarização de referência usado nas conferências DUC [DUC07], que consiste em escolher as primeiras n frases de uma notícia para formar o sumário de 100 palavras, a partir de um documento de entrada único, até então nunca batido por uma diferença significativa [Nen05], Svore *et al.* apresentam uma nova abordagem baseada em redes neuronais [SVB07]. Um conjunto de novas características é avaliado, com base em registos de consultas sobre notícias e entidades da *Wikipédia* [Wik01]. Usando o algoritmo de redes neuronais baseado em pares *RankNet* [BSR+05], um classificador de frases é treinado para classificar cada frase no documento e assim identificar as frases mais importantes. O sistema desenvolvido conseguiu um desempenho significativamente melhor na sumarização de documento único em textos noticiosos que o sistema de sumarização de referência referido.

Métodos de Teoria de Grafos

Outra técnica que tem sido aplicada com sucesso à sumarização automática de texto é a teoria de grafos. A ideia é usar um grafo para representar a estrutura do texto. Neste sentido, os nós do grafo representam unidades textuais e as arestas as relações entre elas. As unidades textuais representadas pelos nós poderão ter granularidades diferentes, podendo ser por exemplo, palavras, expressões, frases ou parágrafos, e as arestas poderão representar diferentes tipos de relação, como semelhança (geralmente avaliada pela repetição de conteúdo), relações léxicas ou semânticas.
Um método pioneiro no uso da teoria de grafos na sumarização automática de texto foi o proposto por Salton *et al.* [SSMB97], que descrevia a representação da estrutura de determinado documento por um grafo não direccionado, onde parágrafos são usados como nós do grafo, sendo as suas arestas representativas da semelhança entre estes parágrafos (repetição de conteúdo textual). Os parágrafos são classificados, no sentido de encontrar aqueles que gozem de maior centralidade, com base no seu grau de conectividade no grafo, ou seja, um parágrafo é considerado tão importante quanto o número de ligações que tiver a outros parágrafos. Os parágrafos mais centrais são então extraídos para formar o sumário.
Erkan e Radev [ER04a, ER04b], consideram uma nova abordagem, em que a importância das frases de um documento é calculada com base no conceito de centralidade de vector próprio,

num grafo composto por nós representativos dessas frases. Neste modelo, os valores da matriz de adjacências que representam conectividade entre frases, ou seja, as arestas do grafo, são calculados pela similaridade do co-seno entre frases. A saliência das frases é calculada, após redução do número de ligações do grafo com base num determinado limite mínimo para o valor de similaridade, por um passeio aleatório no grafo (com a matriz de adjacência, ou seja, a matriz de similaridade vista como uma cadeia de *Markov*).
Motivados pelo sucesso de alguns algoritmos de classificação baseados em grafos, em áreas como a análise da estrutura de ligações da WWW, análise de citações e redes sociais, Mihalcea e Tarau [MT04, Mih05, MT05], propõem usar a mesma linha de raciocínio na sumarização extractiva de texto. Algoritmos de classificação baseados em grafos, como o algoritmo *PageRank* [BP98] ou o algoritmo *HITS* [Kle99], baseiam a avaliação da importância de um nó do grafo na informação a global, calculada recursivamente a partir de todo o grafo, em vez de fazer essa avaliação com base na informação local do nó [MT04].
A ideia base por de trás dos algoritmos de classificação baseados em grafos é a de *recomendação*, ou seja, quando um nó está ligado a outro, o primeiro nó está basicamente a recomendar o segundo. A importância de um nó é assim dada pelo número de outros nós que o recomendem, no entanto, a importância de cada recomendação é determinada pela importância do nó que faz a recomendação.
O modelo proposto pelos autores [MT04], denominado de *TextRank*, foi aplicado a duas tarefas distintas do processamento de linguagem natural, a sumarização extractiva de texto e a extracção de palavras-chave, e este modelo segue independentemente do tipo e características dos elementos adicionados ao grafo, os seguintes passos:

1. Identificar e adicionar ao grafo unidades textuais, como nós, na granularidade que melhor se ajusta à resolução da tarefa em questão;
2. Identificar as relações entre as unidades textuais e adicionar no grafo ligações que representem essas relações (estas ligações podem ser direccionadas ou não direccionadas e pesadas ou não pesadas);
3. Correr o algoritmo de classificação baseado em grafos até que este convirja ou até que seja atingido um número de iterações limite pré estabelecido;
4. Ordenar os nós do grafo pela sua pontuação final e fazer a selecção das unidades textuais associadas aos nós com base nas suas pontuações.

O trabalho relatado nesta dissertação é em parte baseado neste método, pelo que o mesmo é explorado em maior detalhe no próximo capítulo.
Ainda fazendo uso da teoria de grafos, Mani *et al.* [MBG98] descrevem um método que calcula a saliência de elementos num grafo usando um algoritmo de procura baseado em propagação de activação. Este método tinha já sido explorado pelos autores para sumarização de múltiplos documentos [MB99, MB97]. A ideia é construir um modelo geral de coesão através da representação explícita de relações de coesão textual. Nesse sentido, o texto é representado por um grafo em que os nós são ocorrências de palavras em posições distintas. As ligações estabelecidas entre nós do grafo representam relações de coesão extraídas do texto que incluem **adjacência**, **repetição**, **sinonímia**, **hiperonímia**, e **co-referência** entre instâncias de palavras. A saliência dos termos é calculada pelo algoritmo de procura baseado em propagação de activação [CN95], com base nas relações de coesão extraídas. Inicialmente são identificados um conjunto de nós de entrada, que se relacionam com determinado tópico definido por uma consulta de utilizador, posteriormente os valores de activação de nós relacionados com os nós de entrada vão

sendo alterados pelo algoritmo de procura em função do tipo de relação entre os nós, e dos pesos dos seus antecessores. Se não existir um tópico definido, os nós de entrada são encontrados pelos valores $tf - idf$ dos seus termos. O peso de uma frase é determinado pelo valor final da activação propagada, dos seus termos, dados os valores $tf-idf$ iniciais e os termos do tópico.

Métodos Baseados em *Cluster*

Os documentos são normalmente divididos, implícita ou explicitamente, em secções de modo que diferentes tópicos sejam abordados, um após o outro, de uma forma organizada. É por isso intuitivo pensar que um sumário deva abordar os diferentes tópicos do(s) documento(s) original(ais). Uma forma de abordar esta questão é através de técnicas de aglomeração de segmentos textuais ou de documentos completos (no caso sumarização de múltiplos documentos, e em sistemas de recuperação de informação) de acordo com os tópicos que estes representam. No caso da sumarização de múltiplos documentos com temas totalmente diferentes, a aglomeração de documentos torna-se fundamental para gerar um sumário significativo, uma vez que, todos os temas devem ser convenientemente representados no sumário de acordo com a sua importância [GDCY02].

Estas abordagens lidam essencialmente com os conceitos de *redundância* e *diversidade*. De uma forma geral, sistemas de sumarização que sigam esta abordagem procuram numa primeira fase construir vários aglomerados que representem os diversos tópicos presentes no texto. Estes aglomerados são constituídos por conjuntos de frases, não necessariamente contíguas, que são mutuamente semelhantes de acordo com determinado critério. Desta forma é possível representar a diversidade dos tópicos presentes no documento. Para cada tópico representado por um aglomerado é identificada a frase mais importante para ser usada como representante do tópico desse aglomerado. A selecção de apenas a(s) melhor(es) frase(s) de cada tópico permite reduzir a redundância da informação extraída. As frases identificadas são depois usadas para formar o sumário [NM03].

O trabalho pioneiro na aplicação do conceito de diversidade na sumarização automática de texto foi o apresentado por Carbonell e Goldstein [CG98]. O algoritmo desenvolvido, denominado de *Maximal Marginal Relevance* (MMR), pretendia maximizar a *relevância marginal* para sistemas de RI e sumarização automática. O conceito de relevância marginal pretende fornecer uma métrica de medição de novidade relevante, através da medição independente e posterior combinação linear de relevância e novidade. Assim, um documento é dotado de elevada relevância marginal se é relevante com relação a determinado tópico mas revela um nível de similaridade mínimo para com documentos previamente seleccionados.

Para sumarização de um documento único, uma frase tem elevada relevância marginal se é muito relevante para determinado tópico, definido pelo utilizador, mas menos similar a frases já seleccionadas. Este método tenta assim maximizar a relevância das frases incluídas no sumário, enquanto reduz a repetição de informação minimizando a sua redundância.

Este algoritmo goza de grande relevo no campo da sumarização automática de texto, tendo vindo a ser usado em combinação com outras técnicas em diferentes abordagens [GDCY02, NM03, AAHM11].

Métodos de Análise de Semântica Latente

A Análise de Semântica Latente (LSA) [DDF$^+$90] é uma técnica algébrico-estatística, baseada em corpus, para extrair e representar o uso contextual de significados das palavras em passagens do discurso [JS08]. Esta técnica identifica relações de **sinonímia** e **polissemia**, extraídas através da análise dos valores estatísticos da co-ocorrência das palavras no mesmo contexto [Pat07].

A ideia base é que o conjunto de todos os contextos (semânticos) de palavras, em que uma determinada palavra possa ou não aparecer, fornece restrições mútuas que permitem determinar a similaridade de significados de palavras e conjuntos de palavras entre si [JS08]. A análise de semântica latente classifica como próximos palavras e frases ou documentos fortemente relacionadas semanticamente, ou seja, as frases ou documentos são considerados próximos mesmo que não contenham as mesmas palavras [GDCY02, Pat07].
Na aplicação desta técnica à sumarização automática de texto [DDF+90], o processo consiste essencialmente em dois passos. Inicialmente é construída uma matriz palavra-frase, em que cada coluna da matriz representa o vector pesado das frequências de termos de cada frase do documento de entrada. No caso de sumarização de múltiplos documentos, a matriz será do tipo palavra-documento e as colunas da matriz representam os vectores das frequências de termos de cada documento do conjunto de documentos de entrada. No segundo passo do processo, a técnica algébrica de *Decomposição de Valor Singular* (SVD) é aplicada à matriz. Esta técnica encontra dimensões importantes e mutuamente ortogonais de vectores frase na matriz, através da compressão de dimensões semelhantes numa dimensão composta, reduzindo assim a dimensionalidade das matrizes às suas dimensões mais importantes. A escolha de uma frase representante de cada uma das dimensões permite garantir relevância em relação ao documento, enquanto a ortogonalidade entre estas assegura a não redundância entre frases escolhidas [GL01].
A vantagem da utilização de vectores de análise de semântica latente para sumarização automática de texto, em vez de simples vectores de palavras, é que as relações conceptuais, como representadas no cérebro humano, são automaticamente capturadas [LD97], enquanto o uso de vectores de palavras, em que não é aplicada a transformação por análise da semântica latente, exige o emprego de métodos explícitos que permitam revelar estas relações conceituais [GDCY02].
Algumas das abordagens mais relevantes que fazem uso desta metodologia para a sumarização automática tando de documentos únicos como de múltiplos documentos podem ser encontradas em [JS08, For08].

2.3.3 Sumarização Extractiva vs. Sumarização Abstractiva

A qualidade dos sumários produzidos por modelos extractivos é em geral questionável, sendo difícil controlar ou modelar processos de decisão que garantam a sua textualidade (coesão e coerência). A interligação entre as frases não é normalmente tida em conta, o que pode levar a incoerências como por exemplo frases extraídas poderem conter referências anafóricas não resolvidas. No exemplo da Figura 2.3, o pronome pessoal "Ele" encontrado na segunda frase ([F2]) refere-se ao nome "Chomsky" surgido na primeira frase ([F1]), e se apenas a segunda frase for incluída no sumário, a sua referência a "Chomsky" aparecerá fora de contexto. Sistemas de sumarização automática extractiva produzem assim, sumários informativos, mas geralmente com má textualidade.

[F1]: *Noam Chomsky é um linguista norte-americano.*

[F2]: *Ele é conhecido por ter criado a gramática generativa transformacional.*

Figura 2.3: Exemplo de uma referência anafórica.

No entanto, normalmente a produção de sumários abstractivos requer maquinaria pesada (pro-

cessos complexos, corpus anotados manualmente, modelos de conhecimento de grandes dimensões) para a geração de linguagem, e estes são processos de difícil reprodução e extensão a domínios amplos. Em contraste, a simples extracção de frases têm produzido resultados satisfatórios em aplicações de larga escala, especialmente para sumarização de múltiplos documentos, sendo mais eficiente, simples de implementar e menos dependente de género, domínio, e idioma, podendo não o ser de todo.
Portanto, embora a legibilidade e o poder de síntese de informação de sumários abstractivos sejam em geral superiores aos de sumários extractivos, como já referido, a geração de sumários abstractivos é muito mais custosa, exigindo recursos linguísticos e computacionais avançados, nem sempre disponíveis, como analisadores sintácticos, discursivos e semânticos. Em contraste, a sumarização extractiva representa actualmente um maior interesse, em consequência da sua maior simplicidade de implementação e abrangência. Sistemas de sumarização extractiva podem ser independentes de língua natural, domínio e género textual. Além disso, técnicas e métodos de IA, como AA são mais facilmente aplicadas em sistemas de sumarização extractivos.

Problemas com a Sumarização Extractiva

- Sumários extractivos poderão ser propensos a redundância, uma vez que, unidades textuais com conteúdo saliente semelhante obterão pontuações igualmente altas e serão susceptíveis de ser incluídas no sumário. Isto poderá ser especialmente evidente para abordagens puramente estatísticas e em sumarização de conjuntos de documentos do mesmo tópico;

- Frases extraídas tendem normalmente a ser mais compridas do que a média. Assim, segmentos não essenciais poderão ainda assim ser incluídos, consumindo espaço que poderia ser preenchido por conteúdo mais informativo;

- Informação relevante está geralmente distribuída por várias frases, e sumários extractivos não conseguem captar nem lidar com esta questão, a menos que o sumário seja comprido o suficiente para abarcar todas as frases relevantes;

- Informações intensionalmente discordantes ou contraditórias podem não ser apresentadas de forma precisa;

- Métodos de extracção pura muitas vezes levam a problemas de coerência global do sumário, um problema frequente são referências anafóricas relativas a um sujeito ausente do sumário, ou seja, frases muitas vezes contêm pronomes que perdem as suas referências quando extraídas fora de contexto. No pior cenário possível, a justaposição de frases fora de contexto, pode levar a uma interpretação errada de referências anafóricas, o que poderá resultar numa representação imprecisa da informação original;

- Um problema semelhante consiste na extracção de expressões temporais, termos como “ontem”, “hoje”, “no mês passado”, etc., poderão ser problemáticos, uma vez que são relativas ao contexto temporal da criação do texto original, mas não necessariamente da criação do sumário.

Estes problemas tornam-se mais evidentes para sumarização de múltiplos documentos, uma vez que, os sumários extractivos são esboçados a partir de fontes diferentes com contextos diferentes. Uma visão comum sobre como abordar estas questões envolve geralmente o pós-processamento dos sumários, substituindo os pronomes responsáveis por referências anafóricas

pelos seus antecedentes, e substituindo expressões temporais relativas por datas concretas [GL10].

Problemas com a Sumarização Abstractiva

- A síntese de frases não é ainda um campo completamente desenvolvido, e sumários abstractivos gerados automaticamente podem também ser portadores de incoerência, até mesmo dentro das frases, enquanto, para sumários extractivos, problemas de incoerência ocorrem apenas entre frases;
- Foi demonstrado que os utilizadores de sumários automáticos preferem sumários extractivos, em vez de sumários abstractivos demasiado explicados ou artificialmente embelezados [ENH00]. Isto é facilmente explicado pelo facto dos sumários extractivos apresentarem a informação como esta é exposta pelo autor, o que permite aos utilizadores identificar e interpretar informação nas entrelinhas;
- O maior desafio da sumarização abstractiva é um problema de representação da informação ou conhecimento. A capacidade dos sistemas é limitada pela riqueza das suas representações e pela sua capacidade de gerar essas construções, isto é, os sistemas não conseguem resumir conceitos que as suas representações não conseguem captar. Para domínios limitados, pode ser viável gerar estruturas adequadas, mas a construção de uma solução de propósito geral dependerá de análise semântica independente de domínio, e sistemas que consigam de facto compreender linguagem natural estão ainda além das capacidades da tecnologia de hoje [GL10].

Num estudo de utilizadores estas duas abordagens foram avaliadas comparativamente e foi observado que ambas as abordagens têm um bom desempenho em termos quantitativos. Qualitativamente, no entanto, foi observado que as diferentes abordagens têm um bom desempenho por razões diferentes, mas complementares. Foi concluído que métodos de sumarização eficazes deverão combinar as duas abordagens [CNP06].
Dado o contexto descrito, a maior parte do trabalho desenvolvido na área da sumarização automática baseia-se ainda na extracção de segmentos do documento original para formar o sumário, e sendo o foco deste trabalho a produção de sumários puramente extractivos, em vez de ser feita uma revisão exaustiva de todos os métodos e abordagens existentes, serão evidenciados os mais relevantes métodos de sumarização extractiva presentes na literatura. Para visões mais exaustivas das diferentes abordagens e sistemas de Sumarização Automática existentes, o leitor é remetido para as muitas revisões bibliográficas que foram sendo compiladas ao longo do tempo [JS08, SJ07, DM07, SUN11, Man01b, Jon98, Pai90, Zec97, Tuc99].

Capítulo 3

Caracterização do Sistema Desenvolvido

Neste capítulo é descrito o sistema implementado. São expostos os métodos utilizados para levar a cabo as diferentes fases do processo de sumarização automática (Figura 2.1), descritas no capítulo anterior. Decisões tomadas no sentido de alcançar os objectivos traçados são também debatidas e justificadas.

3.1 Visão Geral do Sistema

Foi desenvolvido um sistema de sumarização automática de texto para produção de sumários puramente extractivos, a partir de um texto de entrada único. As unidades textuais extraídas para formar o sumário são frases. É realizada também a extracção de palavras-chave. O sistema implementado é ainda independente do género e domínio dos textos de entrada, tendo a capacidade de processar textos nos idiomas Português e Inglês.
A implementação do sistema foi levada a cabo com resurso à linguagem de programação *JAVA*, estando a sua estrutura geral e as dependências entre os diversos pacotes de classes, representadas na Figura 3.1.

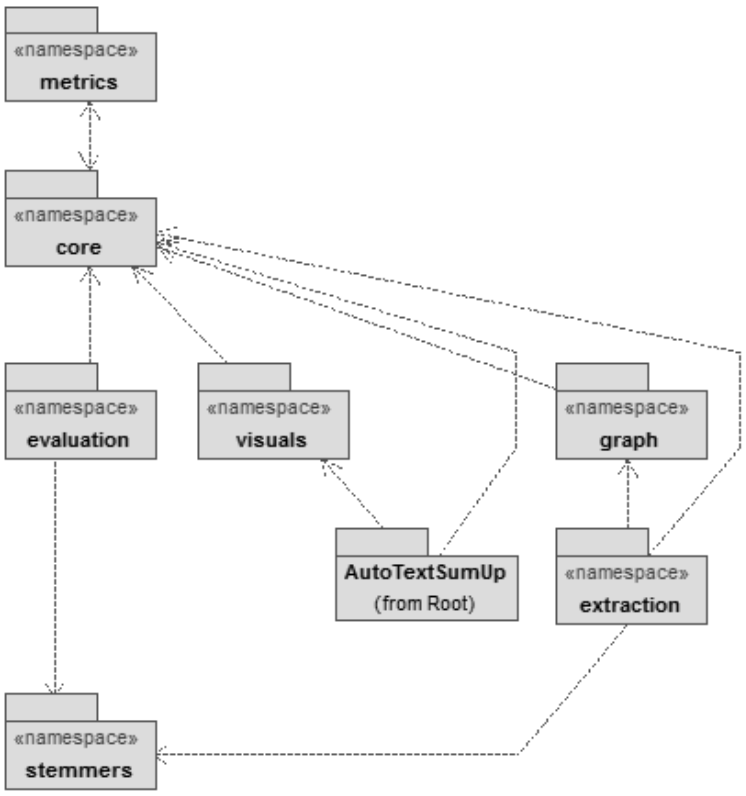

Figura 3.1: Diagrama da estrutura geral do sistema.

No pacote de classes *core* estão contidas classes que permitem manipular os textos a sumarizar e guardar as informações relativas aos mesmos, permitem ainda operações de entrada/saída e processamento de corpus, etiquetagem e filtragem morfossintáctica, e operações com ficheiros. Em *metrics* encontra-se a classe de cálculo da relevância de termos. No pacote *graph* estão encerradas as classes que possibilitam a representação e manipulação de gafos. As classes que permitem operações de lematização estão contidas em *stemmers*. O pacote de classes *extraction* contém as classes que descrevem os métodos de classificação e extracção usados. Em *evaluation* estão contidas as classes usadas para o processamento dos dados de teste usados para a avaliação do sistema. No pacote *visuals* encontram-se as classes responsáveis pela representação visual e definição de funcionalidade da interface do sistema.

O sistema desenvolvido contempla a interpretação e processamento de essencialmente dois formatos de texto, texto simples e texto marcado com hipertexto, através de três métodos distintos de entrada.

O ponto de partida e principal foco do sistema são a interpretação, pré-processamento e sumarização de documentos de aglomerados de notícias, organizados com a estrutura ilustrada na Figura 3.2, através de marcação *XML*.

```
<?xml version="1.0"?>
<news-clusters>
<cluster i="0"  url="endereço de um sítio de notícias">
  <new i="0"  url="endereço da notícia">
              …
              corpo da notícia
              …
  </new>
  <new i="1" url="endereço da notícia">
              …
              corpo da notícia
              …
  </new>
  <new i="2"  url="endereço da notícia">
              …
              corpo da notícia
              …
  </new>
</cluster>
<cluster i="1"  url="endereço de um sítio de notícias">
  <new i="0"  url="endereço da notícia">
              …
              corpo da notícia
              …
  </new>
  <new i="1" url="endereço da notícia">
              …
              corpo da notícia
              …
  </new>
</cluster>
</news-clusters>
```

Figura 3.2: Estrutura de um documento de aglomerados de notícias.

Estes ficheiros são criados, no formato descrito, por uma ferramenta automática de indexação de textos noticiosos [CDB07].

Cada notícia contida num documento de aglomerados de notícias é sumarizada individualmente.

Os restantes métodos de entrada contemplam a recuperação de texto simples da área de transferência do sistema operativo, e a recuperação de textos (notícias, artigos, etc.) ou documentos de texto completos nos formatos mais comuns a partir de endereços WWW. Estes métodos de entrada foram adicionados numa óptica de maior abrangência e facilidade de utilização do sistema.

O carregamento, interpretação e segmentação iniciais, dos ficheiros de aglomerados de notí-

cias, são levados a cabo pela biblioteca *Hultig* [fHLTB]. A recuperação e a análise interpretativa dos conteúdos de ligações WWW, são realizadas pelas bibliotecas *Apache Tika* [Foub], um conjunto de ferramentas para detecção e extracção de metadados e conteúdo textual estruturado a partir de documentos, e *Boilerpipe* [Koh], uma biblioteca que permite encontrar e extrair o principal conteúdo textual (notícia, artigo, etc.) de uma página web.
Para a identificação do conteúdo relevante no texto fonte foram implementados um conjunto de métodos relativos às diferentes abordagens descritas. Foi ainda implementado um método que combina os diferentes métodos numa abordagem híbrida que se caracteriza pela avaliação combinada de diferentes atributos do texto fonte.
De um modo geral a extracção é feita da seguinte forma:

1. Inicialmente o texto fonte é pré-processado para produzir uma representação computável;
2. Classificam-se as diferentes frases do texto original com determinada pontuação, em função da lógica dos métodos implementados;
3. Por fim as frases com melhor pontuação vão sendo incluídas no sumário até que a taxa de compressão desejada seja atingida.

Como já referido, a taxa de compressão de determinado sumário (Equação 3.1), é a relação entre o tamanho do sumário produzido (T_s) e o tamanho do texto original (T_o), dada por:

$$C(T_s) = 1 - \frac{\|T_s\|}{\|T_o\|} \quad (3.1)$$

À medida que a taxa de compressão aumenta o sumário torna-se mais conciso mas mais informação é perdida. Sumários extractivos produzidos de apenas um texto fonte pretendem geralmente ter de 5 a 30 % do tamanho do texto original. No entanto, as metas de compressão para síntese de múltiplas fontes ou na geração de resumos para dispositivos portáteis são muito menores. Estas elevadas taxas de redução são um desafio porque são difíceis de alcançar sem uma quantidade razoável de conhecimento linguístico profundo [HM00].

3.2 Pré-Processamento

A sumarização extractiva de texto é geralmente realizada em duas etapas, sendo a primeira o pré-processamento do documento de entrada, para produzir uma representação estruturada sobre a qual seja possível operar computacionalmente, e a segunda o processamento da representação gerada por métodos que tentam encontrar o conteúdo saliente do texto original.
A etapa de pré-processamento assume assim um importante papel, uma vez que sem uma representação adequada do texto original é impossível fazer uma avaliação do seu conteúdo relevante de forma eficaz e eficiente. Esta fase do processo envolve geralmente as seguintes operações:

- **Análise sintáctica do texto (*Parsing*)**: permite a identificação das palavras, geralmente separadas por espaços, e de características estruturais do texto de entrada, como os limites das frases, geralmente assinalados pela presença de pontos finais. Desta forma as cadeias de caracteres lidas dos documentos de entrada são transformadas em vectores de unidades textuais significativas. No caso de documentos de texto marcados com

hipertexto esta etapa poderá envolver ainda a remoção do hipertexto recuperando a informação que este possa fornecer. No sistema implementado, os documentos de aglomerados de notícias são analisados para que toda a estrutura *XML* possa ser separada do contudo textual propriamente dito;

- **Eliminação de maiúsculas/minúsculas (*Case folding*)**: consiste em converter todas as letras das palavras para o mesmo formato (maiúscula ou minúscula), é uma forma simples de "normalizar" o texto para efeitos de comparação. Um exemplo seria a substituição da letra "I" pela letra "i" na palavra "Informação", resultando numa comparação positiva desta com a palavra "informação", o que poderia não acontecer caso contrário;

- **Eliminação de palavras irrelevantes (*Stopwords*)**: diz respeito à eliminação de palavras comuns, sem valor semântico, que não fornecem por isso qualquer informação relevante, tornando-se insignificantes. Foram compiladas manualmente, listas de palavras irrelevantes para Português e Inglês, para utilização pelo sistema desenvolvido, com comprimentos de 337 e 319 palavras respectivamente;

- **Lematização**: é o processo de redução das palavras à sua forma canónica, através da remoção de sufixos. Isto permite evidenciar a sua semântica. Uma das mais amplamente usadas soluções de lematização é a apresentada por Porter [Por80], inicialmente restrita à língua inglesa, está actualmente disponível para vários idiomas incluindo o Português.

No sistema implementado a análise sintáctica do texto e a redução de maiúsculas é levada a cabo com recurso às ferramentas disponibilizadas pela biblioteca *Hultig* [fHLTB], a remoção de palavras irrelevantes é levada a cabo directamente pelo sistema, com base nas listas compiladas, e a lematização, tanto para Português como para Inglês, é possibilitada pelas diferentes implementações, em linguagem *JAVA*, de lematizadores disponibilizados por Porter [Por80].

3.3 Ferramenta de Processamento de Corpus

A utilização de métodos de sumarização automática de texto que baseiam a sua avaliação, da saliência das frases, na informação das frequências dos termos do texto, cria a necessidade de um registo de tamanho significativo de frequências de termos em geral, para que seja possível fazer a análise da relevância de um termo em determinado documento, quando comparado com essa informação (a sua frequência em geral). Para possibilitar a construção deste tipo de registo foi desenvolvida uma ferramenta de processamento de corpus (Figura 3.3).
Esta ferramenta permite a análise de corpus de documentos de texto simples e de documentos de aglomerados de notícias, como os discutidos anteriormente, no sentido de produzir estruturas onde a frequência generalizada de termos possa ser armazenada. Para que estes valores sejam significativos, o corpus deve ter dimensões consideráveis.
Esta informação é guardada em estruturas derivadas da estrutura *Hashtable* da linguagem *JAVA*, que permite o mapeamento entre os termos e as suas frequências. Esta estrutura faz ainda o registo total de termos analisados, o que permite o cálculo das suas probabilidades.
A ferramenta desenvolvida permite guardar as estruturas criadas no sentido destas poderem voltar a ser utilizadas posteriormente sem necessidade de novo processamento, uma vez que este pode ser um processo moroso, especialmente para corpus de grandes dimensões.
Existe ainda a possibilidade de usar um registo de frequências, previamente guardado, como ponto de partida para um novo processamento. No final de cada processamento é possível fazer

o carregamento da informação do corpus processado para a aplicação, de forma automática, para depois dar início ao processo de sumarização.

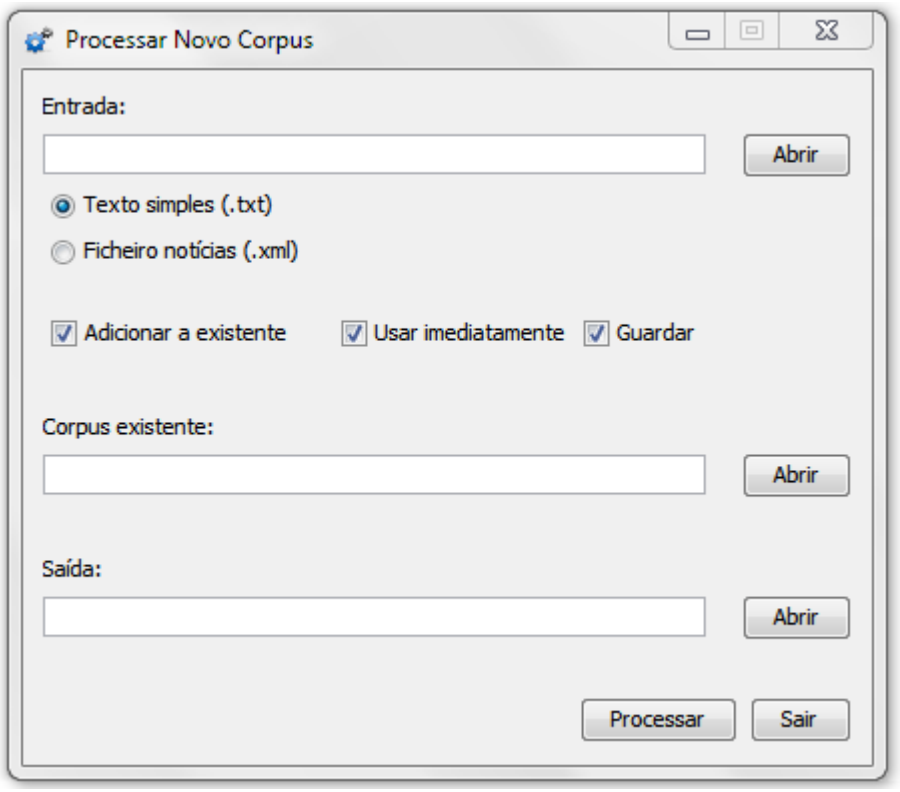

Figura 3.3: Ferramenta de processamento de corpus.

3.4 Métodos de Extracção Implementados

Foram implementados três métodos distintos de sumarização baseados em diferentes abordagens, um método puramente estatístico, um método baseado em teoria de grafos, e finalmente um método que combina diferentes características para fazer a avaliação. Estes métodos serão agora descritos em maior detalhe.

3.4.1 Relevância de Termos

A essência deste método assenta nos princípios clássicos, segundo os quais se considera que termos importantes aparecem frequentemente num documento mas menos frequentemente num contexto geral. Assim, este método faz uso das frequências das palavras no documento de entrada, e das frequências contadas a partir de um corpus de grandes dimensões, como ilustrado anteriormente (Secção 3.3), para avaliar a relevância dos termos do documento de acordo com a Equação 3.2:

$$Re(p|D) = F(p|D) \times \log \frac{P(p|D)}{P(p)} \quad (3.2)$$

Assim, a relevância ($Re(p|D)$) de uma palavra (p) num documento (D), é determinada pela frequência ($F(p|D)$) com que p ocorre em D, influênciada pela relação entre a probabilidade

($P(p|D)$) de p estar presente em D e a probabilidade geral ($P(p)$) de ocorrência de p com base na informação recolhida de um corpus.
Esta métrica pretende evidenciar palavras contidas no documento, que sejam mais relevantes no contexto do documento do que num contexto geral.
Para tornar possível a utilização deste método, foram processados um corpus de Português e um corpus de Inglês com as características descritas na Tabela 3.1.

Tabela 3.1: Dimensões dos corpus processados.

Idioma	Número Total de Termos Únicos	Número Total de Termos Analisados
Português	394 021	76 308 135
Inglês	128 118	12 156 044

Uma vez calculados os valores de relevância dos termos de determinado documento (D) com n_f frases, definido por $D = \{f_1, f_2, \ldots, f_{n_f}\}$, a classificação ($C(F_i)$) da i-ésima frase (F_i) do documento é dada pela Equação 3.3:

$$C(F_i) = \frac{1}{n_p} \times \sum_{j=1}^{n_p} Re(p_j) \quad (3.3)$$

Onde n_p é o número de palavras da frase (F_i) consideradas informativas, ou seja, palavras não presentes nas listas de palavras irrelevantes compiladas, e $Re(p_j)$ é o valor de relevância da j-ésima palavra relevante de F_i.
As frases são depois ordenadas por ordem decrescente das suas pontuações, sendo de seguida escolhidas as frases com melhor pontuação até que seja atingida a taxa de compressão desejada.
Este método tem a vantagem de ser de simples implementação, tendo também um bom desempenho computacional.

3.4.2 *TextRank*

Como já referido (SubSecção 2.3.2), os autores [MT04] deste método propõem a sua utilização em duas vertentes, a extracção de palavras-chave e a extracção de frases, estas duas vertentes foram também exploradas no âmbito desta dissertação.
Como sugerido anteriormente, independentemente da vertente de aplicação que se pretenda explorar, este método segue o modelo geral descrito no capítulo anterior, que tem no seu cerne o uso de algoritmos de classificação baseados em teoria de grafos.
Como observado pelos autores, existem vários algoritmos [MT04, Mih05, MT05] que podem ser usados, não introduzindo qualquer alteração à estrutura geral do método, no caso do sistema desenvolvido foi usado o algoritmo *PageRank* [BP98]. A equação de cálculo sucessivo da pontuação dos nós do grafo usada, no entanto, é uma variante da equação original do algoritmo *PageRank*, esta alteração sugerida em [MT04] permite ter em conta a magnitude das ligações entre unidades textuais representadas no grafo.
No contexto original do algoritmo *PageRank* - a navegação web - é pouco habitual a existência de múltiplas referências, ou referências parciais, de uma página para outra, pelo que uma ligação no grafo é de natureza binária, no entanto, em conteúdo textual baseado em linguagem natural é possível que este tipo de ligações parciais ou múltiplas possa existir, pelo que será útil avaliar magnitude da ligação entre duas entidades textuais representadas no grafo.

Para permitir a representação de determinado documento numa estrutura sobre a qual seja possível agir computacionalmente, define-se formalmente $G = (N, A)$ como sendo um grafo direccionado e pesado, com um conjunto de N nós e um conjunto de A arestas, em que A é um subconjunto de $N \times N$. Para um dado nó N_i, seja $En(N_i)$ o conjunto de nós que apontam pra si, e $Sa(N_i)$ o conjunto de nós para os quais o nó N_i aponta, a classificação ($C(N_i)$) de N_i é dada por:

$$C(N_i) = (1-d) + d \times \sum_{N_j \in En(N_i)} \frac{p_{ji}}{\sum_{N_k \in Sa(N_j)} p_{jk}} \times C(N_j) \quad \textbf{(3.4)}$$

Onde d é um factor de atenuação, que tem o papel de representar a probabilidade do salto de um dado nó do grafo para outro de forma aleatória. Este parâmetro pode assumir valores entre 0 e 1, sendo que geralmente é fixado em $0,85$ [MT04, BP98]. $C(N_j)$ é a classificação de cada um dos nós (N_j) pertencente a $En(N_i)$, ou seja, os nós que recomendam N_i, sendo p_{ji} o peso associado a cada uma dessas recomendações. Os nós (N_k) pertencentes a $Sa(N_j)$, são o conjunto total de nós que cada nó N_j recomenda, sendo o peso destas recomendações dado por p_{jk}.
$G = (N, A)$ providencia assim a representação estruturada do conteúdo textual de um documento, sobre a qual o algoritmo actua computacionalmente de acordo com a Equação 3.4, partindo de valores arbitrários atribuídos aos nós do grafo, o algoritmo faz a computação sucessiva das pontuações de cada nó, até que um valor de convergência, inferior a determinado limite (δ) seja atingido. Os valores inicialmente atribuídos a cada nó não influenciam a pontuação final dos nós, poderão apenas influenciar o número de iterações até que seja atingida convergência.

$$erro = |C^{k+1}(N_i) - C^k(N_i)| \quad \textbf{(3.5)}$$

A convergência é testada, para cada nó, com base no valor do $erro$ entre pontuações sucessivas do nó (Equação 3.5). O valor do erro generalizado ($erro_g$) é calculado, para todos os nós, com base na distância euclidiana, uma vez que, esta métrica simples permite o cálculo da distância entre dois elementos ($C^{k+1}(N_i)$ e $C^k(N_i)$) num espaço n-dimensional, sendo neste caso a dimensão do espaço o número de nós do grafo. É atingida convergência se $erro_g < \delta$. Existem outras métricas que poderão ser exploradas para efectuar o teste de convergência deste tipo de algoritmos [Ber05].
O valor limite usado para o erro foi o de $0,0001$.
Definida a infra-estrutura geral de classificação baseada em teoria de grafos, ver-se-á de seguida de que forma são construídas as estruturas representativas do conteúdo textual que permitem levar a cabo a concretização das duas tarefas para as quais o método foi utilizado. De uma forma geral o conteúdo textual é representado por um grafo (Figura 3.4) em que os nós representam unidades textuais e as arestas entre nós representam um qualquer tipo de relação entre as entidades textuais, representadas pelos nós, que possa ser considerada útil.
A especialização do exemplo apresentado na Figura 3.4, para que este se adeque à realização da tarefa que se pretenda concretizar, significa definir a granularidade das unidades textuais (palavras, frases, etc.) representadas pelos nós do grafo (*UT1* a *UT5*), definir o tipo de relações a considerar e determinar os pesos das ligações definidas em função dessas relações.

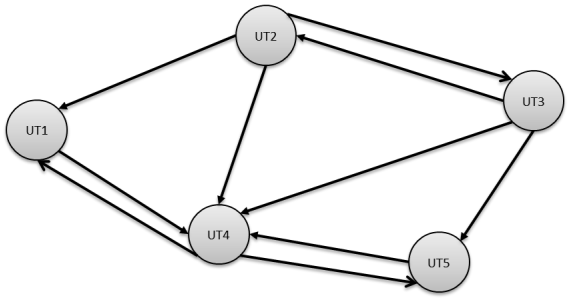

Figura 3.4: Exemplo de um grafo $G = (N, A)$, para $N = 5$ e $A = 10$.

Extracção de Palavras-Chave

Nesta aplicação a ideia é realizar a extracção de termos ou expressões significativas de um dado texto ou documento. Neste sentido, no grafo construído para levar a cabo esta tarefa os nós representam termos do texto, e as arestas do grafo representam relações de co-ocorrência entre esses termos. Este tipo de ligações permite a representação de relações entre elementos sintácticos do texto, que são geralmente indicadores de coesão textual [MT04].

As ligações de co-ocorrência são controladas pela distância entre a ocorrência dos termos, ou seja, dois termos estão relacionados e consequentemente os nós que os representam estão ligados, se estes termos co-ocorrerem dentro de uma janela de N palavras, em que N poderá ser definido entre 2 e 10 palavras.

As unidades léxicas escolhidas para representar nós no grafo podem também ser restringidas através do uso de um filtro sintáctico, que limite os termos escolhidos a determinada classe ou classes gramaticais. A configuração usada contemplou apenas o uso de nomes e adjectivos, e uma janela para co-ocorrências de 2 palavras, uma vez que esta foi a configuração que obteve melhores resultados [MT04].

Para tornar possível a selecção das unidades léxicas descritas é necessário levar a cabo a etiquetagem morfossintáctica dos termos. Esta tarefa é realizada com o auxílio do etiquetador morfossintáctico da biblioteca *OpenNLP* [Foua]. Esta ferramenta permite a etiquetagem de termos em diferentes idiomas, com base em modelos previamente treinados nesses idiomas. No sistema desenvolvido, foram incluídos os modelos de etiquetagem morfossintáctica para Português e Inglês. Para o processo de selecção propriamente dito foi desenvolvido um filtro sintáctico, que com base nas etiquetas morfossintácticas atribuídas aos termos [MSM94, FRB08], decide a sua inclusão ou exclusão do grafo.

A tarefa de extracção de palavras-chave é então levada a cabo, de acordo com a configuração descrita a seguir. Inicialmente o grafo é construído, termos que passem no filtro sintáctico são adicionadas como nós e ligações são criadas para co-ocorrências entre esses termos, dentro da janela prevista. A pontuação dos nós é inicializada com o valor 1. O algoritmo de classificação é corrido sobre o grafo até que seja atingida convergência. Finalmente, os nós são ordenados pela ordem inversa das suas pontuações, e os nós com melhor pontuação são extraídos. No sistema implementado são extraídas do texto fonte 20 palavras-chave.

Extracção de Frases

Na aplicação deste método à sumarização automática de texto, uma vez que as unidades textuais a serem classificadas e extraídas são frases, cada frase do texto é representada no grafo por um nó. A relação entre frases é determinada pela sua semelhança, sendo que esta semelhança é calculada em função da repetição de conteúdo entre frases. Duas frases que se refiram aos

mesmos conceitos são consideradas próximas. Assim, dadas duas frases F_i e F_j, sendo uma frase definida por $F_i = \{p_1^i, p_2^i, \ldots, p_{n_i}^i\}$, onde n_i é o número de palavras da frase, a relação de semelhança entre elas, é dada pela Equação 3.6:

$$Si(F_i, F_j) = \frac{|\{p_k | p_k \in F_i \& p_k \in F_j\}|}{\log(|F_i|) + \log(|F_j|)} \quad (3.6)$$

Para além desta, outras medidas (similaridade do co-seno, subsequência comum mais longa, etc.) podem ser usadas para calcular a semelhança entre frases [MT04, AHS08].
A semelhança entre frases pode ser calculada tendo em conta todos os termos das frases ou pode ser aplicado o filtro sintáctico descrito anteriormente para que a comparação de conteúdo seja baseada apenas em termos de determinadas classes gramaticais.
Depois de construído o grafo, da mesma forma que para a extracção de palavras-chave, a pontuação das frases é inicializada a 1, o algoritmo de classificação é depois corrido no grafo e as pontuações finais são obtidas. As frases são ordenadas por ordem decrescente da sua pontuação e as frases melhor classificadas são extraídas até que a taxa de compressão pretendida seja atingida.
Mesmo sendo capaz de representar relações de coesão textual, através do relacionamento de frases que se refiram aos mesmos conceitos, este método tem a vantagem de não requer conhecimento linguístico profundo.

3.4.3 Método Híbrido

Numa abordagem semelhante à seguida por Edmundson [Edm69], foi implementado um terceiro método que baseia a sua classificação das frases de um texto em quatro características diferentes.
As características consideradas são a relevância dos seus termos, a sua centralidade, as palavras-chave que contém, e a sua posição no texto. As pontuações obtidas em função destas características são depois combinadas linearmente, no sentido de produzir a pontuação final. Esta combinação de pontuações é mediada por pesos associados a cada componente, e é dada pela Equação 3.7:

$$C(F_i) = p_r \times R_i + p_c \times C_i + p_p \times P_i + p_l \times L_i \quad (3.7)$$

Onde $C(F_i)$ é a classificação final da i-ésima frase (F_i) do texto fonte, R_i é a pontuação de F_i em função dos valores de relevância dos seus termos, C_i é a pontuação de centralidade de F_i dada pelo método ***TextRank***, P_i é a pontuação de F_i em função das palavras-chave que esta contém, extraídas pelo método ***TextRank***, e L_i é a pontuação de F_i em função da sua posição no texto, dada por uma heurística de posição. Os pesos p_r, p_c, p_p e p_l permitem controlar a importância atribuída a cada uma das componentes.
Para calcular a pontuação final das frases este método calcula primeiro as pontuações parciais para cada característica. Sendo que as pontuações parciais baseadas na relevância dos termos, e na centralidade da frase são directamente obtidas pelos métodos descritos anteriormente.
A pontuação parcial da frase (F_i) baseada nas palavras-chave que esta contém (P_i), é calculada com base nas palavras-chave extraídas pelo método ***TextRank***, através da Equação 3.8:

$$P_i = \frac{1}{n_p^i} \times \sum_{j=1}^{n_p^i} p_j^i \quad (3.8)$$

Onde n_p^i é o número de palavras-chave (p_j^i) que F_i contém.
A pontuação parcial (L_i) da frase (F_i) em função da sua posição (i) no texto fonte, é calculada com base numa heurística de posição que permite explorar a estrutura textual seguida pelos documentos analisados. No caso de textos noticiosos, foco do trabalho desenvolvido no âmbito desta dissertação, frases que contenham informação mais relevante, ocupam normalmente posições cimeiras no texto [ACA10, Nen05]. No sentido de tirar partido deste facto foi adoptada a heurística de posição sugerida em [PB07], dada pela Equação 3.9:

$$L_i = \frac{1}{\sqrt{i}} \quad (3.9)$$

No entanto, outras heurísticas podem ser exploradas no sentido de melhor explorar outros tipos de estruturas textuais. Por exemplo, para a sumarização de artigos científicos poderiam ser consideradas heurísticas que favorecessem secções particulares como o resumo do artigo ou a sua conclusão.
Calculadas as pontuações parciais das frases, os valores são normalizados para o intervalo $[0-1]$ e a pontuação final é calculada em função de determinada combinação de pesos. As frases são depois ordenadas por ordem decrescente das suas pontuações, e as frases melhor classificadas são extraídas até que seja atingida a taxa de compressão pretendida.
Este método é o mais completo dos três, sendo também o que apresenta um maior custo computacional.

3.4.4 Outros Métodos Investigados

Para além dos métodos implementados descritos anteriormente foram ainda investigados outros métodos com vista à sua inclusão no sistema. Em particular, foi investigado o método de classificação por propagação de activação baseado em teoria de grafos [MBG98], descrito no capítulo anterior. Os grafos baseados em relações semânticas construídos por este método poderiam ser representados pela mesma infra-estrutura de representação de grafos desenvolvida para utilização com o método *TextRank*, e permitiriam explorar a coesão textual com base na sua informação semântica, componente que não é investigada pelos métodos implementados.
Esta abordagem tem a vantagem de identificar afinidades entre entidades textuais que vão para além da simples repetição literal de conteúdo. Muitas vezes os mesmos conceitos são identificados por termos diferentes e a simples comparação literal de termos não consegue captar estas noções.
Através da identificação de relações como anáfora de nomes próprios, repetição, adjacência, sinonímia e hiperonímia, e de expressões, são estabelecidos padrões de coesão textual, representados por um grafo, em que os nós representam termos do documento com informações associadas, como a sua posição no texto, pontuação inicial, etc., e as suas arestas representam as relações semânticas descritas. Esta estrutura serve depois como base da computação ao algoritmo de propagação de informação.

A técnica de procura por propagação de activação em redes semânticas, neste caso, o grafo representativo das ligações semânticas referidas, entre termos do texto fonte, é um processo de procura iniciado através da marcação de um conjunto de nós de origem (termos representativos de tópicos ou conceitos relevantes, geralmente determinados por consulta do utilizador) com valores de "activação" máximos, em seguida, iterativamente dá-se a propagação da activação a nós adjacentes aos nós de origem, em função dos pesos das ligações semânticas e de determinado factor de decadência para as ligações. A procura termina quando determinados critérios de paragem forem atingidos, tais como: um determinado limite para o número de nós activados, todos os nós do grafo terem sido activados, não haver uma alteração significativa dos valores de activação dos nós em iteracções sucessivas, etc.
A extracção de frases seria feita com base nas pontuações obtidas por estas, calculadas com base nos valores de activação finais dos seus termos.
Uma vez que o método original [MBG98] prevê a indicção de tópicos como ponto de partida para a pesquisa por propagação de activação, a ideia seria fornecer automaticamente termos tópico ao método, extraídos previamente do texto pelo método *TextRank*. A pontuação inicial dos nós do grafo seria calculada pelo método de relevância de termos, em vez de pela métrica $ti - idf$.
Este método não foi no entanto incluído no sistema desenvolvido, uma vez que as ferramentas [ABD+95, Kru93] usadas pelos autores para a identificação de expressões e extracção de nomes próprios, a usar na construção do grafo não estavam disponíveis e a procura ou implementação de soluções semelhantes que permitissem a identificação de tais relações semânticas para os idiomas Português e Inglês, visados pelo sistema implementado, não seria possível em tempo útil. Relações como sinonímia e hiperonímia poderiam ser extraídas para o Inglês por implementações *JAVA* do *WordNet* [Mil95] que se encontram disponíveis de forma aberta, no entanto, para o Português o mesmo não seria possível, uma vez que, versões do *WordNet* em outros idiomas não se encontram disponíveis no mesmo formato.
Os autores [MBG98] fazem notar que métodos baseados em elementos de coesão textual obtêm melhores resultados quando é possível explorar os diferentes tipos de relações de coesão referidos, não representado por isso, uma melhoria de desempenho significativa a exploração empobrecida de apenas alguns dos tipos de relações de coesão referidos.

3.5 Aplicação Final

Nesta secção é apresentada a aplicação final resultado do trabalho desenvolvido nesta dissertação, que envolve o sistema de sumarização automática de texto desenvolvido e a interface gráfica que permite a sua utilização simplificada. A Figura 3.5, ilustra a vista primária desta interface, com os seus principais componentes assinalados:

1. Permite o carregamento de um ficheiro de aglomerado de notícias para ser sumarizado;

2. Faz o carregamento do conteúdo textual contido na área de transferência do sistema operativo para ser sumarizado;

3. Permite a recuperação do conteúdo textual central de uma página web (artigo, notícia), ou de um documento de texto (em alguns dos formatos mais comuns), a partir de um endereço web;

4. Possibilita o carregamento dos ficheiros que contêm a informação sobre frequência de termos de corpus processados. Depois de carregado um ficheiro, mostra o nome desse ficheiro;
5. Carrega a tabela de informações de corpus, com os termos e respectivas frequências, para visualização;
6. Permite reverter a selecção de ficheiros que contêm a informação sobre frequência de termos de corpus processados previamente carregados;
7. Indica se devem ou não ser marcadas no texto original as frases seleccionadas para fazer parte do sumário;
8. Assinala se devem ou não ser marcadas, no texto fonte, palavras-chaves extraídas;
9. Permite fazer a escolha do idioma a utilizar na sumarização. A escolha do idioma influencia a escolha da lista de palavras irrelevantes e o filtro sintáctico a usar;
10. Abre a ferramenta de processamento de corpus apresentada anteriormente (Secção 3.3);
11. Mostra uma pequena janela de informação sobre a aplicação;
12. Mostra o número da notícia, de um ficheiro de aglomerados de notícias, que está actualmente a ser mostrada, e o número total de notícias desse ficheiro;
13. Permite procurar, sumarizar e mostrar uma notícia, de um ficheiro de aglomerados de notícias, directamente pelo seu índice;
14. Possibilitam a passagem à notícia anterior ou seguinte, de um ficheiro de aglomerados de notícias;
15. Painel onde é mostrado o texto original;
16. Campo para inserção do endereço web de um artigo, notícia ou documento a recuperar e sumarizar;
17. Permite o início do processo de sumarização do conteúdo recuperado a partir do endereço inserido no campo descrito anteriormente;
18. Painel onde é mostrado o resumo;
19. Permite a selecção da taxa de compressão desejada;
20. Assinala a utilização do método de sumarização baseado em relevância de termos;
21. Indica a utilização do método de sumarização *TextRank*;
22. Marca a utilização do método de sumarização híbrido;
23. Permite guardar o texto original, o seu sumário e algumas informações para ficheiro;
24. Barra de estado que mostra informação sobre as tarefas desempenhadas pela aplicação.

Funcionamento Geral

A aplicação tem essencialmente três modos diferentes, em função do modo de entrada escolhido (componentes 1, 2 e 3, na Figura 3.5). Consoante o modo escolhido a interface sofre ligeiras alterações. Quando se pretende processar um ficheiro de aglomerados de notícias os

componentes 12, 13 e 14 (Figura 3.5) são mostrados enquanto os componentes 16 e 17 (Figura 3.5) são ocultados. Quando o modo de entrada escolhido é um endereço web, acontece o inverso. No caso do modo entrada escolhido ser a área de transferência do sistema operativo, todos os componentes referidos são ocultados. Estas alterações permitem maximizar a área de visualização do conteúdo textual.

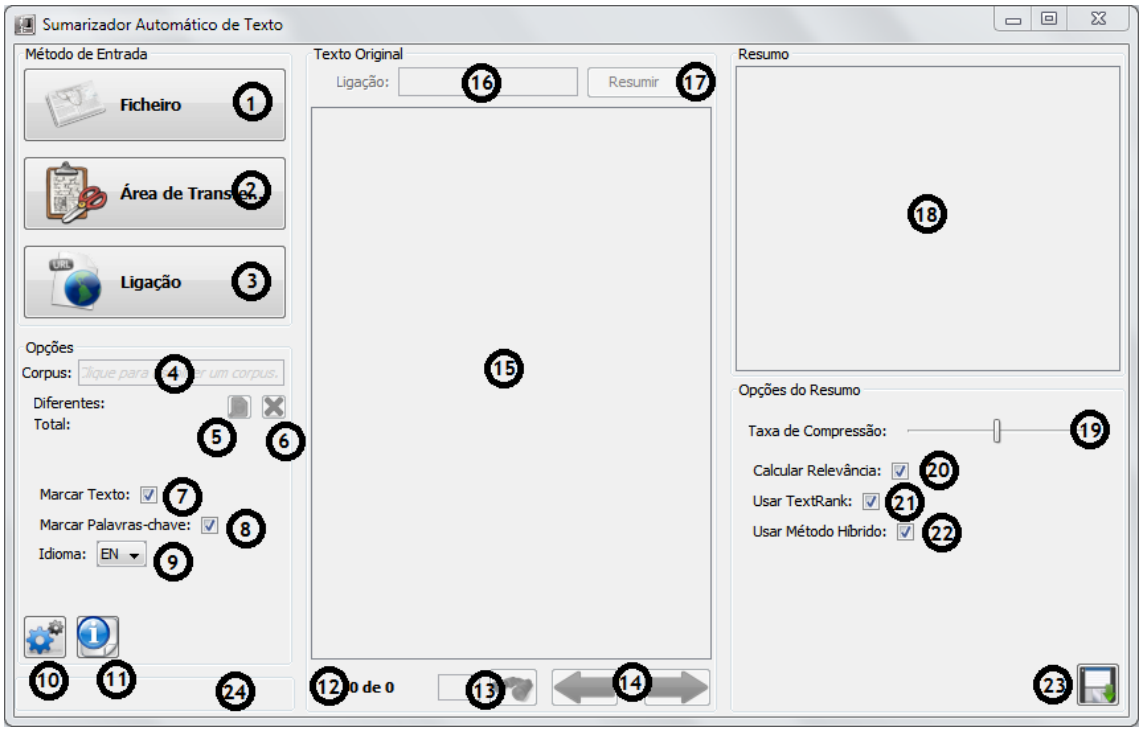

Figura 3.5: Principais componentes da interface da aplicação.

De uma forma geral antes de se iniciar o processo de sumarização, a partir de qualquer um dos modos de entrada apresentados, existem um conjunto de acções que devem ser levadas a cabo. Deve ser carregado um ficheiro de corpus (componente 4 na Figura 3.5), dever ser escolhido o idioma a usar (componente 9 na Figura 3.5), deve ser escolhida a taxa de compressão pretendida (componente 19 na Figura 3.5) e finalmente, deve ser seleccionado o método de sumarização pretendido (componentes 20, 21 e 22 na Figura 3.5).
Para que o processo de sumarização possa ser iniciado estas condições devem ser satisfeitas. Quando se encontram seleccionados vários métodos de sumarização, o método a usar é escolhido pela aplicação pela ordem inversa das suas posições, ou seja, têm maior prioridade o método híbrido e menor prioridade o método baseado em relevância de termos.
Sempre que existir alteração do método de sumarização seleccionado ou da taxa de compressão pretendida, a aplicação calcula automaticamente o novo sumário. Da mesma forma, sempre que forem introduzidas alterações às opções de marcação do texto (componente 7 e 8 na Figura 3.5), essas alterações serão reflectidas automaticamente na visualização do texto.

Exemplo de Utilização

Na Figura 3.6, pode-se observar um exemplo de utilização, neste caso, a sumarização de um ficheiro de aglomerados de notícias, em que a primeira notícia e o respectivo sumário se encontram já expostos.
Na mesma figura é ainda observável a marcação colorida que é feita no texto, através da qual são assinaladas a amarelo as frases extraídas do texto original para formar o sumário, e a tons

de azul as palavras-chave.

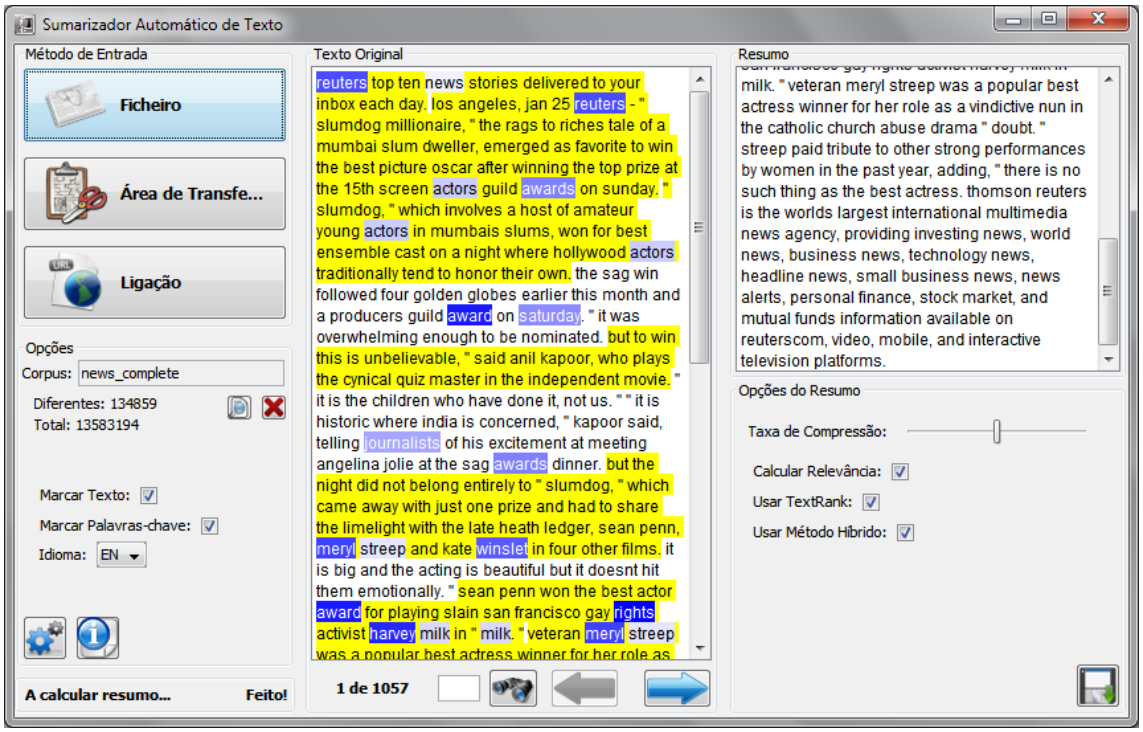

Figura 3.6: Exemplo de utilização.

A intensidade da cor associada à marcação das diferentes palavras-chave é medida pela classificação obtida pelas mesmas, assim, tons de azul mais intensos assinalam palavras-chave mais relevantes. Palavras-chave em posições de adjacência no texto são palavras-chave compostas ou expressões chave. Situações ilustrativas deste tipo de composições, no exemplo apresentado, são "*Meryl Streep*" e "*Harvey Milk*".

A Figura 3.7, ilustra um exemplo de utilização do sistema, apresentanto um sumário produzido pelo mesmo. Neste caso o sumário produzido é obtido a partir do texto "*ft923-6509.txt*" do conjunto de dados DUC 2002 (Secção 4.1), aplicando o método híbrido para uma taxa de compressão de cerca de 75 %.

A partir deste exemplo (Figura 3.7) é possível perceber que as frases escolhidas são bastante informativas da ideia geral do texto original, sendo que neste caso o sumário não sofre de grandes problemas de coerência.

Texto Fonte:

[1] Sir, The argument for an independent central bank is that it should be free of party political interference, not that it should conduct monetary policy without regard to national, regional or global interests. [2] Whatever policy the Kohl government followed over reunification, it was bound to unleash a large federal deficit and inflationary pressures in Germany. [3] As the western German population is now aware, German reunification needs a strong, growing west German economy to finance the subsidisation of eastern Germans' living standards until there is a sustained upswing in the east - even if at the cost of higher German inflation than the Bundesbank would wish. [4] In Europe as a whole, very high real interest rates threaten to turn recession into depression. [5] The governments of Europe have been remarkably patient with the Bundesbank, although it is threatening nearly 10 years of European effort for greater economic and political integration. [6] Surely there must be compromise of some economic habits to achieve the latter? [7] It is ironic that, just when German reunification presented a short-term reason for Germany to shift its anti-inflationary economic and monetary priorities, accommodating European convergence, Bundesbank intransigence is threatening the EMU commitment. [8] Beyond Europe, the dominant economic forces are strongly disinflationary. [9] But Germany is the one economy which, for reasons of reunification, could afford a more relaxed stand on the inflationary front. [10] Yet Mr Otmar Issing, the Bundesbank's chief economist, advises there are no grounds for a cut in Germany's interest rates. [11] This, despite the unreliability of German money supply data. [12] The German government might point out to the Bundesbank that monetary targets might be achievable only at the cost of many more than 3m unemployed in Germany, of the destruction of the EMU programme and global depression. [13] A DM exchange rate adjustment within the ERM is not the answer to the problems. [14] We need a prolonged low-interest-rate German reunification boom. [15] The European currency bloc is overvalued against the dollar and yen as a result of Bundesbank interest rate policies. [16] And this is aggravating recession. [17] A future European-wide central bank must clearly - like the US Federal Reserve - be more accountable in the broader European political and economic interest. [18] The governments and central banks of Europe and the G7 block - including the Kohl administration - should pressurise the Bundesbank, and be prepared to consider constitutional changes if necessary, to halt this idiocy. [19] Howard Flight, managing director, Guinness Flight Global Asset Management, Lighterman's Court, 5 Gainsford Street, Tower Bridge, London SE1 2NE

Sumário:

[3] As the western German population is now aware, German reunification needs a strong, growing west German economy to finance the subsidisation of eastern Germans living standards until there is a sustained upswing in the east even if at the cost of higher German inflation than the Bundesbank would wish. [5] The governments of Europe have been remarkably patient with the Bundesbank, although it is threatening nearly 10 years of European effort for greater economic and political integration. [7] It is ironic that, just when German reunification presented a short-term reason for Germany to shift its anti-inflationary economic and monetary priorities, accommodating European convergence, Bundesbank intransigence is threatening the EMU commitment.

Figura 3.7: Exemplo de um sumário produzido pelo sistema.

Capítulo 4

Avaliação

Uma das questões de maior pertinência mas ao mesmo tempo de maior complexidade da sumarização automática de texto é a avaliação dos resumos produzidos. Esta é uma questão particularmente importante, não só do ponto de vista da avaliação da qualidade dos sumários produzidos ou do desempenho dos sistemas, mas também porque permite ter uma ideia da evolução sofrida por esta área de investigação.

A avaliação da qualidade de um resumo tem provado ser uma questão difícil, isto fica essencialmente a dever-se ao facto de simplesmente não existir uma definição óbvia do que será um resumo "ideal", mesmo quando impostas certas restrições. Como tem sido evidenciado em diversos estudos, existe um baixo grau de concordância na criação e consequentemente na avaliação humana de sumários [Nen06]. Por exemplo, mesmo numa tarefa aparentemente descomplicada como é o resumo de artigos de notícias simples, sumarizadores humanos tendem a atingir níveis de concordância de apenas 60 %, no que diz respeito à sobreposição de conteúdo frásico [RHM02].

Assim, a dificuldade de definir adequadamente o processo de avaliação deve-se essencialmente à subjectividade e falta de concordância inerentes à avaliação levada a cabo por avaliadores humanos. Esta subjectividade na avaliação de sumários está geralmente assente na enorme diversidade de factores que poderão caracterizar os sumários, como a sua legibilidade, inteligibilidade, nível de abrangência da informação fornecida, adequação a tarefas específicas, etc., sendo por isso uma ideia bem estabelecida, a de que não existe um método único e ideal de avaliação.

Num esforço para formalizar processos de avaliação de sistemas de processamento de linguagem natural, Spärck Jones e Galliers [SJG96], sugerem que a avaliação de sistemas de sumarização automática de texto pode ser amplamente classificada como *intrínseca* ou *extrínseca*. No caso da avaliação intrínseca o juízo sobre determinado sistema de sumarização automática de texto é baseado somente no sistema em si e nas características do sumário por si produzido. Por outro lado, a avaliação extrínseca pressupõe a avaliação do sistema em função da sua capacidade de cumprir determinada tarefa, como categorização de documentos, recuperação de informação, ou testes de pergunta-resposta [Man01a].

Na sua maioria os sistemas de avaliação tendem a ser de natureza intrínseca, nestes sistemas a avaliação pode ser feita através da apreciação directa do sumário por avaliadores humanos, de acordo com um conjunto de critérios predefinidos, ou por comparação do sumário produzido automaticamente com um ou mais sumários produzidos por humanos, quer seja por extracção ou abstracção, habitualmente designados de sumários de referência. Dada a referida subjectividade do processo humano de criação e avaliação de sumários, pessoas diferentes criarão sumários diferentes de um mesmo texto fonte, pelo que, de um modo geral são usados mais do que um sumário de referência por forma a obter melhores avaliações. A comparação entre sumários produzidos automaticamente e sumários de referência é geralmente fundamentada na procura de repetição de conteúdo (palavras, expressões ou frases).

A avaliação intrínseca por comparação de sumários é geralmente levada a cabo através do emprego de um conjunto de métricas oriundas do campo da Recuperação de Informação (RI). Estas

métricas são a **Precisão** (*Precision*), **Cobertura** (*Recall*) e **Medida-F** (*F-measure*) [FC99]. Inicialmente usadas como indicadores da relevância de documentos recuperados, no campo da RI, no âmbito da sumarização automática de texto estas métricas podem ser entendidas como, no caso da **Precisão** (Equação 4.1) a relação entre o número de frases do sumário automático presentes no sumário de referência, e o número frases do sumário automático; a **Cobertura** (Equação 4.2) aqui consiste na relação entre o número de frases do sumário automático presentes no sumário de referência, e o número de frases do sumário de referência; a **Medida-F** (Equação 4.3) é uma medida composta que combina os valores de **Precisão** e **Cobertura**, mediante um factor não negativo β que permite ajustar a importância das duas componentes. Considere-se o sumário automático extractivo de um determinado texto, definido por $S_g = \left\{f_1^g, f_2^g, \ldots, f_{n_g}^g\right\}$, em que f_i^g é a i-ésima frase do conjunto de $n_g = \|S_g\|$ frases do sumário. Dado o sumário de referência para o mesmo texto, definido por $S_r = \left\{f_1^r, f_2^r, \ldots, f_{n_r}^r\right\}$ de tamanho $n_r = \|S_r\|$, os valores de **Precisão**, **Cobertura** e **Medida-F** para S_g são dados respectivamente por:

$$P(S_g) = \frac{\|S_r \cap S_g\|}{\|S_g\|} \quad \textbf{(4.1)}$$

$$C(S_g) = \frac{\|S_r \cap S_g\|}{\|S_r\|} \quad \textbf{(4.2)}$$

$$F_\beta(S_g) = \frac{(1+\beta^2) \times P(S_g) \times C(S_g)}{(\beta^2 \times P(S_g)) + C(S_g)} \quad \textbf{(4.3)}$$

Os valores de **Precisão**, **Cobertura** e **Medida-F** variam entre 0 e 1 sendo que valores de **Precisão** máximos significam que todas as frases do sumário gerado automaticamente estão incluídas no sumário de referência, e valores de **Cobertura** máximos significam que todas as frases do sumário de referência estão presentes no sumário automaticamente gerado.

As métricas **Precisão** e **Cobertura** são complementares, variando de forma inversa, quando uma aumenta a outra tende a diminuir. A dimensão desta complementaridade é dada pelo factor β. Valores elevados de β beneficiam a **Cobertura** e valores baixos a **Precisão**. Para valores $\beta = 1$ as duas componentes têm igual importância.

Num estudo dos diversos paradigmas de avaliação, quer intrínsecos quer extrínsecos, Jing *et al.* [JBME98], concluem que diversos factores como, a concordância entre os produtores profissionais na elaboração de sumários de referência, o tamanho do sumário automático produzido, a influência da formulação das métricas de precisão e cobertura, o nível de dificuldade das perguntas, em avaliações do tipo pergunta-resposta, e as características dos documentos, podem influenciar grandemente as avaliações obtidas. Em particular, os autores [JBME98] consideram que as métricas **Precisão** e **Cobertura** poderão não ser indicadas para a avaliação de sistemas de sumarização automática de texto, uma vez que, a troca de uma frase do sumário automático por outra igualmente informativa, mas que não se encontre no sumário de referência, penalizaria injustamente a avaliação do sistema. Outra questão problemática poderá ser variação na granularidade das unidades textuais consideradas para calcular o tamanho do sumário em função da taxa de compressão pretendida.

Tradicionalmente a avaliação da sumarização é levada a cabo por apreciação directa de diferentes métricas de qualidade, como coerência, concisão, gramaticalidade, legibilidade e conteúdo,

por avaliadores humanos [Man01b], no entanto este é um processo custoso, subjectivo e muito dificilmente escalável a aplicações de larga escala, pelo que, a questão de como avaliar automaticamente sumários produzidos, tem vindo a despertar grande interesse na comunidade de investigação da sumarização automática. Este interesse tem vindo a motivar a apresentação de algumas soluções que permitem a avaliação automática de sistemas de sumarização automática, normalmente baseando a sua avaliação na comparação automática dos sumários produzidos, com sumários de referência produzidos por humanos.

Um dos mais relevantes processos deste género é o *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) [LH03, Lin04b, Lin04a]. Um sistema inspirado no sistema de avaliação de sistemas de tradução automática da *International Business Machines* (IBM) denominado *Bilingual Evaluation Understudy* (BLUE) [PRWZ02], que consiste num conjunto de métricas que procuram avaliar a qualidade de sumários produzidos automaticamente, medindo a semelhança entre estes sumários produzido automaticamente e um ou mais sumários de referência através da procura de repetição de n-gramas.

As abordagens automáticas de avaliação intrínseca têm vantagens evidentes, como a sua maior simplicidade e eficiência, e a eliminação da subjectividade humana do processo de avaliação, para permitir avaliações objectivas e uma comparação mais justa entre sistemas de sumarização automática, mas as suas limitações são ainda muitas. Apesar de automáticas na comparação, estas abordagens ainda têm necessidade de sumarizadores humanos especializados, para a produção dos sumários de referência. Outra desvantagem destas abordagens é que não avaliam certas características como a coesão, coerência, e inteligibilidade dos sumários.

Assim, e apesar do enorme progresso que as técnicas de avaliação de sistemas de sumarização automática de texto tem sofrido, existe ainda um enorme potencial de evolução e muitas questões pertinentes a resolver. Avanços neste campo têm-se em muito ficado a dever aos inúmeros seminários e conferências que têm sido organizados no sentido de concentrar esforços na resolução das questões mais pertinentes. Exemplo deste tipo de iniciativas são as conferências TIPSTER SUMMAC (*Text Summarization Evaluation Conference*) [MHK+99], considerada o primeiro grande esforço conjunto nesta área, e *Document Understanding Conferences* (DUC) [DUC07], actualmente denominadas *Text Analysis Conferences* (TAC) [TAC12], que propõem tarefas específicas a serem resolvidas pelos sistemas de sumarização, comparando o desempenho de sistemas propostos da resolução dessas tarefas.

Como já referido, uma parte fundamental do processo de construção de sistemas de sumarização automática de texto é a sua avaliação. Tão importante como tentar definir, teórica ou experimentalmente, um modelo ideal de sumarização automática de texto, é avaliar a eficiência do modelo desenvolvido, por forma a medir o progresso atingido em relação a iniciativas anteriores. No entanto, este processo está longe de ser trivial e a avaliação pode ser feita de diversas formas e com base em diversos factores, como discutido anteriormente.

A avaliação feita do sistema desenvolvido no âmbito desta dissertação pode ser classificada como intrínseca, uma vez que a qualidade dos sumários é avaliada de forma isolada, tendo em conta apenas as suas características próprias, e não em função da sua capacidade de cumprir uma tarefa específica. A avaliação foi feita de forma automática, sem intervenção de avaliadores humanos e tendo em conta apenas o produto final do sistema, o sumário produzido, e não dos seus processos internos.

A avaliação do sistema é obtida através da comparação automática dos sumários por si produzidos com sumários de referência, para um conjunto de textos. Esta comparação é feita através da ferramenta ROUGE.

Neste capítulo são apresentados e discutidos três elementos essenciais do processo de avalia-

ção: o conjunto de materiais de teste, os métodos de avaliação e os resultados obtidos.

4.1 Conjunto de Dados de Teste

Para que seja possível a avaliação do desempenho de um sistema de sumarização automática por meio de comparação dos sumários por si produzidos com sumários de referência, é necessário que existam conjuntos de dados que agreguem um texto original do qual se possam produzir resumos, e um conjunto de resumos de referência do texto original, com os quais se possam comparar os resumos produzidos.
Tais conjuntos de dados são por exemplo os disponibilizados pelas conferências DUC, grandes impulsionadoras de progresso na área da sumarização automática e da sua avaliação. Estes conjuntos de teste amplamente usados para avaliação automática de sistemas de sumarização, são compostos por aglomerados de textos noticiosos e respectivos conjuntos de sumários de referência.
Para avaliação do sistema desenvolvido foram utilizados em particular e de forma parcial os conjuntos de dados DUC gerados nos anos 2001 e 2002, para a tarefa de criação de sumários de 100 palavras a partir de documento único, uma vez que, em anos posteriores esta tarefa deixou de fazer parte da agenda destas conferências.
A não utilização integral dos conjuntos de dados prede-se com o facto de alguns dos textos se repetirem por vários aglomerados, tendo sido feita uma selecção de textos não repetidos. Além disso alguns dos documentos contemplavam anotações no texto que poderiam introduzir ruído nos sumários produzidos, e uma vez que não seria exequível a análise ou remoção deste conteúdo em tempo útil, ficheiros contendo este tipo de anotações foram também excluídos.
O conjunto de dados de teste usado para avaliação do sistema desenvolvido, tendo em conta as contingências referidas, respeita a configuração apresentada na Tabela 4.1.

Tabela 4.1: Informações do conjunto de teste.

DUC	Nº de Textos	Nº Sumários de Referência	Tamanho Médio dos Sumário de Referência [1]
2001	92	3	111
2002	62	2	112

4.2 Método de Avaliação

Como referido anteriormente a avaliação do sistema desenvolvido foi feita através da ferramenta ROUGE. Este é um pacote de métricas de avaliação automática de sistemas de sumarização automática que permite avaliar de forma objectiva a qualidade dos sumários produzidos.
Esta ferramenta de avaliação tem sido amplamente usada, não apenas na avaliação de projectos de investigação independentes, mas também como ferramenta oficial de avaliação em larga escala em algumas tarefas das conferências DUC.
Foi demonstrado em testes realizados nos conjuntos de dados DUC de 2001, 2002 e 2003 que esta ferramenta atinge elevados níveis de correlação com juízos humanos de avaliação da sumarização, particularmente para sumarização de documentos únicos [Lin04b].

[1] Medido em número de palvras.

O pacote de avaliação ROUGE inclui quatro tipos de métricas: ROUGE-N, ROUGE-L, ROUGE-W e ROUGE-S. Estas métricas são baseadas em estatísticas de co-ocorrência de n-gramas, em que o comprimento considerado dos n-gramas pode variar.
Para avaliação do sistema desenvolvido foram usadas especificamente as métricas ROUGE-1 e ROUGE-2, ou seja, a métrica ROUGE-N para n-gramas, com n = 1 e n = 2, uma vez que estas são as métricas que registam melhores níveis de correlação com juízos humanos de avaliação de sumários de 100 palavras, adequando-se assim aos dados de teste seleccionados. Mais ainda, estas são métricas amplamente usadas, permitindo assim a comparação dos resultados obtidos com os de outros sistemas previamente testados.
A métrica ROUGE-N é uma medida de **Cobertura**, no entanto, aqui as unidades sobre as quais se avalia o nível de **Cobertura** são n-gramas e não frases. A medida de **Cobertura** é dada por comparação do sumário gerado automaticamente com um conjunto de sumários de referência criados manualmente. Esta métrica é formalmente definida [Lin04b] pela Equação 4.4:

$$\text{ROUGE-N} = \frac{\sum_{S_r \in R} \sum_{gram_n \in S_r} Count_{match}(gram_n)}{\sum_{S_r \in R} \sum_{gram_n \in S_r} Count(gram_n)} \quad (4.4)$$

Onde $gram_n$ representa cada n-grama de tamanho n de dado sumário de referência (S_r), $Count_{match}(gram_n)$ e $Count(gram_n)$ são respectivamente, o número máximo de de n-gramas presentes simultaneamente no sumário gerado automáticamente que está a ser avaliado e no conjunto dos sumários de referência (R), e o número total de n-gramas do conjunto dos sumários de referência.
Esta métrica representa assim a razão entre o número de n-gramas que co-ocorrem no sumário automático e no conjunto de sumários de referência, e o número total de n-gramas presentes no conjunto de sumários de referência.
Os valores de **Precisão**, **Cobertura** e **Medida-F** são calculados para os sumários automáticos de cada um dos textos fonte do conjunto de dados descrito, com base nos sumários de referência de cada um desses textos. O resultado final de cada métrica é calculado como a média dos valores obtidos para cada texto do conjunto de testes.
Os resultados obtidos para cada uma destas métricas permitem avaliar o nível do conteúdo informativo presente nos sumários gerados automaticamente, quando comparados com um conjunto de sumários de referência, ou seja, estas medidas são indicadores da presença ou ausência do conteúdo que se espera que esteja presente num bom sumário, no sumário gerado automaticamente. No entanto, como já referido, características como a coesão, a coerência, legibilidade, etc., dos sumários automáticos, não são consideradas por estas métricas.

4.3 Resultados

Serão agora apresentados os resultados obtidos nos testes realizados com base na infra-estrutura descrita.
Os conjuntos de textos usados para avaliação do sistema desenvolvido, foram utilizados nas conferências DUC para avaliação de sistemas de sumarização automática de texto, na tarefa de produção de sumário de 100 palavras para um documento único, pelo que o objectivo a concretizar para efeitos de teste, seria a produção por parte do sistema desenvolvido, de sumários dos textos fonte com tamanho de 100 palavras. No entanto, foi verificado que os sumários de referência raramente continham exactamente 100 palavras. Para além disso as

unidades textuais extraídas pelo sistema desenvolvido são frases e não palavras.
No sentido de tentar produzir uma avaliação mais correcta do sistema, o conjunto de sumários de referência foi avaliado, para que pudesse ser determinado com maior exactidão, o tamanho que deveriam ter os sumários gerados automaticamente pelo sistema. Esta informação encontra-se contida na Tabela 4.1.
O tamanho dos sumários produzidos pelo sistema foi ajustado para que este se aproximasse sempre o mais possível do tamanho médio dos sumários de referência de cada conjunto de dados. As frases do sumário não são truncadas, em vez disso, a última frase a adicionar ao sumário que faria com que fosse ultrapassado tamanho desejado (em número de palavras), é adicionada apenas se o número de palavras pelo qual o tamanho desejado seria ultrapassado for inferior à diferença entre o tamanho do sumário sem essa frase e o tamanho desejado.
Os teste realizados contemplaram a avaliação individual de cada um dos métodos implementados, no sentido de permitir uma análise comparativa do desempenho de cada método.
No caso do método híbrido foram testadas algumas combinações de diferentes valores para os pesos das componentes. Nas tabelas de resultados apresentadas a seguir, as diferentes combinações de pesos são representadas no formato **"Híbrido(Pc-Pr-Ppc-Pl)"**, em que **Pc** é o peso atribuído à componente centralidade, determinada pelo método *TextRank*, **Pr** é o peso atribuído à pontuação calculada pelo método de relevância de termos, **Ppc** é o peso atribuído à pontuação obtida pela avaliação das palavras-chave, e **Pl** é o peso atribuído à pontuação resultante da heurística posicional.
Esta combinação de pesos é da forma:

$$[P_c + P_r + P_{pc} + P_l] = 100\%$$

As combinações de pesos para o método híbrido, investigadas no processo de avaliação, pretenderam explorar essencialmente duas vertentes: testar a combinação equilibrada de todas as características consideradas; e tentar apurar a vantagem da inclusão das características não contempladas pelos restantes métodos. Assim, à parte da combinação em que todas as componentes têm pesos iguais, as combinações testadas tendem a favorecer a pontuação de cada frase em função das palavras-chave que contém e da sua posição no texto.
Na Tabela 4.2, são apresentados os resultados dos métodos implementados para a métrica ROUGE-1 no conjunto de textos DUC 2001.

Tabela 4.2: Resultados de Cobertura, Precisão e Medida-F para ROUGE-1 em textos DUC 2001.

ROUGE-1			
DUC 2001	**Cobertura**	**Precisão**	**Medida-F**
Relevância	0,365683606	0,446850048	0,395974274
TextRank	0,413746638	0,409808459	0,409348513
Híbrido(25-25-25-25)	0,417386452	0,433949671	0,422004532
Híbrido(20-20-40-20)	0,394243092	0,423055535	0,403021913
Híbrido(20-20-20-40)	**0,430483149**	**0,459239482**	**0,440906262**
Híbrido(10-10-40-40)	0,416877341	0,447474708	0,42816471

Os resultados apresentados na Tabela 4.2 mostram que de uma forma geral a abordagem híbrida obtém melhores resultados, sendo que, das combinações de pesos testadas para este método apenas uma (**Híbrido(20-20-40-20)**) obtém resultados inferiores aos obtidos pelo método *TextRank*, obtendo ainda assim um valor de precisão mais elevado. O método de sumarização por cálculo da relevância de termos é a abordagem que consegue piores resultados, tendo no

entanto um dos melhores em termos de precisão.
Na Tabela 4.3, são apresentados os resultados dos métodos implementados para a métrica ROUGE-1 no conjunto de textos DUC 2002.
Os resultados alcançados para a métrica ROUGE-1 no conjunto de dados DUC 2002 são de uma forma generalizada melhores do que os obtidos no conjunto de 2001, e confirmam as tendências definidas no teste com conjunto 2001. A combinação de pesos **"Híbrido(20-20-20-40)"** continua a ser a abordagem com os melhores resultados, e o método de sumarização por cálculo da relevância de termos, a abordagem com piores resultados, confirmando-se no entanto a sua propensão a valores elevados de precisão.

Tabela 4.3: Resultados de Cobertura, Precisão e Medida-F para ROUGE-1 em textos DUC 2002.

ROUGE-1			
DUC 2002	**Cobertura**	**Precisão**	**Medida-F**
Relevância	0,372737	0,450028	0,402682
TextRank	0,416197	0,438561	0,425188
Híbrido(25-25-25-25)	0,424881	0,450956	0,435689
Híbrido(20-20-40-20)	0,418122	0,441895	0,427687
Híbrido(20-20-20-40)	**0,433934**	**0,471781**	**0,449823**
Híbrido(10-10-40-40)	0,432316	0,455892	0,441863

Na Tabela 4.4, são apresentados os resultados dos métodos implementados para a métrica ROUGE-2 no conjunto de textos DUC 2001.

Tabela 4.4: Resultados de Cobertura, Precisão e Medida-F para ROUGE-2 em textos DUC 2001.

ROUGE-2			
DUC 2001	**Cobertura**	**Precisão**	**Medida-F**
Relevância	0,121312775	0,145237107	0,130428485
TextRank	0,138153537	0,137586946	0,137227318
Híbrido(25-25-25-25)	0,147522671	0,15351074	0,14939457
Híbrido(20-20-40-20)	0,130639117	0,138413293	0,13319605
Híbrido(20-20-20-40)	**0,16791302**	**0,17939973**	**0,172227094**
Híbrido(10-10-40-40)	0,155625234	0,165819492	0,159578021

O uso da métrica ROUGE-2 no conjunto DUC 2001 obtém resultados proporcionais aos obtidos pela métrica ROUGE-1 no mesmo conjunto, no entanto é possível observar uma maior disparidade entre os resultados obtidos pelos melhores e os piores métodos, enquanto a proporcionalidade do desempenho das diferentes abordagens é mantida, a distância entre os resultados atingidos pelos métodos é mais evidente. De resto, as tendências de desempenho referidas anteriormente são igualmente válidas nesta amostra.
Na Tabela 4.5, são apresentados os resultados dos métodos implementados para a métrica ROUGE-2 no conjunto de textos DUC 2002.
À semelhança do registado para o conjunto DUC 2001, os resultados dados pela métrica ROUGE-2 para o conjunto DUC 2002 corroboram as mesmas tendências de desempenho das diferentes abordagens. De notar que para este conjunto o método baseado na relevância de termos tem um dos piores resultados em termos de precisão.
Os resultados obtidos pelo método *TextRank*, embora estando contidos na gama de resultados obtidos pelos autores do método [MT04], são de uma forma geral inferiores aos seus melhores resultados. Este facto poderá dever-se a um conjunto de factores como a dimensão do con-

junto de testes, diferenças na implementação e configuração do método e da ferramenta de avaliação.

Tabela 4.5: Resultados de Cobertura, Precisão e Medida-F para ROUGE-2 em textos DUC 2002.

	ROUGE-2		
DUC 2002	**Cobertura**	**Precisão**	**Medida-F**
Relevância	0,12985	0,153458	0,13899
TextRank	0,149319	0,157314	0,152503
Híbrido(25-25-25-25)	0,155823	0,164619	0,159482
Híbrido(20-20-40-20)	0,146092	0,153378	0,149025
Híbrido(20-20-20-40)	**0,168254**	**0,183727**	**0,174614**
Híbrido(10-10-40-40)	0,15972	0,169365	0,163671

Observa-se de forma concistente um melhor desempenho por parte da abordagem híbrida, em especial para uma combinação de pesos em que é atríbuida igual importância a cada uma das componentes, e para a combinação de pesos que favorece a classificação das frases obtida através da heurística posicional. Este comportamento era espectável, dado o género dos textos do conjunto de testes, no entanto, para uma utilização generalista, esta característica poderá não obter o mesmo tipo de resultados.
Na Tabela 4.6, são apresentados os tamanhos médios dos sumários gerados pelos diferentes métodos implementados, medido em número de palavras.

Tabela 4.6: Tamanhos médios dos sumários gerados pelo sistema.

Tamanho Médio do Sumário	**DUC 2001**	**DUC 2002**
Relevância	93,91	95,66
TextRank	115,25	110,3
Híbrido(25-25-25-25)	111,47	109,03
Híbrido(20-20-40-20)	107,76	109,17
Híbrido(20-20-20-40)	108,13	106,4
Híbrido(10-10-40-40)	106,97	109,8
Média Geral	**107,25**	**106,73**

De uma forma geral os sumários produzidos pelo sistema desenvolvido, tendo em conta o tamanho limite imposto, têm um tamanho inferior ao tamanho dos sumários de referência, sendo que, o método de relevância de termos é a abordagem que produz os sumários de tamanho mais reduzido.

Capítulo 5

Conclusões e Perspectivas Futuras

A motivação para o desenvolvimento desta dissertação prende-se com a procura de soluções que permitam resolver ou pelo menos atenuar o cada vez mais significativo problema de "sobrecarga de informação". Existem inúmeras abordagens que pretendem resolver este problema, quer sejam baseadas em técnicas superficiais ou em conhecimento profundo, produzindo sumários extractivos ou abstractivos, no entanto, não existe ainda uma solução definitiva. A complexidade inerentemente humana do processo manual de sumarização tem impossibilitado a construção de sistemas computacionais, com a tecnologia actualmente disponível, que permitam a sua modelação adequada. Assim, sistemas mais ricos em conhecimento profundo tendem a ser complexos, custosos e muito limitados em termos de género e domínio dos textos que conseguem analisar com produção de resultados satisfatórios. Por outro lado, sistemas baseados apenas em características estruturais e superficiais do texto, enquanto mais robustos e usufruindo da capacidade de um uso mais geral, produzem ainda sumários com uma textualidade rudimentar.

O objectivo fundamental deste trabalho foi o desenvolvimento de um sistema de sumarização automática de texto, com base na investigação e implementação de diversos métodos de sumarização extractiva. Os métodos explorados incluem um método estatístico de cálculo da relevância de termos, com base nas suas frequências, um método baseado em teoria de grafos e um método híbrido que combina as duas abordagens anteriores com a exploração de características como a presença de palavras-chave e a posição das frases no texto.

O trabalho desenvolvido contemplou ainda a implementação de uma interface gráfica (Secção 3.5) que possibilita a interacção com o sistema, permitindo realizar a sumarização de textos, obtidos a partir de diversos formatos de entrada, através dos métodos de sumarização implementados. Permitindo ainda a comparação visual entre o texto fonte e o sumário, em que elementos extraídos do texto fonte podem ser assinalados com marcação colorida.

Os resultados da avaliação do desempenho do sistema, levada a cabo através do uso da ferramenta ROUGE, permitiram tecer considerações sobre a qualidade dos sumários produzidos quanto ao seu conteúdo informativo, no entanto, este tipo de avaliação não permite ajuizar sobre a coesão e coerência textual dos sumários, não permitindo por isso fazer, desta forma, qualquer apreciação sobre a legibilidade e inteligibilidade dos mesmos. Esta forma de avaliação intrínseca e automática permitiu, no entanto, uma comparação objectiva dos níveis de desempenho dos diferentes métodos, em que se verificou ser o método estatístico o que atinge piores resultados, estando numa posição intermédia o método baseado em teoria de grafos, tendo sido o método híbrido aquele que de forma consistente obteve os melhores resultados nos vários testes realizados. Esta vantagem generalizada da abordagem híbrida deixa a ideia que um sistema de sumarização automática de texto deverá explorar a combinação de diferentes características complementares no sentido de conseguir um melhor reconhecimento do conteúdo relevante de um texto fonte. Ainda assim, o reconhecimento do conteúdo relevante de um texto é apenas parte do problema, para que a produção de resumos que se aproximem daqueles produzidos por humanos possa ser uma realidade, o problema da representação do conteúdo identificado na geração de um resumo em linguagem natural deve ser resolvido, no

entanto, esta é uma tarefa muito complexa e que está longe de ser exequível com a tecnologia de que actualmente dispomos numa perspectiva de sumarização generalizada, liberta de domínio ou género do texto fonte.

A sumarização extractiva aqui realizada representa assim uma abordagem que permite oferecer níveis de desempenho razoáveis e aceitáveis em aplicações em que a principal prioridade do sistema seja a apresentação do principal conteúdo informativo do texto fonte, ainda que fazendo uso de uma textualidade de qualidade limitada.

Numa perspectiva de trabalho futuro, seria interessante explorar a vertente de sumarização multidocumento, e em especial, aplicada a documentos de aglomerados de notícias como os usados no âmbito deste trabalho, isto permitiria produzir um resumo para cada aglomerado do documento, em vez de para cada notícia individualmente, o que seria especialmente interessante se cada aglomerado do documento tratasse de um tópico diferente, outra abordagem seria produzir o resumo de todas as notícias de todos aglomerados de um documento ou mesmo de vários destes documentos. Esta abordagem seria comparativamente mais complexa, uma vez que, o resumo produzido teria de reflectir a possível diversidade de tópicos dos aglomerados ou mesmo documentos completos, tentando reduzir simultaneamente, tanto quanto possível, a redundância da informação incluída no sumário. Outra direcção a seguir seria a sumarização independente de idioma, procurando soluções alternativas aos mecanismos aqui usados para tarefas de remoção de palavras irrelevantes e filtro sintáctico. Também vantajosa seria a investigação de técnicas que tivessem em conta uma maior abrangência em termos de conhecimento linguísticos, permitindo explorar características que possam ir além de características estruturais e de repetição ou co-ocorrência de conteúdo textual, tentando explorar características da semântica textual. Por fim, outra importante direcção a seguir seria a distanciação do modelo de extracção pura para adopção de um modelo de edição e compressão das frases extraídas.

Bibliografia

[AAHM11] Rasim M. Alguliev, Ramiz M. Aliguliyev, Makrufa S. Hajirahimova, and Chingiz A. Mehdiyev. Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Syst. Appl.*, 38(12):14514–14522, 2011. Available from: http://dblp.uni-trier.de/db/journals/eswa/eswa38.html#AlgulievAHM11. 19

[ABD+95] John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. MITRE: description of the alembic system used for MUC-6. In *Proc. of the 6th Conf. on Message Understanding (MUC-6)*, pages 141–155, 1995. 33

[ACA10] Christopher Arnold, Tony Cook, and Elizabeth Angeli. The inverted pyramid structure, april 2010. Available from: http://owl.english.purdue.edu/owl/resource/735/04/. 32

[ACC+03] Laura Alonso, Irene Castellón, Salvador Climent, Maria Fuentes, Lluís Padró, and Horacio Rodríguez. Approaches to text summarization: Questions and answers. *Revista Iberoamericana de Inteligencia Artificial*, 22:79–102, 2003. 15

[ACL12] ACL. Association for computational linguistics workshops and conferences [online]. 2012. Available from: http://www.acl2012.org/. 3

[AHS08] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In Il-Yeol Song, Johann Eder, and Tho Nguyen, editors, *Data Warehousing and Knowledge Discovery*, volume 5182 of *Lecture Notes in Computer Science*, pages 305–316. Springer Berlin, Heidelberg, 2008. 31

[ANS97] ANSI. Guidelines for abstracts. Niso standard (ansi), NISO Press, Bethesda, Maryland, 1997. 6

[AOGL99] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. A trainable summarizer with knowledge acquired from robust nlp techniques. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 71–80+. MIT Press, 1999. 16

[Bax58] P. B. Baxendale. Machine-made index for technical literature: an experiment. *IBM J. Res. Dev.*, 2(4):354–361, October 1958. Available from: http://dx.doi.org/10.1147/rd.24.0354. 2, 7, 15

[BB75] H. Borko and C.L. Bernier. *Abstracting concepts and methods*. Library and information science. Academic Press, 1975. Available from: http://books.google.pt/books?id=6qObzRav_AkC. 11

[Ber05] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1), 2005. Available from: http://www.internetmathematics.org/volumes/2/1/Berkhin.pdf. 29

[BGMP01] Orkut Buyukkokten, Hector Garcia-Molina, and Andreas Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 652–662, New York, NY, USA, 2001. ACM. Available from: http://doi.acm.org/10.1145/371920.372178. 10

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998. Available from: http://dx.doi.org/10.1016/S0169-7552(98)00110-X. 18, 28, 29

[BSR+05] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM. Available from: http://doi.acm.org/10.1145/1102351.1102363. 17

[CDB07] João Paulo Cordeiro, Gaël Dias, and Pavel Brazdil. Learning paraphrases from wns corpora. In Geoff Sutcliffe David Wilson, editor, *Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2007)*, pages 193–198, key West, Florida, USA, May 2007. AAAI Press. DBLP. Available from: http://10.255.0.115/pub/2007/CDB07. 24

[CG98] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM. Available from: http://doi.acm.org/10.1145/290941.291025. 19

[CLW07] Yanmin Chen, Bingquan Liu, and Xiaolong Wang. Automatic text summarization based on textual cohesion. *Journal of Electronics (China)*, 24:338–346, 2007. 10.1007/s11767-005-0188-5. Available from: http://dx.doi.org/10.1007/s11767-005-0188-5. 6

[CN95] H. Chen and T. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist hopfield net activation. *Journal of the American Society for Information Science*, 46:348–369, 1995. 18

[CNP06] Giuseppe Carenini, Raymond Ng, and Adam Pauls. Multi-document summarization of evaluative text. In *In Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 305–312, 2006. 22

[CO01] John M. Conroy and Dianne P. O'leary. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 406–407, New York, NY, USA, 2001. ACM. Available from: http://doi.acm.org/10.1145/383952.384042. 17

[Cre82] E.T. Cremmins. *The Art of Abstracting*. Professional Writing Series. ISI Press, 1982. Available from: http://books.google.pt/books?id=2pBpAAAAMAAJ. 6, 11

[DDF+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990. 19, 20

[DM07] Dipanjan Das and André F.T. Martins. A survey on automatic text summarization. Technical report, Literature Survey for the Language and Statistics II course at CMU, November 2007. 1, 3, 16, 22

[DUC07] DUC. The document understanding conference [online]. 2001-2007. Available from: http://duc.nist.gov. 3, 17, 41

[Edm69] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, April 1969. Available from: http://doi.acm.org/10.1145/321510.321519. 3, 7, 16, 31

[EN98] Brigitte Endres-Niggemeyer. *Summarizing Information*. ISBN: 978-3-540-63735-6. Springer-Verlag, 1998. 6

[ENH00] Brigitte Endres-Niggemeyer and Fachhochschule Hannover. Human-style www summarization, 2000. 22

[ENMS95] Brigitte Endres-Niggemeyer, Elisabeth Maier, and Alexander Sigel. How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management*, 31(5):631 – 674, 1995. Available from: http://www.sciencedirect.com/science/article/pii/030645739500028F. 6

[ER04a] Güneş Erkan and Dragomir R. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, Barcelona, Spain, 2004. 17

[ER04b] Güneş Erkan and Dragomir R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December 2004. Available from: http://dl.acm.org/citation.cfm?id=1622487.1622501. 17

[FC99] T. Firmin and M.J. Chrzanowski. An Evaluation of Automatic Text Summarization Systems. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 325–336. The MIT Press, 1999. 40

[FGL+08] Jianping Fan, Yuli Gao, Hangzai Luo, Daniel A. Keim, and Zongmin Li. A novel approach to enable semantic and visual image summarization for exploratory image search. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 358–365, New York, NY, USA, 2008. ACM. Available from: http://doi.acm.org/10.1145/1460096.1460155. 10

[fHLTB] Center for Human Language Technology and Bioinformatics. The hultiglib. Available from: http://www.di.ubi.pt/~jpaulo/hultiglib/. 25, 26

[Fid86] Raya Fidel. Writing abstracts for free-text searching. *Journal of Documentation*, 42(1):11–21, 1986. 6

[For08] Maria Fuentes Fort. *A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language*. PhD thesis, Universitat Politúcnica de Catalunya, March 2008. 11, 20

[Foua] The Apache Software Foundation. Apache opennlp - a machine learning based toolkit for the processing of natural language text. Available from: http://opennlp.apache.org/index.html. 30

[Foub] The Apache Software Foundation. Apache tika - a content analysis toolkit. Available from: http://tika.apache.org/. 25

[FRB08] Cláudia Freitas, Paulo Rocha, and Eckhard Bick. Floresta sintá(c)tica: Bigger, thicker and easier. In António J. S. Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, and Paulo Quaresma, editors, *PROPOR*, volume 5190 of *Lecture Notes in Computer Science*, pages 216–219. Springer, 2008. Available from: `http://dblp.uni-trier.de/db/conf/propor/propor2008.html#FreitasRB08`. 30

[GDCY02] K. Ganapathiraju, Advisors Dr, Jaime Carbonell, and Dr Yiming Yang. Relevance of cluster size in mmr based summarizer: A report 11-742: Self-paced lab in information retrieval, 2002. 19, 20

[GL01] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 19–25, New York, NY, USA, 2001. ACM. Available from: `http://doi.acm.org/10.1145/383952.383955`. 20

[GL10] Vishal Gupta and Gurpreet Lehal. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268, August 2010. Available from: `http://ojs.academypublisher.com/index.php/jetwi/article/view/0203258268`. 22

[GSG09] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, and Bahareh Gholamzadeh. A comprehensive survey on text summarization systems. In *Computer Science and its Applications, 2009. CSA '09. 2nd International Conference on*, pages 1–6, dec. 2009. xi, 12

[GTGJ01] Pedro Gomes, Sérgio Tostão, Daniel Gonçalves, and Joaquim Jorge. Web clipping: Compression heuristics for displaying text on a pda. *Mobile HCI '01*, 2001. 10

[HL99] E. Hovy and C. Y. Lin. Automated Text Summarization in SUMMARIST. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, 1999. 12

[HL02] Sanda M. Harabagiu and Finley Lăcătuşu. Generating single and multi-document summaries with gistexter. In *In U. Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization*, pages 30–38, 2002. 15

[HM00] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. *Computer*, 33(11):29–36, November 2000. Available from: `http://dx.doi.org/10.1109/2.881692`. 10, 25

[Hov05] E.H Hovy. *Automated Text Summarization*, pages 583–598. The Oxford Handbook of Computational Linguistics. Oxford University Press, 2005. 7, 11

[HSGG99] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 489–498, New York, NY, USA, 1999. ACM. Available from: `http://doi.acm.org/10.1145/319463.319691`. 10

[Hut87] John Hutchins. Summarization: Some problems and methods. In Karen Spärck Jones, editor, *Meaning: The frontier of informatics*. ASLIB, London, 1987. 12

[HWC95] Alexander G. Hauptmann, Michael J. Witbrock, and Michael G. Christel. News-on-demand: An application of informedia technology. Technical report, School of Computer Science, Carnegie Mellon University, 1995. 11

[JBME98] Hongyan Jing, Regina Barzilay, Kathleen Mckeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *In AAAI Symposium on Intelligent Summarization*, pages 60–68, 1998. 40

[Jin01] Hongyan Jing. *Cut-and-Paste Text Summarization*. PhD thesis, Columbia University, 2001. 6, 10, 15

[JM99] Hongyan Jing and Kathleen R. McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 129–136, New York, NY, USA, 1999. ACM. Available from: http://doi.acm.org/10.1145/312624.312666. 14

[Jon98] Karen Spärck Jones. Automatic summarising: Factors and directions. In *Advances in Automatic Text Summarization*, pages 1–12. MIT Press, 1998. 7, 8, 11, 12, 22

[JS08] Karel Ježek and Josef Steinberger. Automatic text summarization (the state of the art 2007 and new challenges). In *In Proceedings of Znalosti 2008, Bratislava, Slovakia*, pages 1–12, February 2008. 1, 11, 19, 20, 22

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999. Available from: http://doi.acm.org/10.1145/324133.324140. 18

[Koh] Christian Kohlschütter. Boilerpipe - boilerplate removal and fulltext extraction from html pages. Available from: http://code.google.com/p/boilerpipe/. 25

[KPC95] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 68–73, New York, NY, USA, 1995. ACM. Available from: http://doi.acm.org/10.1145/215206.215333. 3, 16

[Kru93] G.R. Krupka. SRA: description of the SRA system as used for MUC-6. In *Proc. of the 6th Conf. on Message Understanding*, pages 221–235, 1993. 33

[KV78] Walter Kintsch and T. A. Van Dijk. Toward a model of text comprehension and production. *Psychological review*, 85(5):363, 1978. Available from: http://psycnet.apa.org/journals/rev/85/5/363/. 6

[LD97] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. Available from: /brokenurl#http://publication.wilsonwong.me/load.php?id=233281598. 20

[LFZ11] Hangzai Luo, Jianping Fan, and Youjie Zhou. Multimedia news exploration and retrieval by integrating keywords, relations and visual features. *Multimedia Tools Appl.*, 51(2):625–648, January 2011. Available from: http://dx.doi.org/10.1007/s11042-010-0639-3. 11

[LH03] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/1073445.1073465. 41

[Lin04a] Chin-Yew Lin. Looking for a Few Good Metrics: ROUGE and its Evaluation. In *Working Notes of NTCIR-4*, pages 1–8, June 2004. 41

[Lin04b] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004. Available from: http://research.microsoft.com/~cyl/download/papers/WAS2004.pdf. 41, 42, 43

[Llo08] Elena Lloret. Text summarization: An overview. Technical report, University of Alicante, Alicante, Spain, 2008. 7, 10, 15

[LP12] Elena Lloret and Manuel Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37:1–41, 2012. 10.1007/s10462-011-9216-z. Available from: http://dx.doi.org/10.1007/s10462-011-9216-z. 15

[LPE97] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Communications of the ACM*, 40:55–62, 1997. 10

[Luh58] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958. Available from: http://dx.doi.org/10.1147/rd.22.0159. 2, 7, 15

[Man01a] I Mani. Summarization evaluation: An overview. *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, 2001. 39

[Man01b] Inderjeet Mani. *Automatic Summarization*. ISBN: 978-90-272-4985-2. John Benjamins Publishing Company, 2001. 5, 7, 10, 22, 41

[MB97] Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence*, AAAI'97/IAAI'97, pages 622–628. AAAI Press, 1997. Available from: http://dl.acm.org/citation.cfm?id=1867406.1867503. 18

[MB99] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. *Inf. Retr.*, 1(1-2):35–67, May 1999. Available from: http://dx.doi.org/10.1023/A:1009930203452. 18

[MBG98] Inderjeet Mani, Eric Bloedorn, and Barbara Gates. Using Cohesion and Coherence Models for Text Summarization. In Eduard Hovy, editor, *Intelligent Text Summarization*. AAAI Press, 1998. 18, 32, 33

[Mer97] A. E. Merlino. Multimedia summaries of broadcast news. In *Proceedings of the 1997 IASTED International Conference on Intelligent Information Systems (IIS '97)*, IIS '97, pages 442–, Washington, DC, USA, 1997. IEEE Computer Society. Available from: http://dl.acm.org/citation.cfm?id=846229.848721. 11

[MHK+99] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. The tipster summac text summarization evaluation. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 77–85, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/977035.977047. 41

[Mih05] R. Mihalcea. Language independent extractive summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 49–52, Ann Arbor-MI, United States, June 2005. Association for Computational Linguistics. 18, 28

[Mil95] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. Available from: http://doi.acm.org/10.1145/219717.219748. 17, 33

[Mly06] Angela Mlynarski. Automatic text summarization in digital libraries. Master's thesis, University of Lethbridge. Faculty of Arts and Science, 2006. 11

[MM99] I. Mani and M.T. Maybury. *Advances in Automatic Text Summarization*. Mit Press, 1999. Available from: http://books.google.pt/books?id=YtUZQaKDmzEC. 7, 8, 10

[MMM97] A. Merlino, D. Morey, and M. Maybury. Broadcast news navigation using story segmentation. In *Proc. of the 5th ACM Int'l. Conf. on Multimedia*, pages 381–391, 1997. 10

[MPER01] C.B. Martins, T. A. S. Pardo, A. Espina, and L. H. M. Rino. Introdução à sumarização automática. Technical report, Departamento de Computação da UFSCar RT-DC 002/2001, 2001. 5, 6, 7, 8

[MSM94] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994. 30

[MT04] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004. 18, 28, 29, 30, 31, 45

[MT05] Rada Mihalcea and Paul Tarau. A language independent algorithm for single and multiple document summarization. In *In Proceedings of IJCNLP'2005*, 2005. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.136.1125. 18, 28

[Nen05] Ani Nenkova. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, AAAI'05, pages 1436–1441. AAAI Press, 2005. Available from: http://dl.acm.org/citation.cfm?id=1619499.1619564. 17, 32

[Nen06] Ani Nenkova. Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH*, 2006. 39

[NM03] Tadashi Nomoto and Yuji Matsumoto. The diversity-based approach to open-domain text summarization. *Inf. Process. Manage.*, 39(3):363–389, May 2003. Available from: http://dx.doi.org/10.1016/S0306-4573(02)00096-1. 19

[OKG07] Shiyan Ou, Christopher S. G. Khoo, and Dion H. Goh. Automatic multidocument summarization of research abstracts: Design and user evaluation. *J. Am. Soc. Inf. Sci. Technol.*, 58(10):1419–1435, August 2007. Available from: http://dx.doi.org/10.1002/asi.v58:10. 11

[Pai90] C. D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Inf. Process. Manage.*, 26(1):171–186, April 1990. Available from: http://dx.doi.org/10.1016/0306-4573(90)90014-S. 22

[Pat07] Kaustubh Patil. Automatic text summarization using pathfinder network scaling. Master's thesis, Faculty of Engineering, University of Porto, 2007. supervisor: Pavel Brazdil. Available from: http://10.255.0.115/pub/2007/Pat07. 19, 20

[PB07] Kaustubh Patil and Pavel Brazdil. Sumgraph: Text summarization using centrality in the pathfinder network. *International Journal on Computer Science and Information Systems*, 2(1):18–32, 2007. Available from: http://10.255.0.115/pub/2007/PB07a. 32

[Por80] Martin F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980. 26

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/1073083.1073135. 41

[RHM02] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Comput. Linguist.*, 28(4):399–408, December 2002. Available from: http://dx.doi.org/10.1162/089120102762671927. 1, 11, 39

[RP03] L. H. M. Rino and T. A. S. Pardo. A sumarização automática de textos: Principais características e metodologias. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação*, volume 8 of *III Jornada de Minicursos de Inteligência Artificial (III MCIA)*, pages 203–245, Campinas, SP, Brasil, 2003. 5, 7

[SH02] Mihai Surdeanu and Sanda M. Harabagiu. Infrastructure for open-domain information extraction. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 325–330, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. Available from: http://dl.acm.org/citation.cfm?id=1289189.1289220. 15

[SJ93] Karen Spärck Jones. What might be in a Summary? In Knorz, Krause, and Womser Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Konstanz, DE, 1993. Universitätsverlag Konstanz. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4097. 8

[SJ01] Karen Spärck Jones. Factorial Summary Evaluation. In *Proceedings of the Workshop on Text Summarization (DUC 2001)*, New Orleans, Louisiana, USA, September 2001. Available from: http://www-nlpir.nist.gov/projects/duc/duc2001/agenda_duc2001.html. 11

[SJ07] Karen Spärck Jones. Automatic summarising: The state of the art. *Inf. Process. Manage.*, 43(6):1449–1481, November 2007. Available from: http://dx.doi.org/10.1016/j.ipm.2007.03.009. 3, 7, 11, 15, 22

[SJG96] Karen Spärck Jones and Julia Rose Galliers, editors. *Evaluating Natural Language Processing Systems, An Analysis and Review*, volume 1083 of *Lecture Notes in Computer Science*. Springer, 1996. 39

[SL02] Horacio Saggion and Guy Lapalme. Generating indicative-informative summaries with sumum. *Comput. Linguist.*, 28(4):497–526, December 2002. Available from: http://dx.doi.org/10.1162/089120102762671963. 14

[SR05] Eloize Rossi Marques Seno and Lucia Helena Machado Rino. Rhesumarst: Um sumarizador automático de estruturas rst. Master's thesis, Departamento de Computação, Universidade Federal de São Carlos, 2005. 6

[SSMB97] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193–207, March 1997. Available from: http://dx.doi.org/10.1016/S0306-4573(96)00062-3. 17

[SSZ+05] Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. Web-page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 194–201, New York, NY, USA, 2005. ACM. Available from: http://doi.acm.org/10.1145/1076034.1076070. 11

[SUN11] S. SUNEETHA. Automatic text summarization: The current state of the art. Technical report, Department of Computer Science and Engineering, JNTU, Hyderabad, 2011. 1, 22

[SVB07] Krysta Marie Svore, Lucy Vanderwende, and Christopher J. C. Burges. Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL*, pages 448–457. ACL, 2007. Available from: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2007.html#SvoreVB07. 17

[TAC12] TAC. The text analysis conference [online]. 2008-2012. Available from: http://www.nist.gov/tac/. 3, 41

[Thu24] S. Thurber. *Précis writing for American schools: methods of abridging, summarizing, condensing, with copious exercises*. The Atlantic monthly press, 1924. Available from: http://books.google.pt/books?id=TBFKAAAAIAAJ. 6

[Tuc99] Richard Tucker. *Automatic summarising and the CLASP system*. PhD thesis, University of Cambridge Computer Laboratory, 1999. 22

[Wik01] Wikipedia. Wikipedia, the free encyclopedia, 2001. Available from: http://www.wikipedia.org/. 17

[YCKQ07] Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. Document concept lattice for text understanding and summarization. *Inf. Process. Manage.*, 43(6):1643–1662, November 2007. Available from: http://dx.doi.org/10.1016/j.ipm.2007.03.010. 14

[YW03] Christopher C. Yang and Fu Lee Wang. Fractal summarization for mobile devices to access large documents on the web. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 215–224, New York, NY, USA, 2003. ACM. Available from: http://doi.acm.org/10.1145/775152.775183. 10

[Zec97] Klaus Zechner. A literature survey on information extraction and text summarization. *Computational Linguistics Program*, 1997. 22

[ZW00] Klaus Zechner and Alex Waibel. Diasumm: flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 968–974, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/992730.992786. 10

Glossário

Anáfora	Em Linguística é o processo segundo o qual um termo remete para outro termo anteriormente inserido no mesmo texto.
Corpus	Consiste numa compilação de documentos ou informações relativos a uma disciplina, tema ou domínio. Em Linguística denota um conjunto finito de enunciados representativos de uma determinada estrutura.
Hiperonímia	Relação semântica entre palavras em que o significado de uma, mais genérica (hiperónimo), inclui o significado de outra, mais específica (hipónimo).
Hiponímia	Relação semântica em que uma palavra está num plano hierárquico inferior, uma vez que pertence a uma classe ou espécie que a inclui ao nível do significado. Este facto implica que o significado do hipónimo (etimologicamente significa nome pequeno) é mais específico e mais restrito do que o significado do hiperónimo a que pertence.
Metonímia	Recurso expressivo que consiste no emprego de uma palavra em vez de outra devido a uma relação lógica ou de contiguidade existente entre elas, que se exprime nas relações da causa pelo efeito, do todo pela parte, do continente pelo conteúdo, etc., e vice-versa.
N-grama	No campo da Linguística Computacional representa uma sequência contígua de n itens de uma dada sequência de texto. Este elemento poderá ter níveis de granularidade variados, podendo representar sequências de sílabas, letras, fonemas ou palavras.
Polissemia	Qualidade de uma palavra que possui vários significados contidos numa mesma forma gráfica e fonológica.
Sinonímia	Relação de proximidade semântica entre duas ou mais palavras, que podem, por isso, ser usadas no mesmo contexto sem que haja alteração de significado do enunciado em que ocorrem.
Textualidade	Conjunto de características que permitem perceber um grupo de frases como um texto, isto é, como um todo com sentido.