UNIVERSIDADE DA BEIRA INTERIOR
Engenharia
Informática

Sumarização Personalizada e Subjectiva de Texto

Bruno Miguel Fernandes

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

Orientador: Prof. Doutor João Paulo da Costa Cordeiro

Covilhã, Outubro de 2014

Agradecimentos

Os meus primeiros agradecimentos são naturalmente para a minha família, por toda a ajuda não só económica mas também por todo o apoio demonstrado nos momentos mais difíceis da minha vida, em especial aos meus pais, pois é graças a eles que hoje estou aqui a terminar esta etapa da minha vida.
Quero agradecer ao Professor Doutor João Paulo da Costa Cordeiro, o qual manifestou sempre uma grande disponibilidade, durante a orientação desta dissertação e pelos ensinamento transmitidos, apoio e dedicação manifestados.
Finalmente não podia, também, deixar de agradecer a todos os meus colegas, que durante este percurso no ensino superior me ajudaram não só a nível académico mas também me proporcionaram bons momentos.

Resumo

Um texto pode ser sumarizado ou resumido, isto é, o seu assunto ou conceito pode ser representado de uma forma mais sucinta. A representação mais comum de um sumário é a escrita, pois é constantemente produzida pelas pessoas, quando estas querem descrever uma determinado assunto.
Ao longo dos últimos anos o uso da `Internet` tem vindo a massificar-se e com isso a quantidade de informação disponível nesta enorme rede, aumentou exponencialmente, sendo este acontecimento denominado como sobrecarga de informação. Isto levanta uma série de problemas, entre eles a procura de informação relevante, sobre um determinado tema. Quando alguém procura essa informação pretende encontrá-la de forma eficiente, ou seja, rápido e que aborde diretamente o assunto pretendido. Quanto ao assunto, existem algumas formas de procurar o mesmo, já em relação à celeridade da pesquisa, deparamo-nos com uma quantidade enorme de informação que por vezes difere daquilo que procuramos, sendo muito demoroso o processo de leitura de toda essa informação.
Uma das formas de resolver esse problema é resumir o conteúdo do texto encontrado, para que assim possamos de uma forma mais rápida ter uma noção sobre o tema do texto encontrado. Na área da sumarização existem várias técnicas que possibilitam a obtenção de um sumário mais específico.
Esta dissertação tem como base a combinação de algumas das técnicas estudadas ao longo do tempo, tais como, relevância e informatividade das palavras, objetividade, segmentação em tópicos e no uso de palavras que representem o domínio do texto.
Numa abordagem estatística destacam-se a relevância dos termos de um texto, que é calculada através da frequência dos termos presentes nesse texto e num `corpus`,a extração das palavras-chave que serão encontradas através da sua relevância no texto e a posição das frases no documento que consoante o seu tipo, pode ser calculado de diversas formas, neste caso, sendo avaliado com textos noticioso, foi implementada uma heurística posicional que atribui mais relevância a frases cimeiras. A abordagem baseada na subjectividade de um texto é implementada recorrendo a um conjunto de dados textuais conhecido como *SentiWordNet* [BES10]. Foi ainda implementada uma abordagem híbrida em que se combinam total ou parcialmente os métodos referidos anteriormente.
De modo a proceder à avaliação do sistema foram utilizados dois conjuntos de dados noticiosos. Um destes conjuntos de dados é proveniente da *Document Understanding Conference*, datado de 2001, o outro é o *corpus TeMário*. Para que os sumários produzidos pudessem ser avaliados automaticamente, foi utilizada uma implementação em linguagem *JAVA* da ferramenta *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*). Após a comparação dos resultados do método híbrido com os restantes, com e sem identificação dos tópicos ficou evidenciado que a heurística posicional das frases obtém melhores resultados, pelo que os métodos híbridos onde esta característica tem peso superior às restantes, tanto para quando o texto é separado em tópicos como no caso contrário, de uma forma geral, obtém melhores resultados. O melhor desempenho no total dos resultados é obtido com o método híbrido, atribuindo maior peso à componente da heurística posicional da frase, sem identificação dos tópicos.

Palavras-chave

DUC, Morphadorner, Tópicos, Relevância dos Termos, Subjectividade, Hultiglib, Sumarização Automática de Texto, Sumarização Automática Extractiva, Estatística Textuais, Métodos Híbridos, ROUGE.

Abstract

A text can be summarized or abstracted, ie, its subject or concept can be represented in a more succinct form. The most common representation of a summary is written, because it is constantly produced by people when they want to describe a particular subject.
Over the last years, the use of Internet has come to popularize and therewith the amount of information available in this huge network, has increased exponentially, and this event is called as "information overload". This raises a set of problems, among them the search for relevant information on a given theme. When someone searches for this information he/she want to find it efficiently, ie, fast and directly address the intended subject. For the theme, there are some ways to find it, as compared to the speed of research, we are faced with an enormous amount of information which sometimes differs from what we search, being very slow the process of reading all this information.
One way to solve this problem is to summarize the contents of the text found, so we can a faster way to get a sense on the subject of the text found. In the area of summarization, various techniques exist which allow to obtain a more specific shape.
This dissertation is based on the combination of some techniques, studied over time, such as relevance and informativeness of the words, objectivity, segmentation in topics and in the use a set of words that represent the domain of the text.
In statistical approach is highlighted the relevance of the terms of a text, which is calculated from the frequency of terms present in a text and corpus, the extraction of domain words that will be encountered by their relevance in the text and the position of the phrases in the document, that depending on type, can be calculated in different ways, in this case, being evaluated with news texts, was implemented a positional heuristic that assigns more importance to sentences in the text top. The approach based in subjectivity of a text is implemented using a set of textual data known as SentiWordNet [BES10]. It was also implemented a hybrid approach that combines all or a set of the methods mentioned above.
In order to realize an evaluatiuon of the system, two sets of news data was used. One of these data are from the *Document Understanding Conference*, dated 2001 and other is *TeMário corpus*. For summaries produced could be evaluated automatically, was used an implementation in JAVA language, of tool ROUGE (*Recall-Oriented Evaluation Understudy for Gisting*). After comparing the results of the hybrid method with the other, with and without identification of topics, was showed that the positional heuristic of sentences obtained better results, so that the hybrid methods where this feature has top weight to the others, both when the text is separated into topics or not, in general, performs better. The best performance in overall results are obtained with the hybrid method, assigning greater weight to the positional heuristic phrase, without identification of the component threads.

Keywords

DUC, Morphadorner, Topics, Term Relevance, Subjectivity, Hultiglib, Automatic Text SUmmarization, Extractive Automatic Summarization, Textua Statistics, Hybrid Methods, ROUGE.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

BLEU	Bilingual Evaluation Understudy
DUC	Document Understanding Conferences
EI	Extracção de Informação
IBM	International Business Machines
MMR	Maximal Marginal Relevance
PLN	Processamento de Linguagem Natural
ROUGE	Recall-Oriented Understudy of Gisting Evaluation
SVD	Singular Value Decomposition
TAC	Text Analysis Conference
TF-IDF	Text Frequency - Inverse Document Frequency

Capítulo 1

Introdução

Actualmente existem várias fontes de informação disponíveis, não só em forma digital mas também em forma analógica. Com o avanço da tecnologia é possível quebrar algumas barreiras na consulta da informação pois o acesso à mesma pode ser efectuado de forma mais rápida, com mais precisão e comodidade. A *Internet* veio ajudar nessa consulta e obtenção de informação, sendo utilizada por muitos para distribuir em larga escala informação de qualquer assunto. Por outro lado, existem também pessoas e instituições que necessitam dessa informação e a consultam nos mais variados meios, incluindo a *Internet*. Quando a informação é pouca, a pesquisa é facilitada, mas com os grandes volumes de dados existentes actualmente nem sempre se obtêm os melhores resultados, mesmo quando se encontra o pretendido, os textos muitas vezes são enormes, dificultando a compreensão dessa informação.

De modo a ultrapassar este problema pode proceder-se à condensação da informação disponível [Jon07]. Podem ser feitos resumos do texto, mas esta solução leva-nos a outro problema, que é a dificuldade de construir o sumário de um texto. Como referido anteriormente, existe muita quantidade de informação e produzir sumários manuais para toda essa informação torna-se inviável, não só pelo tempo despendido mas também por outros factores, como por exemplo o nível de conhecimento de quem faz o sumário. Um sumário também pode ser produzido por métodos automáticos, fazendo com que se possam produzir mais sumários do que se fariam com sumarização humana e também se consigam obter informações mais abrangentes, pois neste caso o sumário não estará dependente do grau de conhecimento de um humano e poderá seguir regras específicas para a obtenção da informação presente no texto original.

O Processamento da Linguagem Natural, ou simplesmente *PLN*, é a área que tem como objectivo estudar técnicas automáticas para modelagem de texto. Mais profundamente a Sumarização Automática , ou SA, é a sub-área da PLN responsável pelo estudo das especificidades referentes ao processamento computacional de uma versão abreviada de documentos textuais. Em relação ao resumo de texto não existe uma definição completamente aceite pela comunidade científica, mas existe um conjunto de características aceites, como sendo necessárias para a elaboração de um resumo. Podem ser referidas como características de um resumo o facto deste representar uma forma abreviada de um ou mais conteúdos textuais, terá de preservar os assuntos relevantes do texto-fonte e não deve ultrapassar metade do tamanho do texto original. Estes elementos pertencem à definição de resumo dada por Radev et al. [RHM02] e são citados com frequência na literatura [DM07, GL10, Sun11]. O tamanho do resumo é calculado através de uma taxa de compressão, que é uma media indicadora da proporção entre as dimensões do texto original e da síntese do mesmo. Essa medida pode ser calculada, tendo como base a percentagem do tamanho de texto pretendido em relação ao tamanho do texto-fonte, o número de frases ou palavra pretendidas no sumário. Na figura 1.1 está representado um texto noticioso e duas das suas possíveis sínteses, uma produzida por um humano e outra de forma automática. Este tipo de síntese continua a ser bastante estudada embora não seja uma preocupação recente.

Os humanos sempre produziram resumos de textos científicos, embora nos últimos anos com o avanço da tecnologia se tenham estudado vertentes automáticas para essa tarefa. Quando em 1958, Luhn[Luh58] propôs uma das primeiras abordagens para a produção de sumários, já o

Figura 1.1: Exemplo de Sumarização de uma Notícia

Texto Original:

O número de estudantes candidatos ao ensino superior caiu 3,4% este ano, face a 2013.

Em 2014, há menos 3141 estudantes a candidatar-se à universidade, segundo dados revelados esta segunda-feira pelo Ministério da Educação e Ciência.

Dos 158 566 alunos inscritos para exames do secundário, quase metade não quer prosseguir os estudos no ensino superior. Dos alunos inscritos, apenas 88 358, ou seja 56%, pretende entrar na universidade.

Correio da Manhã 09-06-2014

Resumo Humano:

Em relação a 2013, há menos 3141 alunos a querer seguir o ensino superior.

ResumoExtrativo:

O número de estudantes candidatos ao ensino superior caiu 3,4% este ano, face a 2013.

problema da grande quantidade de informação existia, sendo necessário sintetizar a informação para que esta fosse mais facilmente absorvida, com menos esforço e tempo despendidos. Nessa altura, esta não era a única dificuldade, uma vez que seria necessário passar os textos para uma forma que a máquina pudesse compreender. Também na década de 50, Baxendale [Bax58], elaborou um estudo sobre como tornar a produção de resumos automática, propondo o uso da posição das frases no documento e o número de palavras nos cabeçalhos, títulos e palavras de sinalização que os textos continham.

Após um período de paragem no estudo da sumarização automática de texto, cerca de uma década após as anteriores abordagens, foi proposta uma melhoria ao método de Luhn, a qual consistia em seleccionar computacionalmente palavras com maior probabilidade de transmitir ao leitor a ideia do documento [Edm69]. Desde essa altura tem havido algum interesse na área da sumarização automática, embora só na década de 90 é que tenhamos assistido a um aumento significativo de novos estudos nessa área. Esse crescimento do interesse neste tipo de técnicas deu-se com a proposta de novas abordagens, baseadas em algoritmos de inteligência artificial [KPC95], a sua combinação com outras técnicas, estudadas até então. Uma técnica introduzida neste domínio, e que passou a ser bastante importante, foi a classificação das palavras, através do uso de etiquetadores morfossintáticos[1]. Esta classificação tenta indicar o grau da palavra (plural ou singular),o tipo(nome, verbo, adjectivo, advérbio, determinante, etc.) e a forma (no caso das conjugações verbais).

Com a organização de várias conferências internacionais, no domínio da automatização do processamento de linguagem natural, entre elas a `TAC` (`Text Analysis Conference`) [tac14], que em 2008 agregou um conjunto de tarefas na área da PLN, como por exemplo as tarefas visadas pela `DUC`(`Document Understanding Conferences`) [duc07], as pesquisas e propostas de novas técnicas de sumarização automática, ou simplesmente a combinação das já existentes, ajudaram no crescimento desta área. Além das técnicas de PLN também foram sugeridos alguns métodos de avaliação automática para este tipo de tarefas, de modo a tornar esse processo

[1] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

mais rápido e independente do juízo humano.

1.1 Motivação

Embora já existam ferramentas de sumarização automática, estas não seguem sempre a mesma metodologia. Existem vários métodos automáticos que podem ser aplicados à sumarização e a dificuldade de produzir resumos de forma automática que sejam facilmente compreendidos pelos leitores, para que estes assimilem a informação principal sem ter de proceder à leitura integral do texto original é a maior motivação para esta dissertação.

1.2 Objectivos

O presente trabalho tem como objectivo principal a análise da área da Sumarização Automática de Texto em termos de subjectividade e relevância dos conteúdos textuais.
Neste sentido tenciona-se estudar e combinar diferentes métodos de sumarização automática e desenvolver uma ferramenta que facilite a obtenção de um sumário tendo como base um perfil que caracterize o utilizador. Esta aplicação foi concebida através da implementação e combinação de técnicas e ferramentas já existentes.

1.3 Estrutura da Dissertação

Esta dissertação está organizado em quatro capítulos.

- **Introdução**
  É o presente capítulo, neste é feita uma breve introdução ao trabalho e aos seus objectivos.
- **Estado da Arte** Neste capítulo é feita uma abordagem sobre o trabalho realizado na área da sumarização automática.
- **Abordagens e Experimentação**
  Neste capítulo encontra-se descrita a forma como a aplicação funciona, bem como os métodos utilizados na aplicação e os experimentos realizados.
- **Avaliação**
  Capítulo onde é são apresentadas as diferentes metodologias de avaliação existentes e os resultados obtidos com a aplicação desenvolvida.
- **Conclusão e Trabalho Futuro**
  Neste capítulo serão apresentadas as conclusões e perpectiva futura identificada durante a elaboração do trabalho.

Capítulo 2

Estado da Arte

No geral, a sumarização é uma actividade bastante comum na comunicação entre pessoas. Não só em textos científicos, notícias, prefácios de livros, sinopses de filmes mas também no relato de eventos, não é feita a narração detalhada de todo o acontecimento preferindo, assim, um resumo. Esta representação sucinta é portanto uma acção bastante trivial, para qualquer pessoa, muitas vezes, feita de forma quase sub-consciente [MPER01].
Embora esta seja uma tarefa frequente, é uma acção bastante complexa que envolve processos de raciocínio e memória inerentes à condição humana mas de uma modelação muito complicada a nível computacional. De modo a tentar automatizar este processo, é necessário entender primeiro o comportamento humano na construção destas representações textuais, para que se simule, computacionalmente, de uma melhor forma esta acção.
Neste capítulo são introduzidas metodologias estudadas ao longo dos anos no domínio do Processamento de Linguagem Natural, mais especificamente no campo da Sumarização Automática.

2.1 Definições e Classificações de um Sumário

De forma a definir um objectivo neste tipo de pesquisas é necessário, em primeiro lugar, indicar o que é um sumário. Neste sentido, foram vários os trabalhos que tentaram procurar por uma definição única e abrangente, no entanto, essa tentativa revelou-se fracassada, pois não foi possível chegar a um consenso. Surgiram, então, definições para trabalhos específicos, como é o caso da sumarização automática, em que um sumário é definido como sendo um texto produzido por um ou mais textos-fonte, contendo uma parte significativa da informação mais relevante presente nesse conjunto de textos originais, não ultrapassando metade do tamanho do texto-fonte [FWRZ05, DM07, Hov02]. Mani [Man01] acrescentou que na sumarização multi-documento, também deverá ser eliminada a redundância, tendo em conta as similaridades e diferenças entre os diversos conteúdos. Segundo Hovy [Hov02], a definição de texto inclui documentos multimédia, *on-line*, hipertextos, etc.
Os sumários podem ser classificados de diversas formas, geralmente associadas à sumarização automática, embora estes métodos também possam ser utilizados para a classificação de sumários produzidos por humanos. Spark Jones [Jon98], numa das suas pesquisas, sugeriu que se devem ter em conta alguns factores que possam influenciar o sumário, dividindo-os em três classes:

- **Factores de entrada**: referentes às características presentes num texto-fonte que possam influenciar o seu sumário. Por sua vez este tipo de factores pode ser dividido em três grupos: **formato do texto**, ou seja, a forma como o documento está estruturado; **tipo de tópico**, subdividido em três (comum, específico ou restrito) e o **número de fontes**, quantidade de documentos originais utilizados para o resumo;
- **Factores de saída**: estão incluídos nesta classe factores como a **abrangência**, conteúdo do texto-fonte que o sumário cobre; **formato**, tipo do texto no sumário (corrido ou estru-

turado); **estilo**, objectivo do sumário em relação ao texto fonte em termos de informatividade, indicatividade, crítica ou agregadora;

- **Factores de finalidade**: constituem o grupo mais importante de factores, pois têm em conta a situação, contexto para o qual se usará o sumário; audiência, tipo de leitor para quem é feito o sumário; tipo de uso a que se destina o sumário, de outra forma, objectivo do sumário.

De uma forma mais exaustiva podem classificar-se os sumários consoante as suas características, como presente na tabela 2.1.

Grupos de Classificação	Tipo de Sumário
Abordagem de Geração	Extractiva
	Abstractiva
Valor informativo	Indicativo
	Informativo
	Crítico ou Avaliativo
Teor	Genérico
	Baseado em Consulta
Limitações	Dependente de Domínio
	De Género Específico
	Independente
Número de Fontes	Documento Único
	Múltiplos Documentos
Idiomas	Monolingue
	Multilingue

Tabela 2.1: Tipos de Sumário

Como referido anteriormente estas classificações são geralmente atribuídas à sumarização automática e têm como objectivo indicar as características dos mesmos. Quanto à geração de sumário este pode ser:

- **Extrativo**: formado por unidades textuais, normalmente frases, seleccionadas e copiadas do texto original, na íntegra. Este é um tipo de sumários presesnte em sistemas de sumarização automática que adoptem uma abordagem superficial.

- **Abstractivo**: neste tipo de sumários o texto é composto por versões mais curtas das frases originais e por vezes incorporadas umas nas outras.É por isso mencionado, este tipo de sumários, como fazendo uso da interpretação de um texto. Este tipo de sumários engloba o uso de técnicas linguísticas que permitam analisar e interpretar o texto, para serem encontradas novas formas, humanamente inteligíveis, para expressar os conceitos presentes no texto original.

Em relação ao valor informativo de um sumário, [Hut87] classificou o sumário como sendo:

- **Sumário Indicativo**: estes não podem substituir os textos-fonte, pois nem sempre preservam as informações mais relevantes do texto original, transmitindo assim uma vaga ideia destes. O objectivo associado a este tipo de sumário é o de ajudar o utilizador a escolher quais os testos-fonte mais interessantes. Um dos possíveis usos está relacionado com a indexação de documentos em sistemas de recuperação de informação, onde se torna mais eficiente o uso de sumários curtos que indiquem a ideia chave do documento-fonte;

- **Sumário Informativo**: estes preservam todos os aspectos mais importantes do texto-fonte, podendo assim, substitui-lo. Deverá conter detalhes quantitativos e qualitativos de cada um dos tópicos presentes no texto original;
- **Sumário Crítico ou Avaliativo**: têm como objectivo a avaliação do documento original com outros trabalhos na mesma área, tendo em conta uma descrição objectiva dos tópicos, métodos e dados do texto-fonte. Por isso, captam a opinião do autor sobre um determinado conteúdo.

Quando nos referimos ao teor de um sumário, este poderá ser:

- **Genérico**: classifica todos os tópicos, do texto original, com o mesmo grau de importância e estes devem ser incluídos no sumário;
- **Baseado em Consulta**: tem em conta um tópico ou consulta específica para que se inclua, com maior detalhe, no sumário gerado.

O domínio ou género de escrita de um sumário também representam limitações para a produção deste, sendo elas as seguintes:

- **Sumário dependente de domínio**: é um sumário que tem como objectivo abordar em maior detalhe um domínio específico. Assim, consegue-se produzir um sumário com mais qualidade. Os sistemas automáticos que geram este tipo de sumário não conseguem produzir sumários de qualidade para outros domínios;
- **Sumário género específico**: tem em conta o género de escrita do texto-fonte para que se tire partido das suas características. Este tipo de sumários é útil quando se pretende produzir um texto argumentativo sobre o conteúdo do documento original;
- **Sumário independente de género e domínio**: são sumários que não dependem de um domínio específico nem do género de escrita, logo, utilizam uma abordagem superficial na análise do texto-fonte. Por vezes, sistemas que desenvolvem este tipo de sumários também têm capacidade para explorar conteúdos específicos.

Em termos do número de fontes o sumário pode ser elaborado a partir de:

- **Documento Único**: um sumário produzido através de um único documento original;
- **Múltiplos documentos**: neste tipo de sumários, como entrada,são utilizados vários textos-fonte, levando isso, a uma consulta e remoção da redundância entre os diversos documentos, sendo eles da mesma origem ou não, para que seja garantida coerência e uma melhor qualidade nodo sumário.

Por último, e não menos importante o sumário também está sujeito ao idioma de escrita, neste caso pode ser classificado como:

- **Monolingue**: são sumários em que os textos-fonte estão todos no mesmo idioma, Sistemas que criem sumários teste género, normalmente são dependentes do idioma;
- **Multilingue**: sumários em que a informação de entrada esteja em diferentes idiomas, criando assim um sumário não dependente de um único idioma. Sistemas automáticos deste género exploram características dos testos-fonte que contenham transversalidade linguística.

2.2 Sumarização Humana de Texto

No âmbito da sumarização existem algumas vertentes, entre elas a sumarização textual. Esta forma de redução do conteúdo textual é uma actividade comum na vida de qualquer pessoa. Em qualquer área, social, académica ou profissional a informação presente por meio textual é um instrumento muito útil para a comunicação e actualização da mesma. Por exemplo, no meio académico recorre-se muitas vezes à sumarização de conteúdos textuais para que estes possam ser mais rapidamente assimilados, evitando assim a leitura da bibliografia recomendada, também no meio cientifico se recorre a este tipo de representações textuais para que se possa escolher mais facilmente o material literário a consultar [RP03].

O objectivo da sumarização textual é a redução do texto-fonte, através da distinção entre o conteúdo mais relevante e aquele que possa ser descartado. As etapas que uma pessoa segue para sumarizar o texto são a identificação da sua essência e em seguida a adição de informação presente no texto-fonte para completar o sumário[Man01]. Embora esta pareça uma abordagem simples, não o é, pois existem algumas características que podem influenciar o conteúdo do sumário, tais como as características do autor do sumário, das características do potencial leitor e a importância subjectiva que tanto o leitor como o autor atribuem ao teor do texto. Estas características poderão fazer com que não só o conteúdo mas também a estrutura do sumário possa ser modificada de acordo com a forma na qual se pretende fazer a representação do conteúdo do texto-fonte.

Sínteses de notícias, obras literárias, filmes, trabalhos científicos, entre outros, pertencem a toda uma diversidade se sumários produzidos por humanos. Devido a este grande conjunto de temas, para os quais é possível produzir um sumário, podem ser encontradas vários pressupostos e características específicas, mas também diversos sumários para o mesmo texto-fonte[MPER01]. Por exemplo, um autor de um sumário noticioso envolvendo a actualização das bolsas pode considerar um bom título mencionar o valor da queda das acções de uma empresa. Assim, o seu sumário poderia ser "BES a cair quase 7% na Bolsa de Lisboa". Para o mesmo acontecimento, outro sumarizador humano pode priorizar os efeitos dessa queda no valor das acções.

Estes factores revela um grande diversidade de sumários para o mesmo texto-fonte, visto que ao serem produzidos por diferentes sumarizadores humanos o conteúdo informativo, a multiplicidade frásica ou estrutural e os interesses do autor produzem essas diferenças no sumário[Sen05]. Segundo Chen et al. [CLW07], abordagens cognitivas ajudam na compreensão do processo de estruturação dos sumários produzidos por humanos, das características do texto-fonte que influenciam o sumário ou como a finalidade do mesmo pode modular a sua síntese. Além disso os autores ainda referem que o estudo psicológico da sumarização humana em laboratório tem sido bastante útil para a sumarização automática. Várias experiências deste género foram feitas, algumas revelaram características humanas na elaboração de sumários através da construção de uma hierarquia no discurso que permite obter pistas sobre a recuperação de informação da memória [KvD78]. Outras estudaram o processo de elaboração de sumários por parte de sumarizadores humanos profissionais, como referido em [EN98, ENMS95] estes adoptam uma estratégia de cima para baixo (*topdown*) na estrutura discursiva de um texto. Jing, no seu trabalho[Jin01], analisa a literatura [EN98, ENMS95, Thu26, Fid86] na busca por directrizes para a sumarização textual e conclui que além destas serem gerais ou de muito alto nível também não são consensuais. Devido a este infortúnio Jing [Jin01] para tentar encontrar técnicas utilizadas por sumarizadores humanos e que pudessem ser aplicadas em contexto automático, realizou uma séries de análise em textos-fonte e seus sumários produzidos por sumarizadores profissionais. Tendo em conta essas análisas, concluiu que os sumarizadores profissionais reutilizam frequen-

temente a informação presente no texto-fonte mas não se limitam a extraí-lo. Sumarizadores humanos editam as frases extraídas alterando a sua estrutura e mantendo a ideia original. Isto, já tinha sido verificado em [EN98, ENMS95], quando os autores indicaram o uso de fragmentos das frases do texto original, por parte dos sumarizadores humanos.
O autor [Jin01] identificou um conjunto de seis operações, denominando-as de operações de revisão, podendo ser usada separadamente, sequencialmente ou em simultâneo, que consistem na **redução de frases** (remoção de excertos ou palavras das frases), **combinação de frases** (duas ou mais frases ou excertos das mesmas são combinados), **transformação sintáctica** (é alterada a estrutura sintáctica da frase), **parafraseamento léxico** (alteração das frases por outras com o mesmo significado), **generalização** (reposição das frases por outras com descrição mais geral) e **especificação** (alteração de frases por outras com conteúdo mais específico), e **reordenação** (alteração da ordem das frases extraídas). Além disso, é observado que nem todas as frases do texto-fonte são incluídas no seu sumário.
Em [RP03], Rino et al. destacam também 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.
A modelação computacional do sumário produzido por um humano é uma tarefa difícil, pois o nível de subjectividade e a exigência de representações muito complexas do conhecimento do domínio ainda representa um grande entrave à automatização de sumários escritos. No geral, São adoptados dois tipos de abordagens para a sumarização automática, a **superficial**, em que apenas são tidas em conta as carectirísticas estruturais do texto-fonte, e a **profunda**, na qual a análise é baseada no conteúdo explicito dos textos-fonte bem como nas suas características estruturais.
Spärk Jones [Jon07, Jon98] identifica três fases para a geração automática de um sumário: a interpretação, a transformação e a geração. Estas fases coincidem com os três passos seguidos por sumarizadores humanos, a **interpretação** do texto-fonte, que envolve a leitura e compreensão do mesmo; a **identificação** das informações mais relevantes; a **condensação** do conteúdo linguístico do sumário, podendo isto levar a uma nova estrutura e expressão linguística de texto incluído no sumário.

2.3 Sumarização Automática

O Processamento de Linguagem Natural é um sub-campo das Ciências da Computação, no qual se insere a Sumarização Automática. Esta técnica tem como objectivo a produção de um sumário de forma automática, isto é, uma máquina processa um texto de forma a gerar uma versão condensada do mesmo, tendo em conta alguns requisitos [Man01]. Geralmente esta técnica também é associada à área Linguística e da Inteligência Artificial.
O interesse nesta área surgiu devido à necessidade de indexar a grande quantidade de informação existente, principalmente ao nível das publicações científicas, mas devido à falta de recursos tecnológicos foi decidido que o conteúdo a ser armazenado seriam os sumários e não os textos originais, pois assim ceria mais eficiente a pesquisa e selecção dessa informação.
Embora a ideia fosse boa, problemas como a falta de qualidade dos sumários, a baixa quantidade de documentos num formato legível pela máquina e a dificuldade em transformar o conteúdo para esse tipo de representação [Edm69, Bax58, Luh58], levou a um estagnar da investigação destas técnicas. Mais tarde, com a chegada da *Internet*, como uma grande fonte de informação, o interesse reviveu-se e deu origem à necessidade de pesquisar, seleccionar e assimilar

informação disponível em grande escala de forma eficiente.
Devido à ineficiência humana na síntese de grandes quantidade de informação,tem vindo a verificar-se um melhoramento dos métodos de Recuperação de Informação (RI) tornando o acesso aos documentos cada vez mais rápido e eficaz. Assim, o processamento de uma grande quantidade de informação também poderá seguir os passos da RI.
Segundo a autora [Llo08], os sistemas de sumarização automática são descritos como seguindo um modelo de três fases [Hov02, Jon98, Man99]. Spark Jones apresentou um modelo conceptual de um sumarizador automático de texto. Nele estão presentes as seguintes fases:

- **Interpretação** Através da análise de um ou mais textos-fonte é criada uma representação computável dos mesmos,
- **Transformação** É formulada uma representação do resumo, num formato computável, tendo como ponto de partida o resultado da etapa anterior,
- **Geração** Obtida a representação do sumário, esta é transformada numa representação textual.

Embora esta seja uma representação simples, o processo de sumarização humana é uma tarefa muito complexa. Até meados dos anos 90, houve pouco progresso na área da sumarização automática, sendo apontada como principal razão, por Spark Jones [Jon93], a dificuldade em modelar o processo de sumarização humano, de forma adequada. Em [MPER01], os autores referem que com base nos aspectos linguísticos do texto-fonte é possível explorar até um certo nível de características estruturais e de composição.
Em [Man99], é sugerida uma forma de classificação para analisar o processamento de métodos de sumarização automática. Essa classificação é dividida em três abordagens diferentes, tendo em conta o seu nível **superficial**, das **entidades** ou do **discurso**. A **abordagem superficial** inclui um conjunto de características que visam a representação da informação com características superficiais que sejam combinadas de forma a obter uma função de relevância, de modo a extrair a informação mais relevante presente no texto. Essas característica são:

1. **Características temáticas**: é calculada a frequência dos termos no texto e os mais frequentes são considerados como sendo os mais relevantes. As frases que contenham mais termos relevantes terão uma maior probabilidade de serem escolhidas para o sumário;
2. **Localização**: é tido em conta a localização das frases para decidir se serão ou não transcritas para o resumo. Geralmente são tidas em conta excertos de títulos, cabeçalhos ou frases que ocupem posições predeterminadas no texto, tais como na introdução ou conclusão do mesmo;
3. **Características de fundo ou contexto**: neste caso são consideradas mais relevantes as frases que contenham um maior número de palavras que estejam incluídos nos títulos ou cabeçalhos;
4. **Termos ou expressões sinalizadoras**: é um conjunto de unidades textuais, expressões ou palavras, que revelem um certo nível de relevância, tais como: tópicos específicos: "bónus", "lucro", termos de realce: "importante", "em particular", ou também termos e expressões que indiquem a presença de uma síntese: "em conclusão", "em suma". Esta palavras são denominadas de termos ou expressões sinalizadoras.

A abordagem, ao nível das entidades, constrói uma representação interna do texto tendo em conta os padrões entre elas de modo a ajudar no processo de pesquisa por conteúdo relevante. Em suma, são modeladas as relações entre as entidades presentes no texto. Essas relações incluem:

- **Similaridade**: pode ser calculada por técnicas linguísticas ou pela justaposição do vocabulário, podendo verificar-se em diferentes níveis. Por exemplo dois excertos de texto que partilhem os termos ou duas palavras que partilhem a sua forma canónica;
- **Proximidade**: obtida pela distância entre duas unidades textuais, este é um factor importante para se estabelecerem relações entre excertos textuais;
- **Co-ocorrência**: tem em conta a relação entre dois termos e a sua ocorrência em contextos comuns;
- **Semelhança léxica**: esta relação é conseguida com o auxílio de um dicionário de sinónimos. Algumas relações incluídas nesta abordagem são a hiponímia, a hiperonímia, meronímia, holonímia.
- **Co-referência**: refere-se à identificação de relações entre expressões de referência que se co-referenciem. Com esta relação as unidades textuais são utilizadas na construção de cadeias de co-referencia;
- **Lógica**: inclui relações de concordância, contradição, vinculação e consistência lógica;
- **Sintaxe**: são criadas árvores de análise sintáctica (*parsing trees*) e consideram-se as relações, entre as unidades textuais, nelas presentes;
- **Representação de significado**: são relações semânticas estabelecidas entre as entidades textuais. por exemplo entre o sujeito e o seu predicado.

Quanto ao nível do discurso, esta abordagem modela a estrutura global do texto e a sua relação com os objectivos comunicativos. A coesão e coerência são duas características importantes na estrutura discursiva, quando usado este tipo de abordagem. A informação analisada por este tipo de métodos é:

- **Formato**: disposição do conteúdo (secções, capítulos, etc), contornos do documento ou mesmo através de marcações de hipertexto;
- **Segmentação em tópicos (*Threads of topics*)**: conteúdos com um grau de semelhança elevado poderão ser considerados como parte da mesma ideia;
- **Estrutura retórica**: geralmente representada por relações estruturadas em árvore. Indica qual a estrutura narrativa ou argumentação do texto, fazendo com que a centralidade das unidades textuais nesta estrutura reflita a sua importância.

As abordagens referidas anteriormente são exemplos de técnicas singulares utilizadas para a construção de sumarizadores automáticos. Actualmente, este tipo de sistemas implementa uma mistura dessas técnicas, ficando esses métodos conhecidos como abordagens híbridas. O tipo de sumarização mais comum é a textual mas existem outros tipo como por exemplo, sumarização de áudio [ZW00], imagens [LCH+06], vídeo [TIT+13, dAdLJdAAC08] e gráficos, diagramas ou tabelas.

Além da diversidade de conteúdos sumarizáveis, também podem ser identificadas áres onde este tipo de representação seria muito útil [Man01, Jin01]. Exemplos disso são as áreas dos **dispositivos móveis**, ***Internet***, **Bibliotecas digitais** ou **notícias multimédia**.
Em relação aos **dispositivos móveis**, tecnologia cada vez mais usada em todo o mundo e com ecrãs de dimensões reduzidas é necessário adaptar o conteúdo disponível para estas caracteristicas de visualização [BGMP01, SC06]. Como referido anteriormente, é de extrema importância recuperar e condensar a informação disponível na ***Internet***, pois esta tem vindo a apresentar uma tendência de aumento exponencial. Para melhorar a escolha dos conteúdos disponíveis nesta grande rede de informação torna-se importantíssimo o uso de métodos automáticos que permitam ajudar na pesquisa e consulta da mesma [AP00]. Existem motores de busca que apresentam um sumário das páginas presentes no seu resultado de pesquisa, mas em muitos casos ainda não são úteis devido à fraca qualidade desses mesmos sumários, como por exemplo o Google [goo14], o Yahoo [yah14], o Hotbot [hot14] e o dmoz [dmo14]. Actualmente, as **bibliotecas digitais** são uma grande fonte de informação sobre determinados temas, possuindo diversos conteúdos, tais como livros e artigos. Devido ao crescimento desse tipo de informação é necessário pesquisar por métodos eficientes que permitam ao utilizador obter o que necessita em tempo útil [LDS+13, Mly06, Ou09]. Quanto às **notícias multimédia** é de grande utilidade a apresentação de um sumário com as notícias mais relevantes pertencentes a um tema específico, emitidas durante um determinado período de tempo, pois assim o utilizador não necessitaria de consultar individualmente, todas as notícias para o tema pretendido [LFZ11, MH03].

2.4 Abordagens de Sumarização Automática

A sumarização humana é um processo muito difícil de modelar em termos matemáticos, lógicos e computacionais [JM99]. Os sumários estão dependentes da pessoa que o escreve e são influênciados por diversos factores, como referido na secção 2.1. Entre todos os tipos de classificação referidos, em 2.1, a abordagem quanto à geração do sumário é uma das mais úteis, pois é frequentemente usada para distinguir sistemas de sumarização automática. A geração do sumário pode ser considerada extractiva ou abstractiva.

2.4.1 Abordagens Abstractivas

A abordagem abstractiva aplicada à geração automática de sumários tem vindo a ser bastante estudada e aplicada, durante a última década. No geral, os sistemas que fazem uso desta abordagem extraem unidades textuais e após aplicarem alguns algoritmos de processamento linguagem natural, fundem essas expressões para obter o respectivo sumário abstracto. As técnicas de PLN usadas para gerar os sumários abstractos recorrem ao uso de uma análise profunda do texto, ao nível da sua representação semântica, inferência e geração de linguagem natural, os quais produzem resultados satisfatórios [YsCyKQ07]. Este tipo de abordagem tem como objectivo gerar sumários automáticos que se assemelhem aos gerados por humanos, pelo que tendo em conta a perstectiva humana de um sumário, esta abordagem é a mais indicada [LRFP13].
Segundo [KS14], a sumarização abstractiva pode ser classificada em duas categorias: abordagem baseada na Estrutura e abordagem baseada em Semântica. Métodos baseados em árvores, em modelos, em ontologia, em frases de liderança e corpo, e em regras de extracção incluem-se na abordagem baseada na estrutura. Quanto aos métodos baseados em Semântica podem ser classificados em modelos baseados em semântica multimodal, métodos baseados em itens

informativos e métodos baseados em grafos de semântica.

Abordagens Baseadas em Estrutura

Este tipo de abordagens codifica as informações mais importantes dos documentos através de esquemas cognitivos [GL11] como modelos, regras de extracção, árvores, ontologia e estruturas de frases de liderança e corpo.

- **Métodos Baseados em Árvores**
  Nesta técnica o conteúdo do documento é representado através de árvores de dependência. Barzilay et al. [BME99], propôs um método que fundisse automaticamente frases similares ao longo de artigos noticioso, sobre o mesmo evento. Este método produz texto conciso usando para o efeito, ferramentas de geração linguística. Inicialmente é feito um pré-processamento das frases similares usando um analisador superficial e posteriormente esses resultados são mapeados para uma estrutura de argumento-predicado. De seguida, as frases comuns são detectadas através de um algoritmo de intersecção de temas comparando as estruturas criadas anteriormente. Essas frases são seleccionadas, ordenadas e são adicionadas informações (referencias temporais, descrição de entidades). No final é utilizado um interpretador (*FUF*) e um gerador de textos (*SURGE*) para combinar e organizar as frases seleccionadas para o sumário. Como maior vantagem este método ao fazer uso de um gerador de texto consegue melhorar significativamente a qualidade dos sumários resultantes. No entanto também tem o problema de as frases do sumário ficarem fora de contexto.

  Num trabalho mais recente [BM05], do mesmo autor, foi proposta a fusão de frases através da sobreposição das mesmas, para gerar um sumário com ou sem frases sobrepostas. Na primeira fase, são analisadas as frases para obter a árvore de dependências. Após encontrar o centróide da árvore construída é gerada uma outra, passando esta última a ser a árvore de base, que é aumentada com as sub-árvores pertencentes às outras frases. Finalmente essa árvore é cortada com os constituintes predefinidos. Esta abordagem tem como limitação o facto que não conter uma representação abstracta do conteúdo seleccionado.

- **Métodos Baseados em Modelos**
  Neste método é utilizado um modelo para a representação total do documento. Padrões linguísticos ou regras de extracção são combinadas para identificar secções no texto que serão mapeadas para o modelo. Estas secções indicam o conteúdo do sumário. Um dos sistemas que usa este método é o GISTEXTER [FH02], que produz sumários abstractos de múltiplos documentos através do sistema de Extracção de Informação (EI), denominado CICERO [Sur02]. Este sistema de EI requer um modelo de representação dos tópicos para extrair a informação de multi-documentos. Neste caso, o tópico é representado como um conjunto de conceitos e implementado como uma estrutura ou modelo previamente construída, no caso de se conhecer bem o tópico, caso contrário são gerados novos modelos específicos. Esses modelos são mapeados para excertos de texto dos documentos, de modo a resolver expressões anafóricas. Finalmente os sumários coerentes podem ser obtidos com os excertos previamente formados. Este sistema de sumarização também produz sumários extractivos, além de aceitar como entrada um único documento mudando, neste caso, a forma de processar e gerar o sumário. Para um único documento, as regras de extracção são aprendidas de um corpus de resumos elaborado por humanos e seleccionadas as frases mais importantes, procede-se a uma redução textual, com comprimento máximo de 100

palavras. Esta abordagem tem a vantagem de produzir sumário muito coerentes devido ao uso da informação relevante identificada pelo sistema de EI. este método apenas funciona se as frases do sumário já estiverem presentes no texto original. Não consegue lidar com a tarefa se a sumarização multi-documento necessitar de informação sobre similaridades e diferenças ao longo de vários documentos.

- **Métodos Baseados em Ontologia**
  Para melhorar o processo da sumarização, vários pesquisadores concentraram esforços para usar ontologia (conhecimento sobre um domínio). Na *Internet* existem vários documentos que abordam o mesmo assunto. Cada domínio tem a sua própria estrutura de conhecimento (termos que definem o domínio) que pode ser representado pela ontologia.

  Um sistema baseado neste método, é descrito em [LJH05, LCJ03], funcionando apenas para artigos noticiosos em mandarim, usando para isso a ferramenta *Penn Chinese Treebank* [Xia00]. A ideia passa por modelar informação difusa, ou seja, descrever melhor o conhecimento do domínio do documento. Inicialmente, especialistas no domínio definem os conceitos do domínio para as notícias. De seguida, a fase de pré-processamento do documento calcula os termos significativos do *corpus* de notícias e do dicionário, neste caso em mandarim. Posteriormente, esses termos significativos são classificados pelo classificador de termos com bases nos eventos das notícias. Na fase seguinte, a fase de inferência difusa gera os graus de associação, para cada conceito difuso. Esses graus são associados aos vários eventos do domínio ontológico. Para terminar, o sumário é produzido pelo agente de notícias com base no domínio ontológico. Este método tem como benefício a exploração de conceitos difusos para lidar com informação incerta. Em relação às suas limitações, pode ser referido, o facto de só funcionar com idioma mandarim e ser necessário a intervenção de um especialista de domínio.

- **Métodos Baseados em Frases de Liderança e Corpo**
  Este método é baseado nas operações de inserção e substituição de frases que sejam sintaticamente semelhantes às frases do corpo e de liderança. Uma frase de liderança é uma frase que inicia um artigo, capítulo, ensaio ou notícia, neste caso é a frase de abertura e que introduz um artigo noticioso. Portanto, este método tem como objectivo, reescrever as frases introdutórias de cada paragrafo baseando-se na semelhança sintática entre estas e as restantes frases que constituem o paragrafo.

  Foi proposto por Tanaka et al. [TKK+09] uma abordagem abstractiva que fizesse a revisão das frases introdutórias de um aglomerado de notícias. Ao contrário de outros métodos, este não utiliza a relação de co-referencia das frases nominais. Uma frase nominal é aquela que possui um nome como palavra principal, ou seja, o verbo está subentendido (exemplo: em "Filho criado, trabalho dobrado." o verbo "ser" está subentendido, caso estivesse presente seria "Filho criado *representa* trabalho dobrado."). Na primeira fase, os mesmo excertos (*chunks*, também chamados de *triggers*) são pesquisados nas frases de liderança e do corpo do documento. Depois as frases máximas (candidatas a revisão) de cada excerto são identificadas e alinhadas usando uma métrica de similaridade. As frases do corpo do documento só são substituídas se forem correspondência das frases de liderança e estas últimas forem mais ricas em informação. Por outro lado, as frases do corpo podem ser inseridas no lugar das frases de liderança quando não existe correspondência entre elas. Este método tem como potencial benefício a identificação das revisões semânticas para rever a frase de liderança. As fraquezas deste método são: os erros de divisão (*parsing*)

que degradam a qualidade da frase a nível da sua gramática e repetição; e o foque em técnicas de reescrita, carecendo de um modelo com a representação abstrata para a selecção do conteúdo.

- **Métodos Baseados em Regras**
  Neste método, os documentos a serem sumarizados são representados em termos de categorias e numa lista de aspectos. O modelo de selecção de conteúdo selecciona o melhor candidato através do conjunto gerado pelas regras de extracção, para responder a pelo menos um aspecto da categoria. No final, padrões de geração são utilizados para gerar as frases do sumário.

  A metodologia proposta em [GL12], gera sumários abstractos curtos e bem escritos através de aglomerados de notícias sobre o mesmo evento. essa metodologia baseia-se em esquemas de abstracção. Este esquemas, de modo a criarem as regras de extracção, usam um módulo de extracção de informação, heurísticas de selecção de conteúdo e um ou mais padrões de geração frásica. Cada esquema de extracção lida com um tema ou subcategoria. As regras de extracção de um esquema são geradas determinando similaridades entre vários pares (verbos e nomes) e a identificação do papel da posição sintáctica. O módulo de Extracção de Informação encontra várias regras candidatas para cada aspecto de uma dada categoria. Tendo como base esse resultado, o módulo de selecção de conteúdo selecciona a melhor regra entre as candidatas para cada aspecto. O módulo de geração de sumário recebe as melhores regras, e forma o sumário baseado nos padrões de geração de cada um dos esquemas de abstracção. Actualmente, este é um sistema que consegue criar sumários com o maior grau de informação [KS14]. O facto de as regras e padrões serem escritos manualmente, torna-se num processo demoroso, consumidor de tempo e incompleto, sendo o principal ponto fraco desta metodologia.

Abordagens Baseadas em Semântica

Esta abordagem usa uma representação semântica para "alimentar" o sistema de Processamento de Linguagem Natural. Esta abordagem foca-se na identificação das frases nominais e das frases verbais através do processo de dados linguísticos[SL02].

- **Modelos Baseados em Semântica Multimodal**
  Neste método são capturados conceitos e relações entre conceitos de documentos multimodais, ou seja, documentos que contêm texto e imagens. Esses conceitos e relações são inseridos num modelo semântico que representa o conteúdo dos referidos documentos. Os conceitos relevantes são avaliados com base numa métrica e por fim, forma-se um sumário com frases que expressem esses conceitos.

  Greenbacker [Gre11], propôs uma técnica para gerar um sumário abstractivo através do modelo semântico de um documento multimodal. Esta técnica contém três passos: **primeiro**, é construído o modelo semântico usando representação de conhecimento baseado nos objectos (conceitos) organizados por ontologia (domínio); **segundo**, conteúdo informativo(conceitos) são avaliados através da sua densidade métrica, número de relações com outros conceitos e o número de expressões que mostram a ocorrência do conceito; **terceiro**, os conceitos importantes são expressos em frases. As expressões observadas são guardadas num modelo semântico para expressar os conceitos e relações. A vantagem desta técnica é a boa cobertura do documento original, pois contém o conteúdo gráfico

e textual mais saliente. A limitação desta técnica prende-se com o facto de ser avaliada manualmente.

- **Métodos Baseados em Itens Informativos**
  Os conteúdos de um sumário são gerados através da representação abstracta dos documentos-fonte em vez das frases desse documentos. A representação abstracta é um item de informação que contém o elemento informativo mais pequeno do texto.

  Uma técnica [GL11], proposta com base no método referido anteriormente, para sumarização abstractiva foi apresentada nas conferencias TAC [tac14], edição de 2010, para sumarização de notícias multi-documento. Essa técnica consiste em quatro etapas: (1) recuperação de itens de informação, é definido um objecto, que guarda os dados numa estrutura to tipo sujeito-verbo-objecto (INIT), através de um divisor sintático (*parser*) de texto ; (2) gerador de frases, o INIT é convertido directamente numa frase, através de um gerador de linguagem, natural, o SimpleNLG [GR09]; (3) selecção de frases, é calculado uma média de frequência dos INIT do documento, classificando as frases geradas através deles; (4) gerar sumário, este módulo gera o sumário através das frases com melhor classificação. A maior vantagem deste sistema é o facto de ele produzir sumários curtos, coerentes, ricos em informação e com pouca redundância. No entanto, este método tem várias limitações, destacando-se a rejeição de muitos itens de informação por ser complicado gerar frases significativas e gramaticais, a partir deles e a baixa qualidade linguística dos sumários, devido à má análise (*parsing*) do texto.

- **Métodos Baseados em Grafos de Semântica**
  Este método sumariza documentos após a criação e uma *Rich Semantic Graph* (RSG), reduzindo o grafo semântico gerado e através deste último produzindo o sumário abstracto.

  Moawad et al. [MA12], sugeriram uma abordagem que consiste em três fases: (1) criação do grafo (Rich Semantic Graph); (2) redução do RSG; e gerar texto do sumário. Na primeira fase é gerado o RSG que representa semanticamente o documento de entrada. No RSG os verbos e nomes, do documento, são representados através dos nodos do grafo, enquanto as arestas representam as relações topológicas e semânticas entre as referidas entidades. Na fase seguinte é feita a redução do RSG usando algumas heurísticas. Este grafo reduzido é utilizado na terceira fase, em que através dele é gerado o sumário abstrato. Este método gera frases curtas, coerentes, com pouca redundância e correctas a nível gramatical. A desvantagem deste método é facto de aceitar apenas documentos únicos para a sumarização.

2.4.2 Abordagens Extractivas

No âmbito da Sumarização Automática, além das abordagens abstractivas, ou seja, conjunto de metodologias que tentam gerar sumários idênticos àqueles que são produzidos por humanos, existem também abordagens mais simples e menos custosas para a máquina que executa o sistema de SA, sendo elas denominadas por abordagens extractivas. Este tipo de abordagem consiste na aplicação de um ou mais métodos de cálculo, podendo ser divididos em quatro grandes classes: **métodos clássicos**, fazem uso das abordagens estudadas no inicio das pesquisas sobre sumarização, logo, são puramente estatístico; **métodos de aprendizagem automática**, tiram partida da evolução computacional e estudos em que o objectivo passa por fornecer informação à máquina que ela possa processar e criar conhecimento, sobre a mesma, de forma automática; **métodos de teoria de grafos**, em são usados grafos na representação do texto conseguindo, por

este meio, obter relações entre as diversas unidades textuais; **métodos baseados em cluster**, de modo a identificar e abordar os diferentes tópicos de um texto; e **métodos de análise de semântica latente**, onde se tentam identificar relações de sinonímia e polissemia. De seguida serão abordados mais profundamente os métodos referidos anteriormente.

Métodos Cássicos

Este métodos são derivados de estudos levados a cabo durante as décadas de 50 e 60. Luhn [Luh58], um dos pioneiros nesta área, identificou um padrão que podia ser utilizado na SA, sugerindo que através das frequências das palavras, seria possível obter as palavras mais relevantes e assim construir um sumários extractivo. Inicialmente é feito um pré-processamento do texto, eliminando as palavras insignificantes e reduzindo as restantes à sua forma canónica, mais tarde esta redução foi estudada por Porter, levando ao desenvolvimento de um algoritmo de *stemming* [Por80] e posteriormente a algoritmos de lematização [vHR13, Gal01, PM92]. De seguida, procede-se a uma ordenação decrescente das palavras mais significativas encontradas no texto, com base nas suas frequências. O cálculo da relevância da frase é feito através da contagem de palavras relevantes e da distância entre cada uma das suas ocorrências, calculada através do número de palavras irrelevantes. No final, as frases são ordenadas de forma decrescente e aquelas que possuírem melhor pontuação serão as prioritária para a formação do sumário.
Outro método, apresentado na mesma década, é o descrito por Baxendale [Bax58], em que após uma análise exaustiva de textos, concluiu-se que a localização da frase também pode ser informativa quanto à relevância da mesma, visto que para 85% dos parágrafos analisados a primeira frase era a mais relevante e apenas para 7% a última poderia ter essa consideração. O autor também decidiu que a última frase seria importante para o sumário, mesmo com essa classificação reduzida, pois esta frase normalmente serve como elo de ligação entre dois parágrafos, preservando assim a coesão textual. Devida a este cenário, o autor sugeriu que ambas as frases fossem incluídas no sumário de um texto.
Edmundson [Edm69], cerca de uma década depois, sugeriu uma nova abordagem que incluísse as duas características referidas anteriormente e levasse em consideração o uso de palavras sinalizadoras ou indicativas (como exemplo: "conclusão" , "significativo", etc), tal como, a estrutura do documento, através da presença de palavras no texto que componham elementos salientes, como títulos e cabeçalhos . Foi elaborada uma equação para corresponder a esta sugestão:

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

Em que, F_i representa a pontuação final da frase i, os pesos para cada uma das características são denotados por p_1 a p_4 sendo o seu somatório é igual a 1 e S_i, E_i, T_i, L_i representam respectivamente a pontuação da frase i, em relação à presença de palavras sinalizadoras, palavras relevantes estatisticamente contidas na frase, palavras presentes no texto que estejam contidas nos elementos estruturais e da sua posição.

Métodos de Aprendizagem Automática

Com a massificação dos computadores, conteúdos textuais e com o desenvolvimento de novas técnicas computacionais no âmbito da *Inteligência Artificial*, a sumarização automática presen-

ciou um novo interesse por parte da comunidade científica e voltou a ser pesquisada tendo como base a implementação de métodos de Aprendizagem Automática ao domínio do Processamento de Linguagem Natural.

Em 1995, Kupiec et al. apresentaram um trabalho pioneiro em que se descrevia o uso de um classificador Naive-Bayes, para a selecção das frases a serem incluídas no sumário, indicando qual a função de classificação. Sem pôr de parte as indicações de Edmundson, sugeriu também a inclusão de características como a presença de maiúsculas nas palavras e o comprimento das frase. De seguida, as frases eram extraídas tendo em consideração as melhores classificação e o tamanho pretendido para o sumário. Um *corpus* de 188 pares de documentos serviu para treinar o classificador, sendo a sua avaliação levada a cabo por um *corpus* constituído por documentos técnicos e os seus respectivos sumários, obtidos manualmente. Os autores procederam a uma avaliação manual que visava mapear as frases dos sumários manuais em relação ao texto original para que depois pudessem avaliar os sumários automáticos. Foi concluído que ao considerar apenas as características da posição, sinalização e comprimento das frases se obtinham os melhores resultados.

O método anterior não é único no uso deste tipo de classificador. Aone et al. [AOGL99] também o utilizaram nas suas pesquisas, em conjunto com a métrica *tf-idf* de forma a tentar salientar conceitos fundamentais através das sua palavras indicativas. Esta métrica assenta na relação entre o número de presenças de uma palavra num documento para com a sua frequência num *corpus*. O valor *tf-idf* de uma dada palavra é mais elevado quanto mais frequente seja esta, no documento e mais rara no *corpus*. Ambos, pertencem ao mesmo domínio, sendo que o *corpus* utilizado para o cálculo da métrica deverá possuir grandes dimensões. Além destas características o autor também teve em conta a presença de palavras únicas ou pares de substantivos que se referissem a uma entidade. Para manter a coesão textual, foi implementada uma análise superficial do texto que pudesse reconhecer referências diversas à mesma entidade e com o uso da ferramenta WordNet [Mil95] foram estabelecidas relações de sinonímia.

Em [CO01], foi proposta a utilização de um *Hidden Markov Model* com três características presentes no texto: a localização da frase no documento, comprimento da frase (em número de palavras) e a probabilidade dos termos presentes na frase sabendo os termos que constituem o documento. As frases a serem incluídas no sumário dependem da probabilidade de a frase anterior pertencer ao sumário. Os estados deste modelo, servem para representar numa estrutura sequencial, as frases do documento. Este é dividido em $2s + 1$ estados em que s estados pertencem ao sumário e $s+1$ não pertencem ao sumário. O *corpus* usado para este trabalho contém o mapeamento das frases de sumários produzidos por humanos, com o objectivo de tentar calcular a probabilidade das transições entre os estados do modelo, baseando-se nas ligações estabelecidas entre sumários humanos e texto original.

As conferências DUC [duc07], realizadas entre 2001 e 2007 e nos anos seguintes fazendo parte das tarefas inclusas nas conferências TAC[tac14], foram também causadoras de uma forte motivação no estudo da área da sumarização automática. Foram vários os pesquisadores que tentaram criar as suas metodologias e aplicações tendo como objectivo ultrapassar a eficiência do método de referência usado nas ditas conferências, que foi analisado em [Nen05] e consiste no uso de 100 palavras para gerar o sumário através das *n* primeiras frases de uma notícia. Entre eles, Svore et al. [SVB07] desenvolveram uma abordagem baseada em redes neuronais. Nesta abordagem são utilizados pares RankNet [BSR+05] e um classificador, que após treinado, tenta identificar as frases mais importante. As características avaliadas neste sistema têm origem em registos de pesquisas provenientes das entidades e notícias da Wikipédia [Wik04]. Os autores conseguiram atingir o objectivo que os motivou, obtendo resultados significativamente melhores

que os do sistema de referência.

Métodos de teoria de Grafos

Tal como acontece na sumarização abstractiva, os métodos baseados em grafos também são usados para produzir sumários extrativos. O objectivo é representar a estrutura do texto através de um grafo, em que as suas unidades textuais (palavras, excertos, frases ou parágrafos) estejam representados nos nós do grafo e as arestas representam as relações entre eles, tais como, a similaridade, relações léxicas ou semânticas. A ideia base desta abordagem é a recomendação,ou seja, as ligações entre os nós representam a recomendação de um em relação ao outro. Essa recomendação é determinada pela importância do nó que a faz, enquanto a importância é obtida pelo número de recomendações que esse nó recebe.
Foi sugerido por Saltou et al. [SSMB97a] um dos métodos pioneiros nesta abordagem, em que era descrita a representação de um texto através das relações entre os seus parágrafos, neste caso a semelhança entre eles, calculada pela repetição do conteúdo textual. Nesta técnica os nós continham informação dos parágrafos enquanto as relações eram armazenadas nas arestas. O sumário era obtido após a identificação dos parágrafos mais relevantes. Essa relevância era conseguida através da contagem do número de arestas que cada nó possuía. Quantas mais ligações maior a relevância do parágrafo. Esta classificação de unidades textuais também é conhecido como, centralidade do texto.
Em 2004, [ER04a, ER04b], foi proposta uma nova abordagem, que em relação à anterior, utilizariam-se frases no lugar dos parágrafos, e o calculo da centralidade dessas frases seria indicativa da sua relevância. Nesta abordagem, Erkan e Radev calculam a semelhança entre as frases recorrendo ao uso da similaridade do *co-seno*, entre elas. As frases mais relevantes só são consideradas após uma redução do grafo, tendo em conta um limite mínimo para o valor da semelhança.
Mihalcea e Tarau, após presenciarem o sucesso dos algorítmos baseados em grafos, para a classificação da rede da *World Wide Web*, decidiram propor uma abordagem, que seguisse a mesma ideia, aplicada à sumarização automática [Mih05, MT05, MT04]. A importância de um nó do grafo, em ferramentas como o *PageRank* [BP98], é calculada recursivamente a partir de todo o grafo, contrapondo técnicas em que esse cálculo é efectuado com base na informação local de um nó [MT04]. Os autores do *TextRank* aplicaram o sistema a duas tarefas distintas, a sumarização automática extractiva de texto e a extracção de palavras-chave. Este sistema, consiste nas seguintes quatro fases:

1. **Identificação e adição das unidades textuais ao grafo**
   Dependendo da tarefa a realizar a granularidade da informação é adicionada aos nós, ajustando a sua granularidade;

2. **Identificação e adição das relações entre as unidades textuais**
   As ligações entre os nós representam as relações entre as unidades textuais, podem ser direccionadas ou não direccionadas e pesadas ou não pesadas;

3. **Execução do algoritmo de classificação**
   Este algoritmo é executado até convergir ou então até atingir um limite de iterações pré-estabelecido;

4. **Ordenação e selecção das unidades textuais**
   É feita uma ordenação dos nós, que contêm as unidades textuais, através da sua pontuação final e são seleccionadas as unidades textuais dos nós com melhor pontuação.

Um método explorado para sumarização multi-documento nos finais da década de 90, é descrito por Mani et al. [MBG98] como um método de cálculo para a relevância dos elementos de um grafo a partir de um algoritmo de busca baseado em propagação de activação. O grafo é gerado tendo em conta a coesão textual. Essa característica é conseguida após a representação do texto no grafo ser feito com as relações de coesão (adjacência, repetição, sinonímia, hiperonímia e co-referência) entre as palavras, representadas nas ligações entre os nós, sendo estes últimos as palavras presentes no texto. Como referido anteriormente, um algoritmo de busca baseado em propagação por activação, neste caso descrito em [CN95], é utilizado para calcular a saliência do termos. Esse cálculo passa pelas seguintes fases:

1. **Identificação dos nós de entrada**
   O utilizador define um tópico que será usado para o cálculo das relações entre este e os nós do grafo;

2. **Actualização dos valores dos nós**
   Durante a execução do algoritmo de busca baseada em propagação por activação, os valores de cada nó serão actualizados, tendo como base o tipo de relação entre os termos de cada nó e o peso dos seus antecessores. Caso não se tenham definido os tópicos, utilizam-se os valores *tf-idf* dos termos de cada nó para se proceder ao cálculo dos seus valores;

3. **Cálculo dos pesos de de uma frase**
   Os pesos são calculados com base no valor final da activação propagada, dos seus termos, dados os valores *tf-idf* iniciais e dos termos de cada tópico.

Métodos Baseados em Cluster

Esta metodologia pode ser usada em duas vertentes durante o processamento de um ou vários documentos. Uma dessas vertentes está relacionada com a divisão de que, implícita ou explicitamente, os documentos são alvo. Geralmente um documento pode ser dividido em tópicos, que vão sendo abordados ao longo do mesmo. Sendo o sumário um texto que tenta sintetizar o conteúdo de um documento fonte, é normal pensar que essa versão reduzida aborde todos os tópicos do texto que lhe deu origem. Outra vertente em que se podem utilizar métodos baseados em *cluster* é a aglomeração de documentos sobre um mesmo tema. Esta técnica é usada geralmente para a sumarização de documentos múltiplos com temas diferentes, em que a aglomeração dos documentos se torna fundamental no processo de gerar sumários dos mesmos, visto que, é convenienteabordar os temas de todos os documentos de acordo com a sua importância [GDCY02].
Os sistemas que implementam esta abordagem têm como principal função lidar com a redundância e diversidade de temas, sendo constituídos pelas seguintes fases [NM03]:

- **Representação dos tópicos**
  Nesta fase os tópicos são representados através de aglomerados de frases, não necessariamente consecutivas, mas semelhantes entre si, seguindo um determinado critério;

- **Identificação da frase mais relevante de cada tópico**
  O tópico passa a ser representado pela frase mais importante do aglomerado que lhe deu origem, reduzindo, assim, a redundância da informação extraída;

- **Formação do sumário** As frases seleccionadas anteriormente são justapostas para formar o sumário final.

Carbonell e Goldstein [CG98], foram pioneiros neste tipo de métodos, sugerindo o uso do conceito de diversidade aplicado à sumarização automática. O algoritmo que proposto pelos autores foi denominado por *Maximal Marginal Relevance* (*MMR*), o qual pretende maximizar a relevância marginal para sistemas de Recuperação de Informação e sumarização automática. A relevância marginal é um critério que luta para reduzir a redundância entre frases de um ou mais documentos. Desta forma, um documento possui um valor alto de *MMR* se for relevante para um dado tópico e pouco semelhante a outro documento. Quando aplicado este conceito a uma frase, esta diz-se com elevada relevância marginal se além da pouca similaridade com as restantes frases seleccionadas, for muito relevante para o tópico definido, pelo utilizador. Com este método pretende-se aumentar a relevância de uma frase no sumário reduzindo a sua redundância. Devido a esta propriedade, este método é aceite, em grande escala, para a sumarização automática de texto em combinação com outros métodos de diferentes abordagens.

Métodos de Análise de Semântica Latente

Ježek e Steinberger, em [JS08], definiram análise de semântica latente, (*Latente Semantic Analysis* ou *LSA*, em inglês) como sendo uma técnica algébrica-estatística totalmente automática para extrair e representar o uso contextual de conceitos das palavras em passagens de um discurso. Acrescentando que, a ideia básica é que um agregado de contextos de palavras em que uma palavra possa ou não aparecer fornece restrições que determinam a similaridade de significados de palavras e conjuntos de palavras entre si. Em [Pat07], o autor indica que esta é uma técnica estatística baseada em corpus para descobrir a relação semântica entra as palavras e através das estatísticas de ocorrência das palavras conseguimos identificar as relações de sinonímia e polissemia. Com esta técnica é possível classificar textos (parágrafos, documentos) que sejam considerados próximos mesmo contendo palavras diferentes.
A *LSA* contém dois passos importantes:

1. **Criação de uma matriz termo-frase ou termo-documento**
   Cada coluna desta matriz representa o vector pesado da frequência de termos de cada frase do documento fonte, a diferenciação entre termo-frase e termo-documento relaciona-se respectivamente com a sumarização de um único documento ou de múltiplos documentos;

2. **Aplicação do *Single Value Decomposition* à matriz**
   O *Single Value Decomposition*, ou SVD, (Decomposição de Valor Singular, em português) é uma matriz que tem como função derivar a estrutura semântica do documento, ou seja, reduzir as dimensões da matriz *LSA* às suas dimensões mais importantes.

Segundo [GL01], seleccionar frases baseadas nas pontuações da sua relevância assegura que o sumário cobre a maioria dos tópicos do documento e que ao remover, todas as palavras presentes na frase, do documento assegura-se que existirá o mínimo de redundância no sumário. A vantagem da utilização de *LSA* é que as relações conceptuais são automaticamente capturadas [LD97], ao contrário do que acontece com simples vectores de palavras, que para obterem essa informação necessitam de empregar métodos explícitos para revelar as relações conceptuais[GDCY02].

2.4.3 Aplicações desenvolvidas no âmbito da Sumarização automática

Ao longo das últimas décadas, principalmente nas duas mais recentes, foram desenvolvidas várias aplicações de sumarização automática, tendo como base abordagens extractivas, abs-

tractivas e nalguns casos uma combinação das mesmas. Alguns dos trabalho que se destacaram são os seguintes:

- **COMPENDIUM**: Este sumarizador foi desenvolvido com o objectivo de gerar sumários de artigos na área das ciências biomédicas. Consegue gerar *extracts*, em que gera o sumário seleccionando e extraindo as frases mais relevantes do texto, e também usa uma abordagem orientada a *abstracts*, tendo sido pesquisado, em [LRFP13], a implementação de uma nova estratégia combinando a informação extractiva com alguns excertos de informação previamente comprimidos ou fundidos. A nível de resultados, os autores de ambas as abordagens geram bons sumários, pois conseguem manter a informação dos documentos-fonte, embora a abordagem orientada os *extracts* seja recomendada para a perspectiva humana.

- **GistSumm**: O GistSumm [Par02] é um sumarizador automático baseado na ideia principal de um texto, isto é, selecciona e extrai a frase que melhor identifica a ideia principal do texto. Os autores deste sistema delinearam duas hipóteses para o desenvolvimento do mesmo. A primeira faz uso de métodos estatísticos para identificar a frase mais relevante do texto ou, pelo menos, aquela que se aproxime significativamente dela; a segunda, aplica a justaposição de frases relacionadas com aquela que foi previamente identificada, de modo a produzir uma complementaridade entre as mesmas. Mais recentemente, o sistema foi melhorado tendo em vista a diminuição de uma restrição, presente na proposta inicial, que se baseava na ideia de que um texto tem apenas uma ideia principal. Mesmo após esta melhoria, ainda não é capaz de lidar, satisfatoriamente, com textos que possuam mais de uma ideia principal. Este sistema é simples, no que ao processo computacional diz respeito e visa favorecer a coerência em torno da ideia principal.

- **SuPor**: [Mó03] O SuPor é um sistema automático de sumarização automática desenvolvido para Língua Portuguesa que utiliza o classificador Naive-Bayes seguindo o modelo sugerido por Kupiec [KPC95]. Este sistema permite que o utilizador escolha as características que serão usadas durante o processo de sumarização, nomeadamente, o método de Cadeias Léxicas [BE97], a computação da frequências das palavras [Luh58], localização das frases [Edm69], comprimento das frases, ocorrência de substantivos próprios [KPC95], a representação da coesão do texto através de um mapa de relacionamentos entre parágrafos[SSMB97b] e a identificação das frases mais importantes de cada tópico do texto [NSKF00]. Também foi implementada a escolha de algumas opções de processamento e pré-processamento, tais como, *Stemming* ou quadrigramas para identificar a frequência das palavras, importância dos tópicos e mapas de relacionamentos; e o uso da ferramenta *TextTiling* ou de parágrafos para identificar os tópicos do texto. Este siste obtém bons resultados o que levou ao uso do mesmo como ponto de partida para outros trabalhos. O grau de complexidade no uso do sistema constitui uma das principais desvantagens, visto ser necessário o utilizador conhecer a área da sumarização quanto aos métodos opcionais implementados.

- **NeuralSumm**: Este sistema [PRN03] utiliza uma rede neural do tipo SOM (*Self-Organising Map*) [Koh90] para classificar as frases do texto com base num conjunto de características previamente seleccionadas. A rede neural usada classifica as frases como sendo essenciais (prioritárias na selecção), complementares ou supérfluas (frases que podem ser descartadas), após organizar as informações aprendidas, durante a fase de treino, num grupo de similaridades. Entre as características usadas para Sumarização Automática encontram-se a posição da frase, presença de palavras-chave, presença de palavras indicativas, entre

outras, totalizando um conjunto de oito. O treino da rede foi efectuado com um corpus de dez textos científicos, no idioma português do Brasil e anotado manualmente de acordo com as classificações referidas anteriormente. Segundo os autores, este algoritmo atingiu resultados melhores que os algoritmos *Naive-Bayes* e *C4.5*, para o mesmo corpus. Embora nos capítulo da precisão e abrangência (*recall*) tenha obtido resultados similares a outros sumarizadores desenvolvidos, até então, Tiago Rino, em [PR04] fez uma comparação entre este e outros sistemas para o Português do Brasil e concluiu que o desempenho era inferior a outros sumarizadores supervisionados e não-supervisionados.

2.4.4 Problemas da Sumarização Automática

A sumarização automática como se pode perceber facilmente traz algumas vantagens para o campo geral da síntese de textos, facilitando esse processo e poupando tempo ao utilizador que, como é natural, recorre a este tipo de textos para absorver a informação principal, do texto fonte, no menor tempo possível.Embora esta área esteja bastante desenvolvida, com vários métodos (extractivos e abstractivos) de sumarização automática, em muitos casos ainda apresenta algumas debilitações na obtenção de um sumário coerente e que represente com maior exatidão o conteúdo do texto original. Um exemplo deste tipo de problemas são os sumários extractivos, em que o facto de extraírem literalmente as frases do texto original, podem-se perder algumas informações relevantes. Este tipo de sumário produz um sumário informativo, geralmente com má textualidade, devido a essas perdas de informação. o Exemplo seguinte mostra um desses caso em que ao ser extraída a segunda frase [F2] é perdida a referência ao sujeito da frase [F1], não permitindo, assim, perceber com exatidão, qual o tema do texto.

Texto Original:

[F1]: Japão desenvolve sensor que deteta embriaguez do condutor.

[F2]: Esta tecnologia deteta o som e a vibração produzida pelo sistema cardiovascular.

in DN `http://www.dn.pt/inicio/ciencia/interior.aspx?content_id=4123301` 14/9/2014

Uma solução para o problema da sumarização extractiva é o uso da abordagem abstractiva, em que devido aos métodos usados é possível obter textos mais ricos, a nível da informação. Como contrapartida, a sumarização abstractiva requer muitos recursos da máquina que a produz, como em utilizações de grande escala a sumarização extractiva tem conseguido resultados satisfatórios, principalmente em sumarização de documentos múltiplos, sendo mais fácil a sua implementação e menos dependente do género, domínio ou idioma, tem-se verificado um maior uso desta sumarização, em relação à sua vertente abstractiva.
De seguida serão apresentados em maior pormenor os problemas dos dois tipos de abordagem da sumarização automática.

Problemas da Sumarização extractiva

A seguinte lista indica alguns dos problemas da sumarização automática extractiva, que se podem notar quando se procede a essa tarefa para documentos únicos, tornando-se mais perceptivos quando utilizados multiplos documentos na elaboração de sumários automáticos extractivos.

- **Propensão a redundância**
  Este tipo de sumários, geralmente apresentam muita redundância, principalmente quando

aplicados a vários documentos sobre o mesmo tema, isto acontece porque as várias unidades textuais presentes num conjunto de documentos obtêm pontuações semelhantes quando são igualmente relevantes nos diferentes textos processados, sendo muitas vezes incluídas no sumário;

- **Frases extraídas são longas**
  As frases extraídas por estes tipo de abordagem, são geralmente mais longos que a média de comprimento das restantes, este problema leva a que, por vezes, conteúdo não relevante seja extraído aumentando o tamanho do sumário que em caso contrário poderia conter frases mais informativas;

- **Informações imprecisas**
  Devido à simples extracção de frases algumas informações não são apresentadas de forma precisa, levando a que, por vezes, informações contraditórias estejam mal representadas no sumário;

- **Incoerência do texto do sumário**
  Métodos de extracção pura podem levar à ocorrência de falta de coerência em passagens, ou mesmo na totalidade do sumário. As referências anafóricas são um problema recorrente neste tipo de sumários, onde por vezes, se perdem informações que ajudam a definir o tema, ou seja, certas unidades textuais, como pronomes, estão presentes no sumário sem a devida referência à pessoa ou entidade pretendida. Num caso extremo, a justaposição de frases pode indicar uma interpretação errada do texto. Outro caso recorrente é o aparecimento de expressões temporais que, com a falta de uma referência textual à data pretendida, podem induzir em erro o leitor. Com vista à resolução deste problema foi sugerido, em [GL10], o uso de datas concretas em vez das expressões temporais presentes no sumário.

Problemas da Sumarização Abstractiva

Alguns do problemas verificados nos sumários extractivos também podem ser encontrados em sumários abstractivos, com a agravante de que com o uso de técnicas linguísticas esses problemas possam aparecer também no conteúdo das frases. O principal problema comum é a incoerência do texto, que nos sumários extractivos apenas aparece entre as frases, na abordagem abstravtiva, pode notar-se esse problema na construção de algumas frases.
Após alguns estudos,demonstrados em [EN00], efectuados sobre a preferência dos utilizadores, em relação aos sumários automáticos, concluiu-se que a abordagem extractiva produz sumários mais informativos , indo mais ao agrado dos utilizadores. Estes não são muito adeptos dos sumários abstractivos, pois ao contrário da simples extracção de frases que expõem as ideias do autor, tal como ele as escreve, os sumários automáticos abstractivos, tornam-se muito pormenorizados e alteram a textualidade artificialmente.
Devido a esse problema a sumarização abstracta ainda possui a grande dificuldade na representação da informação ou conhecimento, pois este tipo de sistemas só consegue resumir conceitos que a sua representação consegue captar. A componente tecnológica ainda representa um entrave à melhoria dos sistemas de PLN, levando a que os métodos abstrativos da sumarização automática ainda não consigam fazer uma representação de um qualquer domínio sem o auxílio de um *corpus* específico, logo dependente do domínio.

Em [CNP06], após alguns estudos sobre sumarização, foi concluído que ambas as abordagens apresentam um bom desempenho quantitativo. Em relação à qualidade dos sumários produzidos, diferentes razões, embora complementares, fazem com que ambos os sistemas consigam obter um bom desempenho.

Assim, conclui-se que o uso das técnicas extractivas aplicadas à sumarização automática ainda são as mais utilizadas, devido aos motivos referidos anteriormente.

Capítulo 3

Abordagem e Experimentação

O principal objectivo deste trabalho é a experimentação de técnicas de sumarização automática extractiva tendo em conta características de subjetividade potencialmente presentes no texto e o perfil do utilizador. Foram exploradas diversas técnicas e formas de combinação das mesmas para atingir o propósito pretendido.
Neste capítulo é descrita a forma como o sistema foi implementado. São expostos os métodos utilizados para levar a cabo as diferentes fases do processo de sumarização automática, descritas no capítulo anterior. Também são apresentadas algumas decisões quanto ao uso de algumas ferramentas já disponíveis e à escolha das mesmas, de forma a atingir os objectivos propostos.

3.1 Ferramentas Exploradas

Para a elaboração de um sistema de sumarização automática, inicialmente deve definir-se o conjunto de idiomas a serem usados ou numa abordagem mais geral, pode optar-se por a implementação de métodos independentes do idioma. No caso do presente trabalho foram escolhidos dois idiomas, o português e o inglês, devido à maior facilidade em encontrar ferramentas específicas e experiência em ambos os idiomas. Neste sentido, foi necessário encontrar *corpora* que contivessem textos, nos referidos idiomas e para uma avaliação final, pudessem incluir sumários humanos ou automáticos de forma a classificar os resultados obtidos. Neste domínio existem vários conjuntos de textos com os quais são fornecidos os seus sumários e por vezes indicando o tema, que pode ser usado para definir um utilizador. Exemplos desses *corpora* são:

- Para o idioma **inglês**:
  - **Corpus das conferências DUC[duc07]**:
    Nas conferências `DUC` (`Document Understanding Conferences`) costumam ser compilados conjuntos de textos noticiosos e respectivos sumários. Nas edições de 2001, 2002 e 2003 o principal enfoque das conferências `DUC` era a sumarização de documentos únicos, passando, desde 2004, a focar-se em sumarização de documentos múltiplos, sempre no idioma inglês. Como referido no capítulo 1, estas conferências, passaram a ser incluídas numa tarefa das conferências `TAC` (`Text Analysis Conference`) [tac14], a partir de 2007.
  - **Corpus de treino do *LingPipe***:
    O *LingPipe* é um conjunto de ferramentas para processar texto usando linguística computacional, entre elas um conjunto de *corpora* para treino com várias vertentes incluindo a polaridade e a subjectividade das palavras, baseando-se num conjunto de frases retiradas de páginas de análise de filmes [PL05, PLV02, PL04].
  - **SentiWordNet**:
    O presente corpus é derivado do `WordNet` [Fel98] e inclui palavras classificadas quanto ao seu tipo (nome, verbo, adjectivo ou adverbio), à sua polaridade (dividida em

duas classificações, positiva e negativa), contendo um conjunto de sinónimos e alguns exemplos de uso. No nosso trabalho este *corpus* é utilizado para a classificação das palavras quanto à sua objectividade, sendo abordado em maior pormenor mais à frente.

- Para o idioma **Português**:
  - ***Corpus* TeMario**:
    Este *corpus* foi criado para sumarização automática de textos, com o objectivo de ajudar nos processos de análise linguística, treino de sumarizadores automáticos e sua avaliação. Inclui 100 textos jornalísticos e os seus respectivos sumários. Segundo os seus autores, os sumários presentes no *corpus* são elaborados por um sumarizador profissional, professor e consultor de edição de textos em português [PR03].
  - ***Corpus* CETEMPÚBlico**:
    O CETEMPÚBlico [SR01a] é um *corpus* em português europeu desenvolvido a partir de um conjunto de textos do jornal Público [pub]. Este foi elaborado tendo como foque a construção de um *corpus* de grandes dimensões para o idioma português. A sua estrutura indica alguns temas para as notícias e seus títulos embora o verdadeiro objectivo do seu desenvolvimento não seja a avaliação de sumários, mas com propósitos mais gerais.

Outras ferramentas de grande utilidade são aquelas que permitem fazer o processamento e o pré-processamento do texto. Entre elas podem ser destacadas ferramentas de Análise Sintáctica do texto, redução das palavras (ou lematização), identificação de tópicos e identificação de objectividade.
A análise sintáctica do texto ficou a cargo da biblioteca *Hultig* [fHLTB], que contém um conjunto de ferramentas para processamento de texto. Com essas ferramentas é possível fazer a separação do texto em palavras e frases, manipular e obter algumas estatísticas sobre um grande volume de texto, entre outras opções.
Outras ferramentas úteis no pré-processamento de texto são os *stemmers*, ou métodos de redução de palavras. São denominadas assim porque têm como objectivos reduzir as palavras retirando-lhes os seus sufixos, numa tentativa de aproximar duas palavras da mesma família em termos de grafia. Neste campo podem ser destacadas duas ferramentas, uma com origem no trabalho de Porter [Por80], apelidada *Porter Stemmer*, que tem como função a redução das palavras no idioma inglês. Para português existe uma ferramenta chamada *PTStemmer* [pts] que inclui implementações dos stemmers de Orengo [OH01], Porter [Por80] e Savoy [Sav99].
Existem alguns trabalhos em que era pretendido construir ferramentas para detecção de tópicos num texto, alguns estão incluídos noutras ferramentas, outros têm como finalidade única essa detecção. Três desse trabalhos são os seguintes:

- **Morphadorner**:
  Esta aplicação [Bur13] foi desenvolvida em linguagem *JAVA* e funciona como um gestor de *pipeline* para processos que executam o embelezamento morfológico das palavras num texto. O termo "embelezamento" refere-se a anotação ou marcação, ou seja, anexação de comentários aos textos .Com o *MorphAdorner* são fornecidos métodos para melhorar o texto com ortografia padrão, partes do discurso e lematizadores. Esta aplicação fornece ainda ferramentas para identificação de texto, reconhecimento dos limites das frases, extracção de nomes e lugares. No presente trabalho, o *MorphAdorner* é utilizado para identificar os limites dos tópicos de um texto.

- **TextTlling**:
  O TextTilling [Hea97], e uma técnica para subdividir automaticamente os textos em unidades multi-paragrafo que representem passagens ou subtópicos de um texto. O algorítmo assume que um conjunto de palavras é usado durante um sub-tópico e quando esse conjunto altera significativamente o sub-tópico também muda. São utilizadas coocorrências e distribuição léxica para determinar a segmentação de um texto. As secções do texto são indicadas através de *tiles* que conseguem uma boa correspondência em relação a julgamentos humanos das principais fronteiras entre sub-tópicos de artigos de revistas de ciências.

- **Mallet**:
  Esta ferramenta está implementada em linguagem *JAVA* e foi desenvolvida para o processamento estatístico de linguagem natural, classificação de documentos, aglomeração (*clustering*), modelagem de tópicos, extracção de informação e outras aplicações de aprendizagem automática [McC02, mal]. Inclui uma grande variedade de algoritmos (entre eles Naïve Bayes, Entropia Máxima e Árvores de Decisão) e código para avaliar o desempenho dos classificadores usando algumas métricas comuns.

3.2 Visão Geral do Sistema

O sistema desenvolvido tem como finalidade a produção automática de um sumário seguindo a abordagem extractiva, explicada em 2.4.2, em que, de uma forma geral, se pretende extrair os elementos (neste caso frases) mais relevantes de um texto para gerar o sumário do mesmo. Neste sistema foram implementados métodos estatísticos para identificar as frases mais relevantes, métodos de identificação de tópicos e estruturas que possam indicar termos relevantes para um determinado perfil de utilizador. O sistema suporta os idiomas de Português e Inglês. A implementação foi levada a cabo com recurso à linguagem de programação *JAVA*.

Na Figura 3.1 está representado o diagrama da aplicação, onde, com fundo cinzento e forma oval, estão presentes os documentos de entrada, *corpus* geral, documento do *SentiWordNet* e documento fonte, ou opções escolhidas pelo utilizador opções, como o perfil e a taxa de compressão. O ficheiro de saída, ou seja, o sumário gerado, é representado por um retângulo preenchida a cinzento.

Assim, para produzir um sumário automático no sistema o utilizador terá de indicar o documento de origem, a taxa de compressão, o documento com as classificações da objecdtividade (*SentiWordNet*) e o perfil que pretende usar para a sumarização. O conteúdo textual passa por um pré-processamento em que são identificadas as unidades textuais (palavras e frases), de modo que se possam efectuar as devidas operações de classificação desse conteúdo. De seguida, são calculadas a objectividade, relevância, bónus de localização da frase e identificados os tópicos, caso o utilizador assim o pretenda, para proceder à classificação das frases, tendo em conta essas características. As frases são ordenadas pela sua classificação e é construído um conjunto com as frases melhor classificadas. No próximo passo são extraídas, desse conjunto, o número de frases correspondente à taxa de compressão escolhida e ordenadas pela sua posição no texto original. Finalmente essas frases são apresentadas ao utilizador, representando o sumário final.

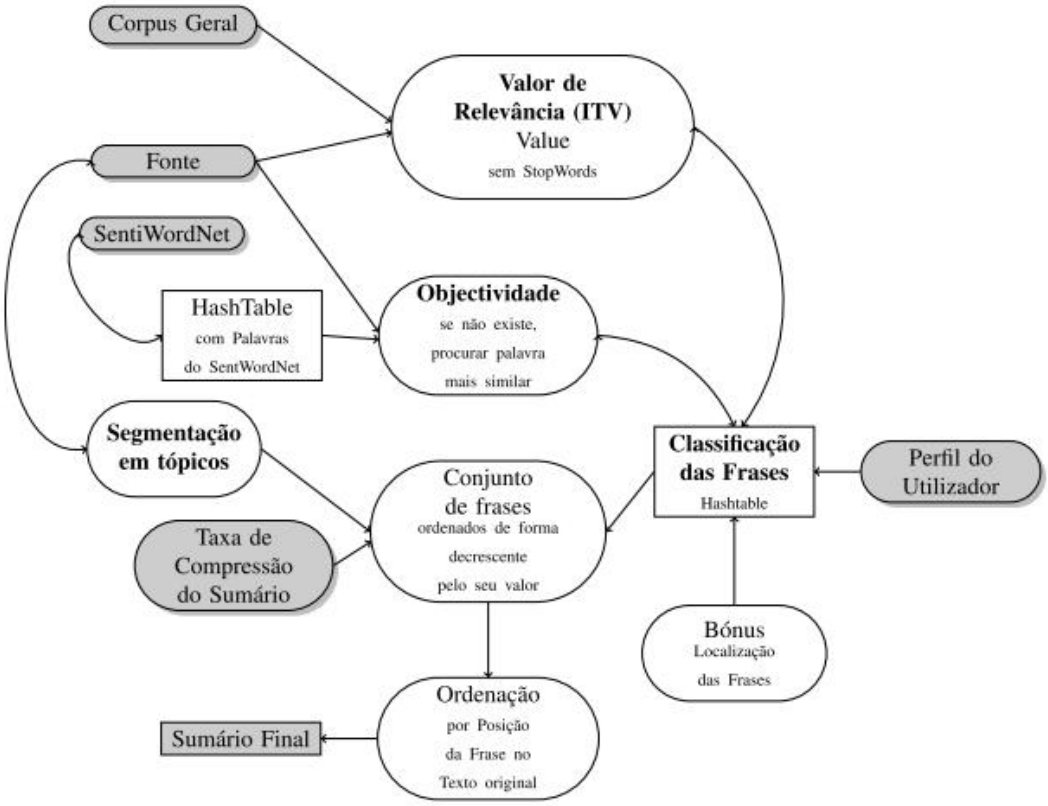

Figura 3.1: Diagrama da Aplicação

3.3 Pre-Processamento

A sumarização extractiva de texto é geralmente realizada em duas etapas, sendo a primeira o pré-processamento do texto ou documento de entrada, onde se procura criar uma representação computacional do conteúdo pretendido, e a segunda o processamento da representação gerada na primeira fase, em que se tenta identificar o conteúdo relevante do texto original para se gerar o sumário.
A etapa do pré-processamento tem um papel muito importante, visto que a representação do conteúdo textual tem de ser bastante precisa e interpretável pelo sistema para que o seu processamento seja eficaz e eficiente. De modo a obter essa representação, esse processo, geralmente, inclui as seguintes fases:

- **Análise Sintéctica do Texto (*Parsing*)**:
  Permite a identificação dos elementos textuais, tais como frases, identificadas pela presença de sinais de pontuação como o ponto final, de exclamação ou de interrogação, e palavras, geralmente, identificadas pela presença de espaços. Assim, o texto é transformado em vectores ou listas de unidades significativas do texto, como palavras, pontuação ou números. Este processo pode resolver os problemas associados a documentos marcados com hipertexto em que para um processamento mais eficiente poderá ser necessário eliminar as marcações;

- **Eliminação de maiúsculas/ minúsculas (*Case Folding*)**:
  Consiste em converter todas as letras das palavras para o mesmo formato (maiúscula ou minúscula), pode ser entendido com um processo de normalização do texto, permitindo

comparações mais eficazes. Assim a palavra "Sumarização" seria convertida para "sumarização" permitindo uma comparação positiva entre ambas;

- **Eliminação de palavras irrelevantes (*Stopwords*)**:
  Esta fase tenta eliminar as palavras mais comuns, ou seja, aquelas consideradas irrelevantes em termos semânticos. Este tipo de palavras pode ser identificado através do cálculo da sua relevância atribuindo um valor mínimo para poder ser considerada uma palavra relevante, ou através de listas pré-compiladas de palavras consideradas insignificantes. Estas listas podem obter-se facilmente na *Web*;

- **Redução das palavras (stemming ou lematização)**:
  Consiste da redução da palavra tentando obter uma forma mais curta, da mesma. Existem dois métodos para efectuar essa redução, um deles é o *stemming* que procede à redução da palavra retirando o seu sufixo, podendo em alguns casos obter-se uma palavra inexistente no dicionário do idioma do texto original, a outra forma, nomeada de lematização, é mais sofisticada e tem como objectivo a redução das palavras à sua forma canónica, tendo em conta que estas existam e estejam relaccionadas com as originais, eliminando assim, os problemas do *stemming*. Um exemplo de *stemming* é a palavra "processando" ser reduzida para "processa" enquanto que um lematizador mais sofisticado conseguiria reduzi-la para "processar" conseguindo facilitar uma comparação entre formas conjugadas do mesmo verbo.

Como referido na secção 3.1 deste mesmo capítulo, o sistema implementado faz uso de ferramentas presentes na biblioteca *Hultig* [fHLTB] para proceder à análise sintáctica do texto e conversão de maiúsculas para minúsculas, a remoção de palavras irrelevantes utiliza os dois métodos referidos anteriormente dependendo da ferramenta usada (Processamento de *corpus*, criação de novo perfil de utilizador ou sumarização), os processos de *stemming*, tanto para o idioma inglês como português são feitos por implementações , em liguagem *JAVA*, do *Porter Stemmer* [Por80] e *PTStemmer* [pts].

3.4 Ferramentas de Auxílio Desenvolvidas

Para auxiliar o processo de sumarização automática são necessárias algumas ferramentas d pré-processamento que contêm informação sobre as frequências das palavras num determinado domínio ou em contexto geral, as classificações a nível da polaridade ou objectividade dos termos que possam estar presentes num texto. De uma forma geral, essa informação é obtida com base num *corpus* de texto de grandes dimensões e pode ser utilizada para diversos fins, como por exemplo o cálculo da relevância e objectividade das palavras ou identificação dos termos mais relevantes num dado tema, para se definir um tipo de utilizador. Para essas tarefas foram desenvolvidas duas ferramentas diferentes, uma para processar um corpus e assim obter as frequências das palavras, outra para processar o conjunto de textos referentes a um domínio, de forma a extrair um conjunto de palavras que possa o definir.
Foi também desenvolvido um método para processar dados do *SentiWordNet*, para mais tarde serem utilizados no cálculo da objectividade. São armazenados os valores dos campos (palavra, polaridade positiva e negativa) presentes num ficheiro de texto fornecido pelos criador do *SentiWordNet*, actualmente na versão 3.0, através do seu site oficial[1]. Este ficheiro, como indicado em 3.5.2, contém informação das palavras a nível da sua polaridade (positiva e negativa)

[1] site oficial do *SentiWordNet*: `http://sentiwordnet.isti.cnr.it`

e alguns exemplos de uso. O processamento do ficheiro do *SentiWordNet* é rápido para o idioma inglês, visto ser o idioma nativo do mesmo. Para português o processo é mais demorado sendo necessário ter ligação à *Internet* de forma a proceder à tradução automática das palavras, de inglês para português. Finalmente, a estrutura gerada é armazenada na localização indicada pelo utilizador de modo a que possa ser utilizada mais tarde. Sendo necessário indicar a localização do ficheiro de entrada e onde será armazenado o resultado, esta ferramenta não executará, o processo de conversão do ficheiro de texto para a estrutura pretendida, sem que estas sejam indicadas.

3.4.1 Ferramenta de Processamento de *Corpus*

Esta ferramenta permite a análise de um conjunto de ficheiros de texto simples, no sentido de produzir uma estrutura de frequências das palavras contidas no *corpus* indicado. Para que esta informação seja significativa é necessário o uso de um conjunto de ficheiros de grandes dimensões.
A estrutura que armazena esta informação deriva de uma presente na linguagem *JAVA*, denominada *Hashtable* e permite o mapeamento dos termos e respectivas frequências. Outra característica desta estrutura é o cálculo da probabilidade dos termos nela presentes, devido ao registo do total dos termos que ela tem.
Visto que este processo de análise se pode tornar bastante moroso, foi desenvolvida uma ferramenta que permita a criação e armazenamento dessa estrutura no computador, para poder ser utilizada mais tarde. Também é possível utilizar um registo de frequência já guardado, como ponto de partida, de modo a completá-lo. Se a opção de Adicionar a Existente, não estiver seleccionada, o botão e caixa de texto referentes ao *Corpus* já existente é desabilitado, indicando assim ao utilizador que não é necessário indicar a localização do mesmo.
Sempre que o utilizador tente iniciar o processamento de um *corpus* e as localizações dos ficheiros necessários ou do local de armazenamento do resultado não sejam indicadas, estas serão pedidas ao utilizador.

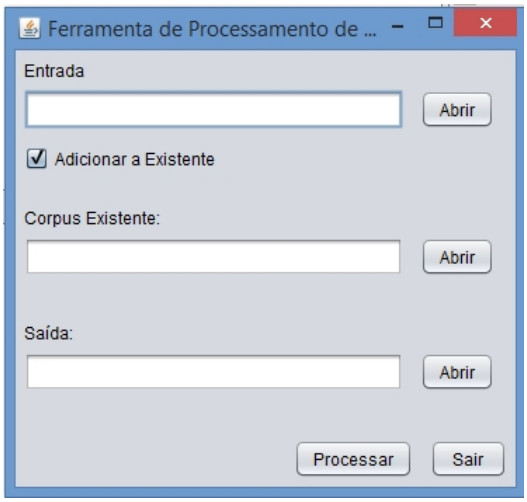

Figura 3.2: Ferramenta de Processamento de *Corpus*

3.4.2 Ferramenta de Criação de Novo Perfil

Esta ferramenta tem como objectivo ajudar na criação de um novo perfil de utilizador. Para tal, é necessário especificar a localização do corpus de ficheiros de texto sobre o domínio pretendido, o local e nome pretendidos para armazenar o ficheiro e indicar o idioma desse conjunto de palavras. Se o utilizador não indicar um ou mais elementos necessários, serão pedidos, pela ferramenta, os elementos em falta. O modo de operar é muito idêntico ao da ferramenta anterior, neste caso é carregada a informação da estrutura que contem a frequência das palavras no contexto geral, obtida pela ferramenta da subsecção 3.4.1, analisados os documentos presentes na localização indicada, para obter uma estrutura idêntica que represente os termos e as suas frequências, pertencentes a um domínio mais específico.De seguida são retiradas as palavras irrelevantes com auxílio a uma lista de *stop words* no idioma indicado pelo utilizador. Com base nas duas estruturas são calculadas as relevâncias das palavras, tal como referido em 3.5.1. Por fim, um conjunto com as 50 palavras mais relevantes, e que representam o tema processado, são guardadas numa *Hashtable* para permitir uma pesquisa eficaz e essa estrutura é armazenada na localização indicada pelo utilizador.

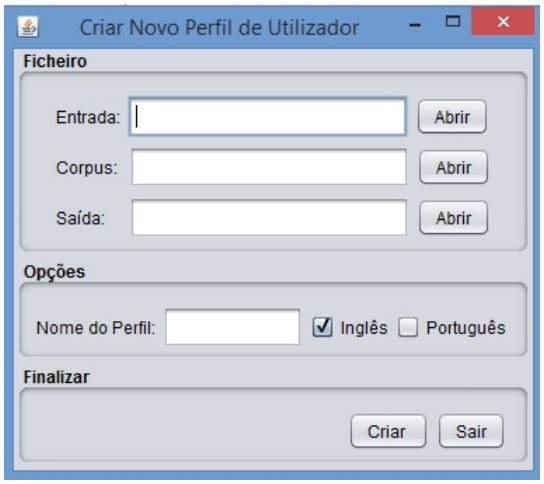

Figura 3.3: Ferramenta de Criação de Novo Perfil de Utilizador

3.5 Métodos de Extracção Implementados

As palavras podem ser classificadas tendo em conta algumas das características, como por exemplo a sua sintaxe, através de marcadores morfossintáticos, a informatividade das mesmas, através da sua relevância e o grau de emotividade que pode ser obtido através de conjuntos de dados pré-processados com informação sobre a objectividade / emotividade dos termos.

3.5.1 Relevância da palavra

A relevância da palavra pode ser calculada tendo em conta o número de vezes que esta aparece no texto. Embora a frequência de uma palavra seja útil, não revela o nível de informatividade de uma palavra, pois ao longo de um texto é possível encontrar um conjunto de termos que não transmite qualquer tipo de informação visto que podem estar presentes nos mais variados assuntos descritos num texto. Estas palavras são chamadas de *stopwords*. Lista com palavras deste género podem ser encontradas em [sto], contendo 571 palavras para o idioma inglês e 356 para português.

Existem algumas técnicas estudadas neste domínio, tais como `tf.idf`[SYY75], que avalia a importância de uma palavra num documento baseado na frequência e distribuição da mesma num conjunto de documentos, `tf.isf`[DAL07], o qual avalia cada palavra em termos de distribuição no mesmo documento. No caso do `tf.idf`[SYY75] a equação que indica o valor de uma palavra é dada por:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \quad (3.1)$$

Supondo que temos uma colecção de N documentos. f_{ij} representa a frequência (número de ocorrências) do termo i no documento j. Essa frequência pode ser normalizada se a dividirmos pelo máximo número de ocorrências de qualquer termo (excluindo *stop words*). Assim o valor da métrica, representado por TF_{ij}, estará dentro do intervalo $[0, 1]$.
É importante a referência à métrica anterior pois neste trabalho é implementada uma abordagem semelhante mas que utiliza as frequências gerais das palavras, indicada na subsecção 3.4.1, não tendo por isso referência a um conjunto de documentos. Posto isto, a relevância é calculada da seguinte forma:

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

Em que, 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, influenciada pela relação entre a probabilidade ($P(p|D)$) de p estar em D e a probabilidade ($P(p)$) de p obtida através de um *corpus*. O seu valor (Equação 3.2) é normalizado para o intervalo $[0, 1]$, tendo em conta a relevância máxima de uma palavra presente no texto, ou seja, a relevância final é calculada da seguinte maneira (Equação 3.3):

$$Re(p|D) = \frac{Re_p(p|D)}{max(Re(p|D))} \quad (3.3)$$

O objectivo desta métrica é evidenciar as palavras contidas num texto ou documento, que sejam mais relevantes no contexto do mesmo do que num geral. A utilização deste método é possível devido à pré-compilação de estruturas onde são armazenadas as palavras e as suas frequências.

Idioma	Número de Termos Únicos	Número Total de Termos Analisados
Português (CTEMPúblico)	494 519	84 208 898
Português (TeMario)	8 965	151 952
Inglês	286 036	229 568 854

Tabela 3.1: Tamanho do *Corpora* explorado.

Após o cálculo da relevância das palavras é necessário proceder ao somatório das classificações de todos os termos da frase para se decidir o quão relevante ela é. Efectua-se uma normalização dos valores de relevância de cada palavra para que, assim, todos os termos do texto se apresentem no intervalo $[0, 1]$. A equação que representa o somatório é:

$$Re(F_i) = \frac{1}{n_p} \times \sum_{j=1}^{n_p} \frac{Re(p_j)}{max(Re(p_{ji}))} \quad (3.4)$$

Onde, a classificação ($Re(F_i)$) da frase i é calculada tendo em consideração a relevância normalizada ($\frac{Re(p_j)}{max(Re(p_{ji})}$) da palavra ($p_j$), em que a relevância máxima de uma palavra da frase (F_i) é representada por $max(Re(p_{ji})$.
Independentemente, da identificação dos tópicos presentes no texto (secção 3.5.3) estas classificações são ordenadas de modo a que possam ser extraídas as frases com maior relevância. O conjunto das frases extraídas pode assim depender do número de tópicos presentes.
Supondo o conjunto de três frases, classificadas com relevância 0.50 (1), 0.23 (2) e 0.61(3), apresentadas de seguida:

1. Finanças apontaram para uma "deterioração de 389 milhões de euros" do défice. ($Re = 0.50$)

2. Um mês depois a situação inverteu-se. ($Re = 0.23$)

3. Défice público volátil melhorou para 769 milhões de euros. ($Re = 0.61$)

Após a ordenação decrescente das mesmas, quanto à sua relevância, obtemos o conjunto $[3, 1, 2]$. Durante o processo de extracção, se a compressão a ser usada for de 30%, ou seja, um terço, apenas seria extraída a frase número 3, no caso de a taxa de compressão permitir a extracção de duas frases iriam ser extraídas as frases 3 e 1, e posteriormente ordenadas pela sua posição original no texto, obtendo assim um sumário com o conjunto $[1, 3]$.

3.5.2 Objectividade da Palavra

A objectividade de uma palavra está relacionada com a sua emotividade [BES10], ou seja, esta pode ser classificada como negativa ou positiva em determinadas situações ao longo de um texto. Essa classificação é encontrada em diversas ferramentas já existentes, algumas feitas manualmente outras de forma automática, como o `SentiWordNet`[BES10]. Esta ferramenta é derivada de uma outra chamada `WordNet` [Fel98], que consiste num conjunto de dados léxicos, em língua inglesa, agrupados em conjunto com os seus sinónimos, expressando cada um deles o seu significado distinto. Em relação ao `SentiWordNet`, este engloba também um conjunto de notações para a classificação de uma palavra como sendo positiva, negativa ou neutra. A classificação é expressa em valor decimal no intervalo $[0, 1]$, sendo que a soma das componentes positiva e negativa também está contida no mesmo intervalo. Segundo os seus autores, a objectividade pode ser calculada através da seguinte equação:

$$Obj_i = 1 - (pos + neg) \quad \textbf{(3.5)}$$

Em que Obj_i se refere ao valor da objectividade da palavra i, pos e neg são respectivamente os valores das componentes positiva e negativa da polaridade para uma palavra. Segundo os autores, a objectividade de uma palavras está associada à polaridade neutra, isto é, uma palavra é mais objectiva quanto menor for a soma das suas polaridades, positiva e negativa. A polaridade transmite assim opinião, sentimento ou subjectividade. Portanto, um termo que não seja associado a uma opinião terá objectividade máxima, sendo a sua subjectividade inversamente proporcional à sua objectividade.
Uma palavra (i) com uma determinada objectividade (Obj_i) é integrante de uma frase (F_i) que vê o cálculo da objectividade ($Obj(F_i)$) conseguido através da soma das objectividades

normalizadas ($\frac{Obj(p_j)}{max(Obj(p_{ji})}$) dos termos que a compõem, tal como representado na seguinte equação:

$$Obj(F_i) = \frac{1}{n_p} \times \sum_{j=1}^{n_p} \frac{Obj(p_j)}{max(Obj(p_{ji}))} \quad \textbf{(3.6)}$$

Os valores obtidos com este método são utilizados para a classificação das frases, $Obj(F_i) \in [0, 1]$, podendo ter em conta os tópicos do texto identificados como indicado em 3.5.3. De seguida os valores podem ou não ser combinados com outras técnicas, sendo extraídas as melhores frases após uma ordenação decrescente dos valores obtidos por cada uma. O conjunto das frases extraídas pode assim depender do número de tópicos presentes.

3.5.3 Segmentação em Tópicos

Um texto contém diferentes tipos de informação, podendo por vezes introduzir temas novos. A segmentação de texto em tópicos tem como objectivo lidar com esta característica textual. Ao longo do tempo foram efectuados alguns estudos com vista a esta divisão. Entre eles podem encontrar-se ferramentas, tais como, o *MorphAdorner*[Bur13], *TextTling*[Hea97], *GistSumm*[PRdGVN03], já referidas na secção 3.5.

No presente trabalho a ferramenta usada para a subdivisão do texto em tópicos foi o *MorphAdorner*[Bur13]. Com esta funcionalidade é possível extrair frases que representem uma maior abrangência do conteúdo textual analisado.

De um modo geral, esta ferramenta após uma separação das unidades textuais (*parsing*) identifica um novo tópico através de uma mudança significativa dos termos presentes numa frases.

Inicialmente, é necessária a segmentação do texto, em listas de frases (cada uma contendo uma lista de palavras), que pode ser obtida através do segmentador C99 [Cho00] contido no *MorphAdorner*. Essa segmentação ficou a cargo da biblioteca *Hultig* [fHLTB], para uma maior compatibilidade entre o resultado gerado por ambos.

A identificação dos tópicos é indicada através de uma lista onde estão armazenadas as posições das frases que iniciam um tópico novo.

Após determinar os tópicos do texto as suas frases são classificadas quanto à sua relevância e objectividade, utilizando para isso os métodos indicados em 3.5.1 e 3.5.2, respectivamente. De seguida são ordenadas e extraídas as melhores frases de cada tópico tendo em conta um número máximo de frases a extrair, calculado com base na taxa de compressão.

Supondo um texto com seis frases $Fr_i = \{A, B, C, D, E, F\}$, com classificação final $C(Fr_i) = \{0.6, 0.3, 0.2, 0.7, 0.5, 0.1\}$ e com três tópicos $T_1 = \{A, B\}$, $T_2 = \{C, D, E\}$ e $T_3 = \{F\}$, após a ordenação das frases em cada tópico, obtêm-se os seguintes: $T_1 = \{A, B\}$, $T_2 = \{D, E, C\}$ e $T_3 = \{F\}$. Tendo em conta uma taxa de compressão de 50%, seriam extraídas quatro quatro ($Sum = \{A, D, E, F\}$), pois a taxa utilizada é empregue em cada uma dos tópicos separadamente, assim cada tópico é reduzido para metade do tamanho, em número de frases, caso esse tamanho seja um valor real (por exemplo 0.6) será extraída uma frase a mais desse tópico. Sendo assim o tópico T_2 terá duas frases extraídas. De seguida, as frases extraídas ($Sum = \{A, E, D, F\}$) são ordenadas, quanto à sua posição no texto original obtendo a seguinte lista: $Sum = \{A, D, E, F\}$.

3.5.4 Identificação do Perfil de Utilizador

Entre as várias caracterísiticas que classificam um texto, o domínio (ou tema) é muito importante, pois fornece uma primeira análise ao leitor sobre o conteúdo do texto. Por esta razão é de grande importância adaptar o processo de sumarização automática àquilo que o leitor gosta. Uma forma de o fazer é construindo um conjunto de palavras que sejam significativas nesse domínio. Assim, pode atribuir-se uma bonificação às frases que contêm algumas dessas palavras, para que estejam mais enquadradas com as preferências do leitor, do sumário. Este conjunto de palavras pode ser identificado como um perfil de utilizador, que por sua vez pode ser obtido através de vários métodos, como um conjunto de palavras que o leitor indica, a análise de palavras significativamente relevantes num conjunto de textos que ele lê ou através de métodos linguísticos que forneçam relações entre um tema, indicado pelo utilizador, e o conteúdo de um texto.
O perfil do utilizador é representado neste trabalho por uma estrutura *Hashtable*, onde são armazenadas palavras (cerca de 50) que definem um determinado domínio. Essas palavras são identificadas através do método de cálculo da relevância, descrito na secção 3.5.1, em que um *corpus* com artigos noticiosos pertencentes a um domínio (por exemplo política) são tratados pelo método como sendo o documento de entrada e o *corpus* geral é aquele que contém as frequências das palavras num contexto geral. Assim o cálculo da relevância é feito com base nas equações 3.2 e 3.4 em que a noção de documento é substituída pelo *corpus* de um domínio específico. Foi escolhido um conjunto de 50 palavras, visto que a obtenção automática desse conjunto , mesmo com textos de apenas um domínio, pode identificar algumas palavras que não representem devidamente o domínio pretendido e embora com o aumento quantitativo desse conjunto o problema seja potencialmente maior, o número de palavras que representam correctamente o domínio tenderá a aumentar, diminuindo assim a possibilidade de uma má classificação, devido às possíveis incorrecções.

O resultado é armazenado num ficheiro para que possa ser utilizado mais tarde sem perder tempo a processá-lo quando necessário.

3.5.5 Método Híbrido

Neste trabalho, foi considerado que uma frase deve ter um tamanho igual ou superior a sete, para poder ser calculada a sua informatividade. A escolha deste valor deve-se ao facto de o *parsing*, divisão do texto em palavras, efectuado pela ferramenta *Hultig* [fHLTB], considerar as palavras, números e sinais de pontuação como partes integrantes para o cálculo do tamanho de uma frase. Assim, a frase "Seguro, recorde-se, demitiu-se do seu cargo.", tem tamanho igual a 9, embora contenha apenas 6 palavras. Pois os sinais de pontuação também integram a frase, quanto ao seu tamanho.
Os métodos referidos anteriormente, Relevância (3.5.1), objectividade (3.5.2) e Identificação do perfil de utilizador (3.5.4) podem ser combinados, seguindo uma abordagem idêntica à de Edmundson [Edm69], obtendo assim uma classificação para cada frase, baseada em 3 características distintas. Tendo em conta o tamanho de uma frase, podemos considerá-la pequena e com pouca informatividade, quando esta tem um número reduzido de termos.
A combinação das pontuações é mediada por pesos associados a cada uma das componentes (relevância e objectividade) e é dada pela seguinte equação:

$$Re.Obj(F_i) = \alpha \times Re(F_i) + \beta \times Obj(F_i)\ ,\ \alpha + \beta = 1 \quad (3.7)$$

Onde, $Re.Obj(F_i)$ representa a combinação entre a relevância ($Re(F_i)$) e objectividade ($Obj(F_i)$) de uma frase, mediada pelos pesos atribuídos a cada componente, respectivamente α e β. O valor obtido por esta medida é usado na combinação de todas as técnicas

$$C(F_i) = \phi \times Re.Obj(F_i) + \psi \times L(F_i) + \gamma \times T(F_i)\ ,\ \phi + \psi + \gamma = 1 \quad (3.8)$$

Assim a classificação final ($C(F_i)$) de uma frase i, é determinada pela soma de cada uma das suas componentes de avaliação, a combinação entre Relevância ($Re(F_i)$), objectividade ($Obj(F_i)$), representada por $Re.Obj(F_i)$, a localização ($L(F_i)$) e número de palavras do perfil de utilizador presentes na frase i ($T(F_i)$), mediadas pelos seus pesos, ϕ , ψ e γ, respectivamente.

A pontuação da heurística posicional $L(F_i)$ é calculada em função da localização (i) da frase (F_i) no texto fonte, o que permite estudar a estrutura textual do texto analisado. Neste trabalho são utilizados textos noticiosos para gerar sumários. Segundo, Nenkova [Nen05], este tipo de textos contém informação mais relevante nas suas frases iniciais. Neste sentido, em [Pat07], foi sugerido o uso de uma heurística posicional, dada pela equação 3.9:

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

Esta, é uma das formas que pode ser utilizada para o cálculo da heurística posicional. Também foi explorada um metodologia que se baseia na Distribuição de *Poisson*. Neste caso era pretendido atribuir maior pontuação às frases das extremidades do texto, ou seja, as frases em posições iniciais e finais, teriam maior pontuação que as do meio. Esta pontuação era obtida com a equação: $L(F_i) = 1 - P(x)$, em que $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, onde x representa a localização da frase e λ o tamanho do texto original, em frases. Esta abordagem seria útil caso o texto analisado fosse um artigo científico, onde os conteúdos mais importantes estariam posicionados no início e no fim do documento, assim uma heurística que se foque nesse posicionamento, é mais indicada.

A classificação final ($C(F_i)$) é ordenada de forma decrescente e pode ainda combinada com a identificação de tópicos no texto (3.5.3). A diferença entre o uso e não uso dos tópicos está relacionada com a posição das frases extraídas. Assim, quando não são identificados os tópicos, as frases extraídas serão as que obtiveram melhor classificação no texto, caso contrário são extraídas as melhores frases de cada tópico, tendo em conta a taxa de compressão pretendida para o sumário, para calcular o número de frases extraídas em cada tópico. Neste último caso, podem ser extraídas frases em diferentes quantidades dependendo do tamanho total do tópico. Por exemplo, num texto com dez frases e três tópicos, se um deles (tópico 2) for constituído por seis frases e cada um dos outros por três, com uma taxa de compressão de 30%, o tópico 2 terá duas frases extraídas dele, enquanto cada um dos outros contribuirá com apenas uma frase para o sumário final.

3.6 Aplicação Final

Para que o utilizador possa usar o sistema com maior facilidade foi desenvolvida uma interface de interacção com o utilizador. Esta é ilustrada pela figura 3.4

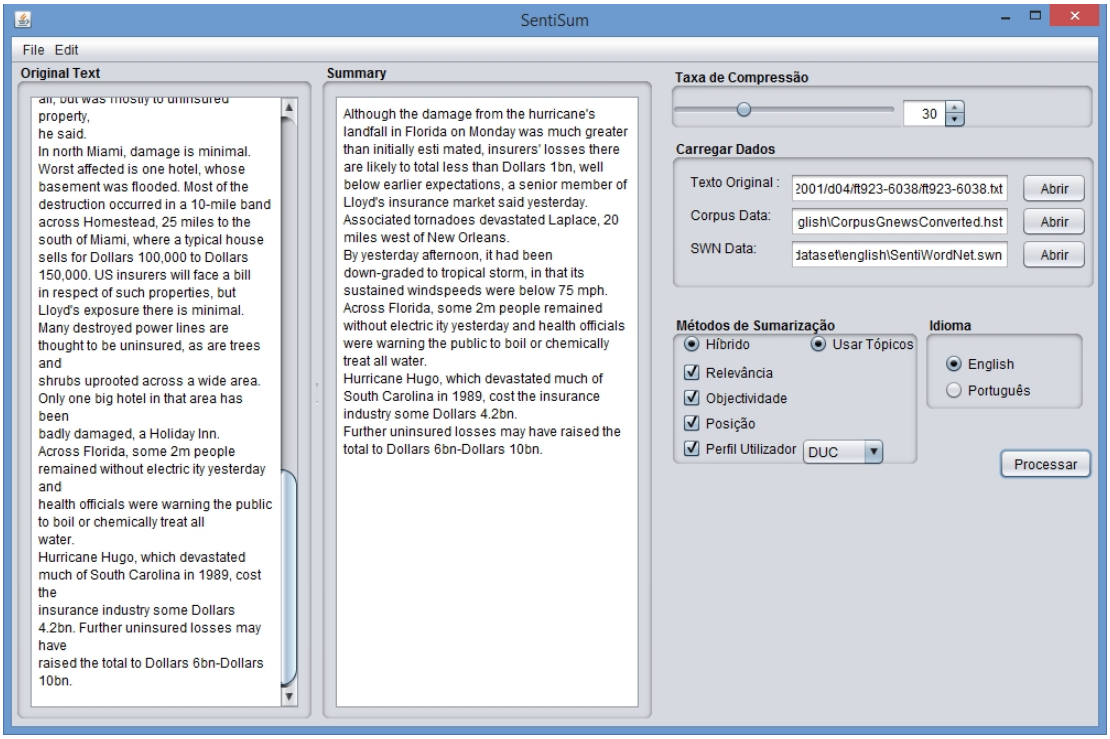

Figura 3.4: Interface da aplicação de Sumarização Automática

Funcionamento Geral

Antes de iniciar o processo de sumarização automática, o utilizador deve indicar o texto ou documento a ser sumarizado, e a localização do *corpus* de domínio geral e das classificações das palavras do *SentiWordNet*, a taxa de compressão e o idioma a usar. Para iniciar o processo de sumarização é necessário efectuar os passos anteriores, por isso, caso algum ainda não tenha sido realizado a aplicação sugere ao utilizador, que o faça. Após se reunirem as condições necessárias, a aplicação gera o sumário do texto indicado utilizando o método híbrido (secção 3.5.5). Podem ser escolhidos, um ou vários métodos de sumarização implementados, relevância, objectividade , localização, perfil de utilizador ou híbrido (agrupa os restantes quatro). O utilizador poderá ainda escolher se prefere que o sumário seja gerado tendo em conta os tópicos do texto.

Exemplo de Utilização

Na figura 3.4, pode verificar-se um exemplo de utilização, com um ficheiro de texto incluído no *DUC 2001*, em que se encontram expostos o texto original e o seu sumário. Na mesma figura são ainda observáveis todas as escolhas feitas para gerar o sumário, como o idioma, métodos utilizados, perfil de utilizador e a taxa de compressão, correspondente a 30% do tamanho, em frases, do texto original.
A figura 3.5, ilustra um exemplo de utilização do sistema, apresentando o conteúdo do ficheiro *ft923-6038.txt*, incluído na versão de 2001 do corpus de documentos *DUC*, como o texto original

e o sumário (Figura 3.6) produzido pelo sistema, aplicando o método híbrido, ou seja, utilizando os cálculos da relevância e objectividade das palavras, posição das frases no texto e um perfil de utilizador, neste caso contendo o conjunto de termos mais relevantes presentes em todo o conjunto de dados *DUC 2001* para uma taxa de compressão correspndente a 30% do texto original.
Pode ainda perceber-se, através deste exemplo (Figura 3.6) que as frases extraídas apresentam um bom grau de informatividade em relação ao texto original, embora nalguns casos sofram de problemas de incoerência, devido à presença de termos temporais.

Texto Original:

[1] HURRICANE Andrew, claimed to be the costliest natural disaster in US history, yesterday smashed its way through the state of Louisiana, inflicting severe damage on rural communities but narrowly missing the low-lying city of New Orleans. [2] The storm, which brought havoc to southern Florida on Monday and then headed north-west across the Gulf of Mexico, had made landfall late on Tuesday night some 60 miles south-west of the city in the agricultural Cajun country. [3] Although the damage from the hurricane's landfall in Florida on Monday was much greater than initially esti mated, insurers' losses there are likely to total less than Dollars 1bn, well below earlier expectations, a senior member of Lloyd's insurance market said yesterday. [4] In Louisiana, the hurricane landed with wind speeds of about 120 miles per hour and caused severe damage in small coastal centres such as Morgan City, Franklin and New Iberia. [5] Associated tornadoes devastated Laplace, 20 miles west of New Orleans. [6] Then, however, Andrew lost force as it moved north over land. [7] By yesterday afternoon, it had been down-graded to tropical storm, in that its sustained windspeeds were below 75 mph. [8] Initial reports said at least one person had died, 75 been injured and thousands made homeless along the Louisiana coast, after 14 confirmed deaths in Florida and three in the Bahamas. [9] The storm caused little damage to Louisiana's important oil-refining industry, although some plants had to halt production when electricity was cut. [10] The Lloyd's member, in close contact with leading insurers in Florida, said that damage to insured property was remarkably small. [11] More than Dollars 15bn of damage may have been caused in all, but was mostly to uninsured property, he said. [12] In north Miami, damage is minimal. [13] Worst affected is one hotel, whose basement was flooded. [14] Most of the destruction occurred in a 10-mile band across Homestead, 25 miles to the south of Miami, where a typical house sells for Dollars 100,000 to Dollars 150,000. [15] US insurers will face a bill in respect of such properties, but Lloyd's exposure there is minimal. [16] Many destroyed power lines are thought to be uninsured, as are trees and shrubs uprooted across a wide area. [17] Only one big hotel in that area has been badly damaged, a Holiday Inn. [18] Across Florida, some 2m people remained without electric ity yesterday and health officials were warning the public to boil or chemically treat all water. [19] Hurricane Hugo, which devastated much of South Carolina in 1989, cost the insurance industry some Dollars 4.2bn. [20] Further uninsured losses may have raised the total to Dollars 6bn-Dollars 10bn.

Figura 3.5: Exemplo de um texto lido e apresentado pelo sistema

Sumário Automático (taxa = 30%):

[3] Although the damage from the hurricane's landfall in Florida on Monday was much greater than initially esti mated, insurers' losses there are likely to total less than Dollars 1bn, well below earlier expectations, a senior member of Lloyd's insurance market said yesterday. [5] Associated tornadoes devastated Laplace, 20 miles west of New Orleans. [7] By yesterday afternoon, it had been downgraded to tropical storm, in that its sustained windspeeds were below 75 mph. [18] Across Florida, some 2m people remained without electric ity yesterday and health officials were warning the public to boil or chemically treat all water. [19] Hurricane Hugo, which devastated much of South Carolina in 1989, cost the insurance industry some Dollars 4.2bn. [20] Further uninsured losses may have raised the total to Dollars 6bn-Dollars 10bn.

Figura 3.6: Exemplo de um sumário produzido pelo sistema

Capítulo 4

Avaliação

Uma questão muito importante e também com um grau de complexidade enorme na sumarização automática de texto é a avaliação do sumário produzido. Esta pode lidar com a qualidade dos sumários gerados pelo sistema e com o seu grau de eficiência durante o processo de sumarização. A avaliação da qualidade de um sumário tem provado ser bastante complexa, muito por causa dos problemas que afectam a construção de um sumário, tais como a público a que se destina, conhecimento do domínio do texto original, por parte de quem o sumariza, entre outros factores. Portanto, uma avaliação automática pode ajudar bastante nesta tarefa, fazendo com que os métodos de avaliação sejam os mesmos, sem a intervenção de um avaliador humano.
Muitos estudos foram publicados mas ainda não se conseguiu chegar a um consenso em relação às caracterísiticas que um sumário deverá ter para ser considerado "ideal", mesmo quando se impõem algumas restrições. Assim, o grau de concordância para criar e consequentemente avaliar um sumário é bastante baixo [Nen06]. Dragomir et al., no estudo que realizaram chegaram à mesma conclusão, indicando um valor de cerca de 60% de concordância entre sumários produzidos por humanos[RHM02].
Por esta razão, a dificuldade na avaliação de sumários deve-se essencialmente à subjectividade e falta de concordância entre os avaliadores humanos. A subjectividade desse processo está relacionada com características dos sumários como a legibilidade, abrangência, domínio do texto, qual a taxa de compressão ideal para a avaliação adequada dos seus sumários, como se deveria avaliar automaticamente a qualidade e o nível de informação presente nos sumários, como identificar um bom juiz humano, ou seja, que tipo de julgamento humano deve ser usado, qual o perfil que este deve ter, dificuldade em encontrar um número de especialistas em técnicas de avaliação suficiente para avaliar os sumários, o tempo e a complexidade necessárias para uma boa avaliação, e a grande subjectividade existente no julgamento humano, o que torna difícil chegar a uma conclusão.
Em [RP03], os autores elaboraram um estudo sobre as principais caracterísiticas e metodologias para a sumarização automática de textos, apresentando várias formas de avaliação de um sistema deste género, onde constam, a possibilidade de medir o grau de utilidade, a adequação a certas tarefas e a validade da metodologia utilizada, entre outras. Este processo também pode passar por uma avaliação do sistema quanto ao seu desempenho, à sua usabilidade, ao grau de satisfação em relação aos resultados produzidos. Quanto ao desempenho do sistema, tem-se em conta, por exemplo, o uso da memória do computador, tempo de execução, a complexidade do algoritmo principal. Em relação à usabilidade do sistema é criticada a clareza da interface ou o grau de intuição necessário para o seu uso ou então, a consistência e flexibilidade no que toca a possíveis configurações [Man99]. Na avaliação dos resultados produzidos pelo sistema, pretende-se verificar se os resultados são os esperados, correctos ou adequados. Num trabalho elaborado por Mani et al., em [Man99], foram estudadas as dificuldades da sumarização automática, em que destacaram os seguintes desafios:

- **Identificação do que seria um resultado correcto para um sumarizador automático**: Um sumário depende de muitos factores, como o público alvo, o grau de instrução ou preferências do sumarizador, etc., por isso para um texto fonte podem existir vários sumários

diferentes;

- **Identificação de uma taxa de compressão ideal**:
  Uma taxa de compressão alta tende a diminuir o conteúdo informativo do sumário e vice-versa. Além disso, consoante o seu nível de conhecimento do sumarizador, poderão ser incluídas diferentes quantidades de conteúdo informativo, mesmo para uma taxa de compressão igual. Isto leva a que não seja fácil encontrar o valor exacto para essa taxa;

- **A forma como a qualidade e a informatividade podem ser avaliadas automaticamente**:
  A falta de um avaliador automático que consiga substituir adequadamente um avaliador humano, quando se tentam avaliar aspectos como a qualidade e a informatividade de um sumário leva a que este processo seja bastante demorado, pois é necessário que os avaliadores humanos comparem o texto original com os respectivos sumários;

- **Identificação da situação e da forma como se utilizar o avaliador humano**:
  A avaliação feita por humanos tem desvantagens, como o tempo de avaliação, a escolha do perfil do avaliador e o custo envolvido. è necessário um número significativo de avalizadores humanos, para que o processo seja mais rápido, além disso uma avaliação robusta e abrangente torna esse processo mais lento e complexo, e o alto grau de subjectividade no julgamento humano também constitui um problema neste tipo de avaliação.

Embora existam muitos métodos de avaliação, os mais usados visam avaliar os resultados do sistema de sumarização automática, isto, porque não existe uma robustez suficientemente boa nos métodos que avaliam o desempenho ou a interface do sistema.
Spärck Jones e Galliers [JG96] sugerem que uma avaliação de sistemas de sumarização automática deva ser classificada como *intrínseca* ou *extrínseca*. Quanto à forma da avaliação *intrínseca* acontece quando é avaliado o desempenho do sistema através da verificação da qualidade e informação presente nos sumários produzidos, enquanto que no caso da avaliação *extrínseca*, é avaliada a adequação do sistema quando este é usado em tarefas específicas fora do âmbito da sumarização automática. Continuando à procura de directrizes para a classificação de sistemas de avaliação, os autores propõem ainda variações de classificação quanto ao tipo de julgamento usado, sendo *on-line* ou *off-line*, quanto ao tipo de resultados avaliados, podendo ser *black-box* ou *glass-box* e quanto à forma de comparação, sendo esta, *comparativa* ou *autónoma*. Uma avaliação diz-se *on-line* quando é utilizado o julgamento humano, no caso contrário, denomina-se *off-line*. No que toca à avaliação dos resultados obtidos, no caso de esses resultados serem obtidos no final da execução do sistema, denomina-se avaliação *black-box*, no caso de serem avaliados resultados intermédios, resultantes da execução de cada processo intermédio do sistema, chama-se avaliação *glass-box*. Em relação à comparação feita nas avaliações, podem-se utilizar outras aplicações que tenham a mesma finalidade de modo a compará-las com o sistema, chamando-se, neste caso, de avaliação *comparativa*, em caso contrário denomina-se avaliação *autónoma*. Não é necessário executar todos os tipos de avaliação, esta depende do objectivo pretendido, podendo proceder, apenas, a um tipo de avaliação, por exemplo, a avaliação quanto ao tipo de julgamento usado não depende dos outros tipos.

A avaliação pode ser feita através da apreciação dos sumários por avaliadores humanos, tendo em conta um conjunto de critérios pré-estabelecidos, ou da comparação automática entre um conjunto de sumários produzidos por humanos e um sumário gerado pelo sistema de sumarização automática. Alguns conjuntos de dados desenvolvidos para avaliação incluem mais do que um sumário, por documento, devido à grande subjectividade humana referente a esse processo,

para que assim a classificação seja melhor. O processo de avaliação assenta na procura de conteúdo repetido, palavras, frases, ou expressões, entre o sumário automático e os sumários de referência. Algumas métricas criadas no campo da Recuperação de Informação, também são utilizadas na comparação de sumários, durante o processo de avaliação *intrínseca*. Essas métricas são a **Precisão** (*Precision*), a **Abrangência** ou **Cobertura** (*Recall*) e a **Medida-F** (*F-Measure*) [FC99]. Nesse campo estas medidas eram entendidas como indicadores da relevância de documentos recuperados. No âmbito da sumarização automática, estas medidas podem ser entendidas e expressas da seguinte maneira:

- **Precisão**: medida que tenta indicar a relação entre o número de frases do sumário automático presentes no sumário de referência, varia de 0 a 1, sendo que o valor máximo indica que todas as frases do sumário automático estão incluídas no sumário de referência;

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

- **Abrangência**: medida que tenta indicar a relação entre o número de frases do sumário de referência presentes no sumário automático, o seu valor está contido no intervalo $[0, 1]$, em que o valor máximo se refere à presença no sumário gerado, de todas as frases do sumário de referência;

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

- **Medida-F**: é uma medida que combina os valores da **Precisão** e **Abrangência**, mediante um factor não negativo β, que permite regular a importância entre as duas componentes, pois ambas são complementares e variam, geralmente, de forma inversa. Varia entre 0 e 1.

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

Nas equações anteriores a **Precisão**, **Abrangência**, **Medida-F** de um sumário gerado (S_g) são representados por $P(S_g)$, $R(S_g)$ e $F_\beta(S_g)$, respectivamente.
Jing et al. [JBME98], elaboraram um estudo sobre avaliação de sumários e chegaram à conclusão que diversos factores podem influenciar a qualidade dessa avaliação. Assim, o tamanho do sumário produzido, as métricas de **precisão** e **abrangência**, a dificuldade das perguntas (em avaliações do tipo pergunta-resposta), caracterísiticas dos documentos e a concordância entre os sumários produzidos por humanos são entraves à descoberta de um método ideal para avaliação. No mesmo estudo, referem ainda que as métricas de **precisão** e **abrangência** poderão não ser as mais indicadas para a avaliação da qualidade dos sumários produzidos automaticamente, visto que a troca de uma frase por outra igualmente informativa, mas que não esteja presente no sumário de referência, prejudicará a avaliação. A escolha das unidades textuais consideradas para o cálculo do tamanho do sumário, tendo em conta a taxa de compressão, também é um factor que pode influenciar negativamente a avaliação.
Segundo o trabalho de Mani [Man01], a avaliação é tradicionalmente levada a cabo por sumarizadores humanos em termos da sua coerência, concisão, gramaticalidade e conteúdo. Devido ao custo e demora desse processo de avaliação, o interesse dos investigadores tem aumentado, levando à apresentação de novas metodologias na comparação de sumários automáticos como sumários elaborados por humanos.

Entre os métodos de avaliação de sistemas de sumarização automática, o *Recall-Oriented Understudy for Gisting Evaluation* (*ROUGE*) [yL04, Lin04, LH03] é um dos mais importantes. Este, foi inspirado no sistema *Bilingual Evaluation Understudy* (*BLEU*)[PRWZ02] da *IBM* (*International Business Machines*), que consiste num conjunto de métricas de avaliação de textos traduzidos. Esta ideia foi trabalhada para que pudesse avaliar um sumário produzido automaticamente com um conjunto de sumários elaborados por humanos, consistindo assim, num conjunto de métricas que pretende avaliar o grau de semelhança entre dois textos, neste caso sumários (um gerado automaticamente e outro produzido por um humano), através da procura de repetição de *n-gramas*.
Apesar das vantagens deste sistema de avaliação, como a eliminação da subjectividade humana neste processo, a sua simplicidade e eficiência, estas abordagens ainda têm algumas limitações, são necessários sumários, de referência, produzidos por humanos especializados, e não avaliam características como a coesão, coerência e inteligibilidade dos sumários. Estes problemas afectam todos os métodos de avaliação baseados em abordagens intrínsecas.
Tem existido um grande progresso nesta área, embora ainda o seu potencial de evolução ainda seja enorme e tenham muitas questões pertinentes a resolver. Assim, conferências e seminários, nesta área têm contribuído imenso para o seu avanço, como as conferências *TIPSTER SUMMAC* [MHK+99], que é considerada pioneira no esforço para a evolução da área de avaliação de sistemas de sumarização automática, a *DUC* [duc07], actualmente fazendo parte de uma tarefa na *TAC* [tac14], que consiste na proposta de tarefas específicas a serem resolvidas por sistemas de sumarização automática, comparando os sistemas proposta na resolução destas tarefas.
No sistema de sumarização, desenvolvido no âmbito desta dissertação, a qualidade dos sumários é avaliada de forma isolada, tendo apenas em consideração as suas característica próprias, em vez de tentar resolver uma tarefa específica. A avaliação foi efectuada de forma automática, sem intervenção de avaliadores humanos, ou seja *off-line*, tendo em conta o julgamento, e tendo em consideração apenas o sumário final gerado pelo sistema, sendo assim considerada, também, uma avaliação *black-box*. É feita a comparação entre sumários de referência e um sumário gerado pelo sistema, através da ferramenta *ROUGE*.
Neste capítulo são apresentados três elementos essenciais no processo de avaliação: o conjunto de testes (dados utilizados para avaliação), os métodos de avaliação e os resultados obtidos.

4.1 Conjunto de Testes

Uma avaliação de sistemas de sumarização automática através da comparação entre sumários de referência e um sumário gerado pelo sistema, necessita de um conjunto de dados que contenha o um ou mais sumários para cada texto original, nele presente. Existem vários conjuntos de dados deste género, uns disponibilizados por conferências como a *DUC* [duc07], em inglês, e outros que foram construídos no âmbito de trabalhos fora do âmbito das conferencias, como o conjunto de documentos *TeMário* [PR03] ou *CETEMPublico* [SR01b].
Para a avaliação do sistema desenvolvido no âmbito desta dissertação, foram usados os conjuntos TeMário, de forma parcial, e DUC da edição de 2001, com todos os documentos nele contidos. A escolha destes conjuntos deve-se ao facto de em ambos os casos estão incluídos sumários gerados a partir de documentos únicos. Em relação ao *corpus TeMário* foram escolhidos 100 documentos, de modo a eliminar repetições. No conjunto de documentos da conferência *DUC*, alguns contêm marcações, pelo que foi necessário um pré-processamento dos mesmos, durante o processo de leitura.

Na Tabela 4.1 são apresentadas as configurações de cada um dos conjuntos de dados utilizados para teste.

Tabela 4.1: Informação dos conjuntos de teste.

Corpus	Nº de Textos	Nº de Sumários de Referência	Tamanho médio dos Sumários de Referência (em palavras)
DUC 2001	400	3	119
TeMario	100	1	220

4.2 Método de Avaliação

A avaliação do sistema foi levada a cabo pela ferramenta *ROUGE*, que consiste num conjunto de métricas para avaliar de forma objectiva, sem interferência humana, os sumários gerados por sistemas de sumarização automática. Esta ferramenta é usada em grande escala, com principal incidência nas conferências *DUC* e *TAC*, mas também em projectos de investigação independentes.
O desenvolvedor desta ferramenta, num dos seus estudos [yL04], demonstrou que os níveis de correlação entre juízos humanos e os desta ferramenta são bastante altos, utilizando os conjuntos de dados das edições de 2001 a 2003, das conferências *DUC*. Esta ferramenta consiste num pacote de quatro métricas: *ROUGE-N*, *ROUGE-L*, *ROUGE-W* e *ROUGE-S*, que permitem avaliar os sumários através da co-ocorrência de *n-gramas*, podendo, para isso, alterar o número de *n-gramas* a considerar.
A métrica *ROUGE-N* é uma medida de **Cobertura** ou **Abrangência**, em que as unidades textuais consideradas são *n-gramas*, em vez de frases. O seu valor é calculado através da comparação entre um sumário gerado, pelo sistema de sumarização automática, e um conjunto de sumários de referência, produzidos por humanos. A equação seguinte (Equação 4.4) apresenta o cálculo desta métrica, apresentada em [yL04].

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

Onde, n significa o comprimento do $n-gram$, $gram_n$, e $Count_{match}(gram_n)$ é o máximo número de co-ocorrência de *n-grams* no sumário gerado e o conjunto de sumários (S_r) de referência (R), e $Count(gram_n)$ representa o número total de *n-grams* do conjunto de sumários de referência. Assim, esta métrica avalia a relação entre o número de *n-gramas* do sumário gerado presentes no sumário de referência, e o total de *n-gramas* presentes no sumário de referência. Nesta dissertação é utilizada a métrica *ROUGE-1*, ou seja, considerando *n-gramas* individuais, visto que é a métrica que apresenta melhores níveis de correlação com juízos humanos. Através dessa métrica são calculados os valores de **Precisão**, **Abrangência** e **Medida-F** para os sumários automáticos, em relação aos respectivos sumários de referência. Com os resultados obtidos para um sumário gerado, se existir mais de um sumário de referência são armazenados os valores do melhor, ou seja, daquele que é mais parecido ao sumário gerado. Após o cálculo de todos os valores para o conjunto de testes utilizado, é efectuada uma média, dos mesmos, obtendo assim a **Precisão**, **Abrangência** e **Medida-F** referentes ao conjunto de dados completo.
Esta avaliação permite identificar a qualidade do sumário gerado em termos de informatividade,

ou seja, estas medidas indicam a presença ou ausência de conteúdo no sumário gerado, o qual se espera que esteja presente num bom sumário. Como referido anteriormente (neste capítulo), a coesão, coerência e inteligibilidade dos sumários não são avaliados por estas métricas.

4.3 Resultados

Nesta secção, serão apresentados os resultados obtidos nos testes realizados como base no sistema desenvolvido e métodos de avaliação descritos na secção anterior. Foram utilizados dois conjuntos de textos, *DUC 2001*, para inglês, e *TeMário*, para o idioma português. No conjunto de dados do *DUC* foi necessário proceder a um pré-processamento durante a leitura, visto que alguns documentos apresentam marcações no texto. Foram utilizados todos os documentos e sumários presentes no conjunto, ou seja, existem alguns documentos repetidos, embora para diferentes tarefas, contendo sumários de 50, 100, 200 e 400 palavras. Em relação ao corpus *TeMario* é apenas utilizada a secção sem título, visto que este é dividido em três conjuntos, cada um com 100 documentos, estando separados por título e fonte, só por título, ou não sendo indicado nem título nem fonte. Os testes são realizados para dois dos domínios especificados no conjunto: internacional e opinião.
As unidades textuais extraídas pelo sistema apresentado são frases. O tamanho do sumário gerado pelo sistema foi ajustado, através do cálculo do número de frases necessárias para gerar um sumário idêntico, em número de palavras, aos sumários incluídos em cada um dos conjuntos de dados.
Em relação ao *DUC*,a taxa de compressão ($TC(D)$) para um documento, tem em conta a taxa de compressão ($TC_O(S|D)$) de um sumário (S), em numero de palavras, correspondentes ao mesmo documento (D), sendo o tamanho pretendido indicado no nome do sumário (presente no *DUC*), o número total de frases ($F_t(D)$) e o total de palavras ($P_t(D)$) no documento, sendo esse cálculo efectuado através da seguinte equação (Equação 4.5):

$$TC(D) = TC_O(S|D) * \frac{F_t(D)}{P_t(D)}; \quad \textbf{(4.5)}$$

Sabendo o número aproximado de palavras que o sumário de referência tem ($TC_O(S|D) = 100$), geralmente indicadas no nome do ficheiro de sumário ou na localização do ficheiro, no conjunto *DUC*, o total de palavras ($P_t(D) = 400$) e frases ($F_t(D) = 20$) contidas no documento fonte, a equação 4.5 indica que será necessário extrair 5 frases do texto, para atingir cerca de 100 palavras no sumário gerado.

Quanto ao conjunto de dados *TeMário*,a taxa de compressão é fixada em 30%, pois é o valor indicado pelos autores [PR03], para a extracção de sumários ideais.
Após o cálculo da taxa de compressão, é identificado o tamanho do texto, em frases, e aplicada a taxa a esse tamanho, obtendo assim o número de frases necessárias para atingir o tamanho, em palavras, próximo ao pretendido.
Durante a obtenção de resultados foi testado um método híbrido com diferentes conjuntos de pesos, para as suas componentes. Nas tabelas de resultados o método híbrido é representado por **Híbrido(Rel-Pos-Tem-RelObj)**, em que **Rel** é o peso associado à relevância (o peso da objectividade é calculado da seguinte forma: $Obj = 1 - Rel$, visto que ambas são combinadas na equação 3.7), **Pos** é o peso associado à componente da posição da frase, **Tem** refere-se ao peso da componente temática, determinada pelo perfil de utilizador, e **RelObj** é o peso da relação

entre a relevância e a objectividade (Equação 3.7).
A combinação entre os pesos das componentes referias anteriormente é conseguida da seguinte forma: **[Rel, Pos, Tem, RelObj] = 1.0** .
No processo de avaliação foram investigados vários valores para os pesos das componentes já mencionadas, para se explorarem duas vertentes: testar a combinação mais equilibrada de todas as caracterísiticas consideradas, e tentar identificar a vantagem no uso de métodos que combinem as várias caracterísiticas em relação aos métodos individuais. Todos os testes foram efectuadas duas vezes, tendo em conta a inclusão, ou não da identificação de tópicos no texto. Para um dado texto, inicialmente é calculado o valor da relação entre relevância (equação 3.4) e objectividade (equação 3.6) através da equação 3.7, com os pesos $\alpha = Rel$ e $\beta = 1 - Rel$, pois estes pesos são complementares. De seguida, são calculados os valores da componente posicional (equação 3.9) e da temática (perfil de utilizador, subsecção 3.5.4. Por fim, a classificação final de **Híbrido(Rel-Pos-Tem-RelObj)** é obtida através da equação 3.8 com os pesos **Pos**, **Tem** e **RelObj**, correspondentes às componentes: heurística posicional ($L(F_i)$), perfil de utilizador ($T(F_i)$) e relação entre relevância e objectividade ($Re.Obj(F_i)$).

4.3.1 Resultados com Identificação de Tópicos

Tabela 4.2: Resultados para o conjunto de documentos DUC 2001, Com Tópicos

Rouge-1			
DUC 2001	**Com Tópicos**		
Medidas	**Precisão**	**Abrangência**	**Medida-F**
Relevância	0.36907992	0.43935894	0.37055911
Objectividade	0.39484503	0.45770723	0.39348578
Perfil Utilizador	0.39408951	0.45586064	0.39208257
Posição	*0.41642000*	*0.51170219*	*0.42944004*
Híbrido(0.8-0.2-0.2-0.6)	0.42730006	0.52054142	0.43897766
Híbrido(0.5-0.4-0.2-0.4)	0.42280883	0.51843953	0.43566417
Híbrido(0.5-0.2-0.2-0.6)	0.42344217	0.51592419	0.43478368
Híbrido(0.5-0.0-0.0-1.0)	0.41233833	0.46577286	0.40567427
Híbrido(0.2-0.2-0.2-0.6)	0.41554507	0.50312193	0.42542589
Híbrido(0.8-0.5-0.2-0.3)	**0.42532483**	**0.52320247**	**0.43931816**

Nos resultados da Tabela 4.2, em que são identificados tópicos do texto, pode verificar-se que, de forma isolada (relevância, objectividade, perfil de utilizador e posição), a heurística posicional é a que obtém melhores resultados, de seguida temos o método do perfil de utilizador (conjunto de palavras que definem um domínio), e objectividade com quantidades de informatividade muito idênticas em relação aos sumários de referência, embora a cobertura (ou abrangência) e precisão do sumário seja mais baixa, mesmo assim a sua classificação final é boa, em termos individuais. Para o conjunto de dados utilizado, o número de palavras que definem o perfil pode ser algo reduzido, o que poderá prejudicar o seu resultado final, visto que este conjunto possui textos noticiosos de diversas áreas. Também é possível verificar uma maior significância da objectividade das palavras em relação à relevância.
Quando se combinam as diferentes métricas, o resultado tende a melhorar caso se considere um peso mais elevado para a componente da heurística posicional, obtendo valores muito próximos nos restantes casos, excepto quando os pesos das componentes de relevância e objectividade são iguais e não se consideram outras características.

Tabela 4.3: Resultados para o conjunto de documentos TeMário, com temática Internacional, Com Tópicos

Rouge-1			
TeMário	**Sem Tópicos**		
Medidas	**Precisão**	**Abrangência**	**Medida-F**
Perfil Internacional			
Relevância	0.46402044	0.40797807	0.42443694
Objectividade	0.46256936	0.44178855	0.43874602
Perfil Utilizador	0.47254500	0.43399709	0.43993727
Posição	*0.46969262*	*0.49499224*	*0.47502230*
Híbrido(0.8-0.2-0.2-0.6)	0.47300915	0.47977905	0.47044541
Híbrido(0.5-0.4-0.2-0.4)	0.47323409	0.48382443	0.47203401
Híbrido(0.5-0.2-0.2-0.6)	0.47374018	0.48407476	0.47251935
Híbrido(0.5-0.0-0.0-1.0)	0.46780381	0.44957268	0.44799760
Híbrido(0.2-0.2-0.2-0.6)	0.47218290	0.46893596	0.46319394
Híbrido(0.8-0.5-0.2-0.3)	**0.47538487**	**0.49431762**	**0.47935753**
Perfil de Opinião			
Perfil Utilizador	0.46394838	0.43948730	0.44174750
Híbrido(0.8-0.2-0.2-0.6)	0.46837177	0.47197588	0.46343419
Híbrido(0.5-0.4-0.2-0.4)	0.46233646	0.47355795	0.46068807
Híbrido(0.5-0.2-0.2-0.6)	0.46669791	0.47390187	0.46404963
Híbrido(0.2-0.2-0.2-0.6)	0.46312751	0.47575164	0.46254433
Híbrido(0.8-0.5-0.2-0.3)	**0.47492602**	**0.48822201**	**0.47650188**

Com a utilização do corpus *TeMário* (Tabela 4.3) pretende-se avaliar as variações, nos resultados, de cada método que tenha em conta a componente do perfil de utilizador. Assim, é possível identificar que com o uso de tópicos não há grande variação nos resultados, quando se altera o perfil de utilizador, ou seja, é utilizado um conjunto de palavras obtido para um domínio diferente, daquele presente no texto original. Embora, utilizando o perfil correcto, a precisão melhore ligeiramente e a abrangência desça, a qualidade geral é idêntica. Isto pode revelar, que a tentativa de encontrar as palavras mais relevantes num texto, não é a mais correcta, quando se pretende definir um perfil de utilizador.

4.3.2 Resultados sem Identificação de Tópicos

Tabela 4.4: Resultados para o conjunto de documentos DUC 2001, Sem Tópicos

Rouge-1			
DUC 2001	**Sem Tópicos**		
Medidas	**Precisão**	**Abrangência**	**Medida-F**
Relevância	0.35717528	0.43132657	0.35717528
Objectividade	0.38863613	0.45716874	0.38958276
Perfil Utilizador	0.38266800	0.45542019	0.38407909
Posição	*0.42637936*	*0.56376681*	*0.45727700*
Híbrido(0.8-0.2-0.2-0.6)	0.41862675	0.55606815	0.44874283
Híbrido(0.5-0.4-0.2-0.4)	**0.42612276**	**0.56540744**	**0.45785898**
Híbrido(0.5-0.2-0.2-0.6)	0.42469430	0.56193978	0.45519158
Híbrido(0.5-0.0-0.0-1.0)	0.41686562	0.47154959	0.41035092
Híbrido(0.2-0.2-0.2-0.6)	0.42418506	0.54787219	0.44952790
Híbrido(0.8-0.5-0.2-0.3)	0.42280094	0.55220319	0.4504053

Nas tabelas 4.4 e 4.5, são apresentados os resultados da execução do algoritmo de sumarização automática, sem identificar os tópicos do texto original. Quando utilizado o conjunto de dados do *DUC 2001* (Tabela 4.4), é verificável que a posição é o factor mais importante na extracção

das frases. Em relação ao método híbrido, a característica referida continua a desempenhar um papel importante, obtendo os melhores resultados quando é utilizado um peso igual ao da relação entre relevância e objectividade.
Avaliando nas mesmas condições, os textos do conjunto de dados *TeMário* (Tabela 4.5), verifica-se uma vantagem , embora ligeira, no uso de um conjunto de palavras que representem um domínio correcto, ou seja, usando um perfil de utilizador referente ao tema do texto analisado. Nestas condições, sem usar identificação de tópicos, verifica-se, de uma forma subtil, uma melhoria no método híbrido, para quase todos os conjuntos de pesos, obtendo melhores resultados quando a objectividade e relevância têm o mesmo valor e o peso da sua relação é maior que o das restantes características.

Tabela 4.5: Resultados para o conjunto de documentos TeMário, Sem Tópicos

Rouge-1			
TeMário	**Sem Tópicos**		
Medidas	**Precisão**	**Abrangência**	**Medida-F**
Perfil Internacional			
Relevância	0.44631444	0.42662791	0.42882487
Objectividade	0.45465846	0.47105124	0.45141007
Perfil Utilizador	0.46481147	0.46939635	0.46121589
Posição	*0.45974098*	*0.52097356*	*0.48282553*
Híbrido(0.8-0.2-0.2-0.6)	0.47493482	0.50759973	0.48593305
Híbrido(0.5-0.4-0.2-0.4)	0.46032493	0.51832265	0.48170691
Híbrido(0.5-0.2-0.2-0.6)	**0.47334346**	**0.51449429**	**0.48793322**
Híbrido(0.5-0.0-0.0-1.0)	0.45951981	0.48582803	0.46520060
Híbrido(0.2-0.2-0.2-0.6)	0.46013580	0.52288476	0.48311404
Híbrido(0.8-0.5-0.2-0.3)	0.45581768	0.53365490	0.48693408
Perfil de Opinião			
Perfil Utilizador	0.45353598	0.45489973	0.44544713
Híbrido(0.8-0.2-0.2-0.6)	0.46082988	0.51504472	0.47969900
Híbrido(0.5-0.4-0.2-0.4)	0.45765703	0.52843772	0.48358272
Híbrido(0.5-0.2-0.2-0.6)	0.45668164	0.52164664	0.48042802
Híbrido(0.2-0.2-0.2-0.6)	0.45238274	0.51777915	0.47494558
Híbrido(0.8-0.5-0.2-0.3)	**0.46224721**	**0.52581277**	**0.48583421**

Ao comparar as abordagens em que se identificam os tópicos do texto com aquelas que não têm em consideração os tópicos, as segundas obtêm melhores resultados com quase todos os conjuntos de pesos, podendo concluir que, com a implementação efectuada neste trabalho, os tópicos pioram a qualidade dos sumários. Excepções são os casos em que se utiliza, de forma individual, a relevância, objectividade ou perfil de utilizador, para o idioma inglês. No caso dos documentos em língua portuguesa os resultados melhoram quando não se usa a verificação dos tópicos. Assim, pode concluir-se que o método implementado para a identificação dos tópicos depende do idioma do documento.

4.3.3 Comparação com outras Abordagens

O método explorado no âmbito desta dissertação obteve resultados idênticos a outros trabalhos no idioma inglês, com o conjunto de dados *DUC 2001* [SVB07, Wan10]. Em relação a textos em língua portuguesa, resultados expostos em [LRPN07], revelam que ferramentas como o *SuPor-2* ou o *TextRank* combinado com *Thesaurus* ou com *stemming* e remoção de *stopwords* conseguem resultados superiores, sendo a primeira ferramenta a que consegue melhores resultados (**$ROUGE - 1 = 0,5839$**).

Capítulo 5

Conclusão e Trabalho Futuro

Neste capítulo serão apresentadas e debatidas as conclusões obtidas no final deste trabalho tendo em conta os métodos experimentados e resultados obtidos. Também serão indicadas algumas abordagens que poderão ajudar a melhorar os resultados e respectivo trabalho futuro. A motivação para o desenvolvimento desta dissertação está relaccionada como a atenuação de um problema recorrente, devido à grande quantidade de informação disponível actualmente. Assim, pretendeu-se estudar soluções para tornar a pesquisa de informação mais eficaz, a nível do tamanho do seu conteúdo. Estão apresentadas várias abordagens nesta área, com vista à resolução deste problema, sedo algumas baseadas em métodos superficiais e outras em metodologias de conhecimento profundo, gerando sumários extractivos ou abstractivos. As formas estudadas para lidar com este problema não permitem o desenvolvimento de uma solução definitiva, pois as vantagens de umas fazem parte das desvantagens de outras. Esta dificuldade no encontro da solução prende-se com o facto de o processo de sumarização humano ser bastante complexo e como alguma falta de recursos computacionais, principalmente em trabalhos particulares. Deste modo, sistemas que façam uso de uma abordagem mais profunda tornam-se mais complexos, custosos e limitados a nível do domínio dos textos utilizados para sumarização, que consigam atingir resultados satisfatórios. Por esses motivos, recorre-se ao uso de metodologias menos complexas, baseadas na análise de características estruturais e superficiais de um texto, embora produzam sumários com baixa qualidade textual.

Tendo em consideração a subjectividade de um texto, tentou-se processar essa característica usando ferramentas que, de alguma forma, pudessem ajudar a ultrapassar a dificuldade na sumarização do mesmo. Assim, o objectivo fundamental deste trabalho foi o desenvolvimento de um sistema de sumarização automática, considerando as características do perfil do utilizador, subjectividade e identificação de tópicos num texto. Neste sentido, investigaram-se e implementaram-se metodologias que visam analisar, conteúdo de um documento, em características como a relevância e objectividade das palavras, localização das frases no texto, identificação de palavras que possam definir um dado domínio, e os tópicos presentes no texto, que podem ajudar na identificação de novos assuntos no conteúdo textual. Foi também explorada uma metodologia que englobasse a combinação dos métodos referido anteriormente. Com esta abordagem, pretendia-se estudar importância das características anteriormente indicadas e qual delas tem maior peso na identificação do conteúdo informativo de um documento.

Para uma maior facilidade na interacção com o sistema, foi desenvolvida uma interface gráfica que permite a produção de um sumário através dos métodos já referidos e a comparação visual deste com o texto original.

Com a avaliação utilizada, do tipo *intrínseca* e automática foi possível uma comparação objectiva entre os resultados do desempenho dos vários métodos, a nível de conteúdo informativo, embora esta não permita ajuizar sobre a coesão e coerência dos sumários. Sendo assim características como a legibilidade e inteligibilidade dos sumários não são consideradas. As experiências

realizados neste trabalho, tiveram como finalidade a exploração de várias vertentes do texto. Assim, utilizando o método híbrido, com diferentes conjuntos de pesos, é possível identificar quais as características, já referidas, que identificam com maior eficiência conteúdo informativo do texto analisado. Foi testada criação de sumários considerando, de forma isolada, cada uma das características do texto, relevância, objectividade e a combinação entre ambas, heurística posicional e perfil de utilizador (domínio) .
Esta avaliação foi levada a cabo pela ferramenta *ROUGE* e permitiu verificar que o método da relevância é o que obtém piores resultados, enquanto a posição da frase melhora a correlação entre os sumários gerados e os sumários de referência. Por isso, métodos híbridos que tenham em maior consideração esta característica do texto conseguem obter melhores resultados. Assim, com esta vantagem, da heurística posicional e com as melhorias verificadas com o uso de métodos que combinem diferentes métricas, pode-se concluir que os sistemas de sumarização automática devem explorar uma abordagem híbrida em que seja considerada a localização das frases, para que se consiga identificar com maior qualidade o conteúdo mais relevante de um texto-fonte.
Mesmo assim, representar o conteúdo informativo para gerar um sumário em linguagem natural, continua a ser uma tarefa muito complexa e difícil de conseguir com a tecnologia existente, principalmente, tendo em conta uma sumarização que não considere um domínio específico. A abordagem extrativa, explorada neste trabalho, permite a obtenção de resultados satisfatórios, para a apresentação do conteúdo mais relevante de um texto, mesmo que a sua textualidade seja limitada. Esses resultados são conseguidos através de uma abordagem híbrida, em que é atribuída maior importância à localização da frases e não são utilizadas informações sobre os tópicos do texto. Em comparação com outros trabalhos, os resultados para o conjunto de documentos do *DUC* são idênticos, não existindo uma variação significativa. Para o idioma português, ferramentas como o *SuPor-2*, combinação entre o *TextRank* e *Thesaurus* ou *TextRank*, *Stem* e remoção de *stopwords* obtiveram melhores resultados, do que a abordagem explorada neste trabalho.

Numa perspectiva de trabalho futuro, seria interessante a implementação de métodos que permitam a sumarização de documentos múltiplos, ou vários textos como os presentes em documentos de aglomerados de notícias. Assim, seria possível gerar um sumário que representasse um conjunto de documentos, com os tópicos mais relevantes e eliminando possíveis conteúdos repetidos, ou seja, reduzindo a redundância da informação desse conjunto.

Tendo em vista a abordagem extractiva presente neste trabalho, seria de grande importância a implementação de um método que substituísse expressões anafóricas, melhorando assim a textualidade do sumário.
Numa abordagem idêntica ao perfil do utilizador, poderia ser interessante e um factor de melhoria para o resultado final, a identificação dos termos mais relevantes no documento analisado, de modo a ajudar na classificação do documento e obter frases ainda mais relevantes no contexto específico do documento e não apenas no domínio em que este se insere.
Por fim, a implementação de uma abordagem abstractiva, poderia ser outro método importante a ser incluído neste trabalho, explorando assim características mais profundas do texto que permitam melhorar a textualidade do sumário e a redução das frases extraídas.

Bibliografia

[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. 18

[AP00] Einat Amitay and Cécile Paris. Automatically summarising web sites - is there a way around it? In *CIKM*, pages 173--179. ACM, 2000. Available from: http://dblp.uni-trier.de/db/conf/cikm/cikm2000.html#AmitayP00. 12

[Bax58] P. B. Baxendale. Machine-made index for technical literature -- an experiment. *IBM Journal of Reasearch and Development*, (2):354–361, 1958. 2, 9, 17

[BE97] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10--17, 1997. 22

[BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010. Available from: http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf. v, vii, 35

[BGMP01] Orkut Buyukkokten, Hector Garcia-Molina, and Andreas Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. pages 652--662. ACM, 2001. Available from: http://ilpubs.stanford.edu:8090/511/1/2001-45.pdf. 12

[BM05] Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297--328, September 2005. Available from: http://dx.doi.org/10.1162/089120105774321091. 13

[BME99] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 550--557, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/1034678.1034760. 13

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 107--117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V. Available from: http://dl.acm.org/citation.cfm?id=297805.297827. 19

[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. 18

[Bur13] Philip R. Burns. Morphadorner v2: A java library for the morphological adornment of english language texts. Technical report, Evanston, IL. Northwestern University, 2013. Available from: http://picard.at.northwestern.edu/morphadorner/download/morphadorner.pdf [cited 13 Setembro 2014]. 28, 36

[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. 21

[Cho00] Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26--33, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=974305.974309. 36

[CLW07] Yanmin Chen, Bingquan Liu, and Xiaolong Wang. Automatic text aummarization based on textual cohesion. In *Journal of Electronics(China)*, 2007. Available from: http://link.springer.com/article/10.1007%2Fs11767-005-0188-5 [cited 21 Julho 2014]. 8

[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. *J. Am. Soc. Inf. Sci.*, 46(5):348--369, June 1995. Available from: http://dx.doi.org/10.1002/(SICI)1097-4571(199506)46:5<348::AID-ASI6>3.0.CO;2-1. 20

[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*, pages 305--312, 2006. Available from: http://www.aclweb.org/anthology/E06-1039.pdf [cited 14 Setembro 2014]. 25

[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. 18

[dAdLJdAAC08] Sandra Eliza Fontes de Avila, Antonio da Luz Jr., Arnaldo de Albuquerque Araújo, and Matthieu Cord. Vsumm: An approach for automatic video summarization and quantitative evaluation. In *SIBGRAPI*, pages 103--110. IEEE Computer Society, 2008. Available from: http://dblp.uni-trier.de/db/conf/sibgrapi/sibgrapi2008.html#AvilaLAC08. 11

[DAL07] Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation.

In *AAAI*, pages 1334--1339. AAAI Press, 2007. Available from: http://dblp.uni-trier.de/db/conf/aaai/aaai2007.html#DiasAL07 [cited 2014-05-22]. 34

[DM07] Dipanjan Das and André F. T. Martins. A survey on automatic text summarization, 2007. 1, 5

[dmo14] Dmoz, 2014. Available from: http://www.dmoz.org/. 12

[duc07] Document understanding conferences [online]. 2007. Available from: http://duc.nist.gov/ [cited 18 Junho 2014]. 2, 18, 27, 46

[Edm69] H. P. Edmundson. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16:264--285, 1969. 2, 9, 17, 22, 37

[EN98] B. Endres-Niggemeyer. *Summarizing Information*. Springer, 1998. 8, 9

[EN00] Brigitte Endres-Niggemeyer. Human-style www summarization, 2000. 24

[ENMS95] Brigitte Endres-Niggemeyer, E. Maier, and Alexander Sigel. How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Inf. Process. Manage.*, 31(5):631--674, 1995. Available from: http://dblp.uni-trier.de/db/journals/ipm/ipm31.html#Endres-NiggemeyerMS95. 8, 9

[ER04a] Güneş Erkan and Dragomir R. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, Barcelona, Spain, 2004. 19

[ER04b] Günes 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. 19

[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. 45

[Fel98] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998. 27, 35

[FH02] Sanda Harabagiu Finley and Sanda M. Harabagiu. 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. 13

[fHLTB] Center for Human Language Technology and Bioinformatics. The hultiglib. Available from: http://www.di.ubi.pt/~jpaulo/hultiglib/ [cited 19 Setembro 2014]. 28, 31, 36, 37

[Fid86] Raya Fidel. writing abstracts for free-text searching bibtex. *Journal of Documentation*, 42(1):11--21, 1986. Available from: http://faculty.washington.edu/fidelr/RayaPubs/WritingAbstractsforFreeText.pdf [cited 22 Julho 2014]. 8

[FWRZ05] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang. Tapping into the power of text mining. 2005. This article is accepted for publication at the Communications of ACM. Available from: http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf [cited 2 Setembro 2014]. 5

[Gal01] Leo Galambos. Lemmatizer for document information retrieval systems in java. In Leszek Pacholski and Peter Ruzicka, editors, *SOFSEM*, volume 2234 of *Lecture Notes in Computer Science*, pages 243--252. Springer, 2001. Available from: http://dblp.uni-trier.de/db/conf/sofsem/sofsem2001.html#Galambos01. 17

[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. 20, 21

[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. 21

[GL10] Vishal Gupta and Gurpreet Singh Lehal. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), August 2010. 1, 24

[GL11] Pierre-Etienne Genest and Guy Lapalme. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 64--73, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=2107679.2107687. 13, 16

[GL12] Pierre-Etienne Genest and Guy Lapalme. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 354--358, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=2390665.2390745. 15

[goo14] Google, 2014. Available from: http://www.google.com/. 12

[GR09] Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90--93, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=1610195.1610208. 16

[Gre11] Charles F. Greenbacker. Towards a framework for abstractive summarization of multimodal documents. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 75--80, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=2000976.2000990. 15

[Hea97] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 1(23):33--64, 1997. 29, 36

[hot14] Hotbot, 2014. Available from: http://www.hotbot.com/. 12

[Hov02] Eduard Hovy. Automated text summarization. In R. Mitkov (ed), Oxford University Handbook of Computational Linguistics. Oxford: Oxford University Press, 2002. 5, 10

[Hut87] John Hutchins. Summarization: Some problems and methods. In *Meaning: The Frontier of Informatics*, pages 151--173. Aslib, 1987. 6

[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. 45

[JG96] Karen Sparck Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. 44

[Jin01] Hongyan Jing. Cut-and-paste text summarization, 2001. 8, 9, 12

[JM99] Hongyan Jing and Kathleen McKeown. The decomposition of human-written summary sentences. In *SIGIR*, pages 129--136. ACM, 1999. Available from: `http://dblp.uni-trier.de/db/conf/sigir/sigir99.html#JingM99`. 12

[Jon93] Karen Sparck Jones. What might be in a summary? In *Information Retrieval*, pages 9--26, 1993. Available from: `http://dblp.uni-trier.de/db/conf/ir/ir93.html#Jones93`. 10

[Jon98] Karen Sparck Jones. Automatic summarising: factors and directions. *CoRR*, cmp-lg/9805011, 1998. Available from: `http://dblp.uni-trier.de/db/journals/corr/corr9805.html#cmp-lg-9805011`. 5, 9, 10

[Jon07] Karen Sparck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449--1481, 2007. Available from: `http://dx.doi.org/10.1016/j.ipm.2007.03.009`. 1, 9

[JS08] Karel Ježek and Josef Steinberger. Automatic text summarization (the state of the art 2007 and new challenges), 2008. 21

[Koh90] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464--1480, 1990. 22

[KPC95] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. 1995. 2, 22

[KS14] Atif Khan and Naomie Salim. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1), 2014. 12, 15

[KvD78] Walter Kintsch and T. A. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85:363--394, 1978. 8

[LCH+06] Cheng-Hung Li, Chih-Yi Chiu, Chun-Rong Huang, Chu-Song Chen, and Lee-Feng Chien. Image content clustering and summarization for photo collections. In *ICME*, pages 1033--1036. IEEE, 2006. Available from: `http://dblp.uni-trier.de/db/conf/icmcs/icme2006.html#LiCHCC06`. 11

[LCJ03] Chang-Shing Lee, Yea-Juan Chen, and Zhi-Wei Jian. Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Systems with Applications*, 25(3):431 -- 447, 2003. Available from: http://www.sciencedirect.com/science/article/pii/S0957417403000629. 14

[LD97] Thomas K Landauer and Susan T. Dutnais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211--240, 1997. 21

[LDS+13] Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *EMNLP*, pages 747--757. ACL, 2013. Available from: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2013.html#LiakataDSBR13. 12

[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, 2011. Available from: http://dblp.uni-trier.de/db/journals/mta/mta51.html#LuoFZ11. 12

[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. 46

[Lin04] Chin-Yew Lin. Looking for a Few Good Metrics: ROUGE and its Evaluation. In *Working Notes of NTCIR-4*, pages 1--8, June 2004. 46

[LJH05] Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang. A fuzzy ontology and its application to news summarization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 35(5):859--880, Oct 2005. 14

[Llo08] Elena Lloret. Text summarization : An overview. Technical report, Dept. Lenguajes y Sistemas Informáticos Universidad de Alicante, Alicante, Spain, 2008. 10

[LRFP13] Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88(0):164 -- 175, 2013. Available from: http://www.sciencedirect.com/science/article/pii/S0169023X13000815. 12, 22

[LRPN07] Daniel S. Leite, Lucia H. M. Rino, Thiago A. S. Pardo, and Maria das Graças V. Nunes. Extractive automatic summarization: Does more linguistic knowledge make a difference? In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 17--24, Rochester, NY, USA, 2007. Association for Computational Linguistics. Available from: http://www.aclweb.org/anthology/W/W07/W07-0203. 51

[Luh58] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159--165, 1958. Available from: http://www.research.ibm.com/journal/rd/022/luhn.pdf. 1, 9, 17, 22

[Mó03] Marcelo Módolo. SuPor: um ambiente para a exploração de métodos extrativos para a sumarização automática de textos em português. Dissertação de mestrado, Universidade Federal de São Carlos, Junho 2003. 22

[MA12] I. F. Moawad and M. Aref. Semantic graph reduction approach for abstractive text summarization. In *Computer Engineering & Systems (ICCES)*, pages 132--138. Seventh International Conference, 2012. 16

[mal] Machine learning for language toolkit. Available from: `http://mallet.cs.umass.edu/` [cited 19 Setembro 2014]. 29

[Man99] Inderjeet Mani. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA, 1999. 10, 43

[Man01] Inderjeet Mani. *Automatic Summarization*, volume 3 of *Natural Language Processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2001. 5, 8, 9, 12, 45

[MBG98] Inderjeet Mani, Eric Bloedorn, and Barbara Gates. Using cohesion and coherence models for text summarization. In *AAAI Symposium Technical Report SS-989-06*, pages 69--76. AAAI Press, 1998. 20

[McC02] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002. Available from: `http://mallet.cs.umass.edu`. 29

[MH03] Sameer Maskey and Julia Hirschberg. Automatic summarization of broadcast news using structural features. In *INTERSPEECH*. ISCA, 2003. Available from: `http://dblp.uni-trier.de/db/conf/interspeech/interspeech2003.html#MaskeyH03`. 12

[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`. 46

[Mih05] Rada Mihalcea. Language independent extractive summarization. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, pages 49--52, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. Available from: `http://dx.doi.org/10.3115/1225753.1225766`. 19

[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`. 18

[Mly06] Angela Mlynarski. Automatic text summarization in digital libraries. Master's thesis, University of Lethbridge. Faculty of Arts and Science, LETHBRIDGE, ALBERTA, CANADA, 2006. Available from: `https://www.uleth.ca/dspace/bitstream/handle/10133/270/MR17413.pdf` [cited 2 Setembro 2014]. 12

[MPER01] C. B. Martins, T. A. S. Pardo, A. P. Espina, and L. H. M. Rino. Introdução à sumarização automática. Technical Report RT-DC 002/2001, Departamento de Computação - Universidade Federal de São Carlos, 2001. 38 p. 5, 8, 10

[MT04] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 404--411, Barcelona, Spain, July 2004. Association for Computational Linguistics. 19

[MT05] Rada Mihalcea and Paul Tarau. A language independent algorithm for single and multiple document summarization. In *In Proceedings of IJCNLP'2005*, 2005. 19

[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`. 18, 38

[Nen06] Ani Nenkova. Summarization evaluation for text and speech: Issues and approaches. Technical report, 2006. 43

[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`. 20

[NSKF00] Joel Larocca Neto, Alexandre Denes Santos, Celso A. A. Kaestner, and Alex Alves Freitas. Generating text summaries through the relative importance of topics. In Maria Carolina Monard and Jaime Simão Sichman, editors, *IBERAMIA-SBIA*, volume 1952 of *Lecture Notes in Computer Science*, pages 300--309. Springer, 2000. Available from: `http://dblp.uni-trier.de/db/conf/sbia/sbia2000.html#NetoSKF00`. 22

[OH01] V.M. Orengo and C. Huyck. A stemming algorithm for the portuguese language. In *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings.Eighth International Symposium on*, pages 186--193, Nov 2001. 28

[Ou09] Shiyan Ou. *Text Summarization in Digital Libraries: Development and Evaluation of a Multi-document Summarization Method for Research Abstracts*. VDM Verlag Saarbrücken, Germany, 2009. 12

[Par02] T. A. S. Pardo. GistSumm: Um sumarizador automático baseado na idéia principal de textos. Série de Relatórios do NILC NILC-TR-02-13, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, Setembro 2002. 22 p. 22

[Pat07] Kaustubh Raosaheb Patil. Automatic text summarization using pathfinder network scaling. Master's thesis, Faculty of Engineering in collaboration with Faculty of Economics University of Porto, Porto, Portugal, 2007. Available from: `http://repositorio-aberto.up.pt/bitstream/10216/11625/2/Resumo.pdf` [cited 14 Setembro 2014]. 21, 38

[PL04] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271--278, 2004. 27

[PL05] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115--124, 2005. 27

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79--86, 2002. 27

[PM92] Hans Paulussen and Willy Martin. Dilemma-2: A lemmatizer-tagger for medical abstracts. In *ANLP*, pages 141--146, 1992. Available from: `http://dblp.uni-trier.de/db/conf/anlp/anlp1992.html#PaulussenM92`. 17

[Por80] Martin Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130--137, 1980. 17, 28, 31

[PR03] T. A. S. Pardo and L. H. M. Rino. TeMário: Um corpus para sumarização automática de textos. Série de Relatórios do NILC NILC-TR-03-09, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, Outubro 2003. 11 p. 28, 46, 48

[PR04] T. A. S. Pardo and L. H. M. Rino. Descrição do GEI - Gerador de Extratos Ideais para o Português do Brasil. Série de Relatórios do NILC NILC-TR-04-07, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, Agosto 2004. 8 p. 23

[PRdGVN03] Thiago Alexandre Salgueiro Pardo, Lucia Helena Machado Rino, and Maria das Graças Volpe Nunes. Gistsumm: A summarization tool based on a new extractive method. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, and Maria das Graças Volpe Nunes, editors, *PROPOR*, volume 2721 of *Lecture Notes in Computer Science*, pages 210--218. Springer, 2003. Available from: `http://dblp.uni-trier.de/db/conf/propor/propor2003.html#PardoRN03`. 36

[PRN03] T. A. S. Pardo, L. H. M. Rino, and M. G. V. Nunes. NeuralSumm: Uma abordagem conexionista para a sumarização automática de textos. In *Anais do IV Encontro Nacional de Inteligência Artificial - ENIA*, pages 1--10, Campinas-SP, Brasil, 2 a 8 de Agosto 2003. 22

[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`. 46

[pts] Ptstemmer - a stemming toolkit for the portuguese language. Available from: `https://code.google.com/p/ptstemmer/` [cited 19 Setembro 2014]. 28, 31

[pub] Publico. Available from: `http://www.publico.pt` [cited 19 Setembro 2014]. 28

[RHM02] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399--408, 2002. 1, 43

[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 VIII: III Jornada de Minicursos de Inteligência Artificial*, pages 203--245, 2003. 8, 9, 43

[Sav99] Jacques Savoy. A stemming procedure and stopword list for general french corpora. *J. Am. Soc. Inf. Sci.*, 50(10):944--952, July 1999. Available from: http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:10<944::AID-ASI9>3.3.CO;2-H. 28

[SC06] Simon O. Sweeney and Fabio Crestani. Effective search results summary size and device screen size: Is there a relationship? *Inf. Process. Manage.*, 42(4):1056--1074, 2006. Available from: http://dblp.uni-trier.de/db/journals/ipm/ipm42.html#SweeneyC06. 12

[Sen05] Eloize Rossi Marques Seno. RHeSumaRST: Um Sumarizador Automático de Estruturas RST. Master's thesis, UNIVERSIDADE FEDERAL DE SÃO CARLOS, São Carlos, Basil, 2005. Available from: http://www.bdtd.ufscar.br/htdocs/tedeSimplificado//tde_busca/arquivo.php?codArquivo=885 [cited 21 Julho 2014]. 8

[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. 15

[SR01a] Diana Santos and Paulo Rocha. Evaluating cetempÚblico, a free resource for portuguese. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 450--457, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. Available from: http://dx.doi.org/10.3115/1073012.1073070. 28

[SR01b] Diana Santos and Paulo Rocha. Evaluating cetempublico, a free resource for portuguese. In *ACL*, pages 442--449. Morgan Kaufmann Publishers, 2001. Available from: http://dblp.uni-trier.de/db/conf/acl/acl2001.html#SantosR01. 46

[SSMB97a] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193--207, 1997. 19

[SSMB97b] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193--207, 1997. Available from: http://dblp.uni-trier.de/db/journals/ipm/ipm33.html#SaltonSMB97. 22

[sto] Ir multilingual resources at unine. Available from: http://members.unine.ch/jacques.savoy/clef/index.html [cited 21 Setembro 2014]. 33

[Sun11] S. Suneetha. Automatic text summarization: The current state of the art. *International Journal of Science and Advanced Technology (ISSN 2221-8386)*, 1(9), November 2011. 1

[Sur02] Mihai Surdeanu. Infrastructure for open-domain information extraction. In *In Proceedings of the Human Language Technology Conference (HLT*, pages 2002--325330, 2002. 13

[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. 18, 51

[SYY75] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33--44, 1975. 34

[tac14] Text analysis conference (tac) 2014 [online]. 2014. Available from: http://www.nist.gov/tac/2014/index.html [cited 18 Junho 2014]. 2, 16, 18, 27, 46

[Thu26] Samuel Thurber. *Précis writing for American schools: methods of abridging, summarizing, condensing, with copious exercises*. Little, Brown and Company, 1926. Available from: http://books.google.pt/books?id=TBFKAAAAIAAJ [cited 21 Julho 2014]. 8

[TIT+13] Shuhei Tarashima, Go Irie, Ken Tsutsuguchi, Hiroyuki Arai, and Yukinobu Taniguchi. Fast image/video collection summarization with local clustering. In Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann, editors, *ACM Multimedia*, pages 725--728. ACM, 2013. Available from: http://dblp.uni-trier.de/db/conf/mm/mm2013.html#TarashimaITAT13. 11

[TKK+09] Hideki Tanaka, Akinori Kinoshita, Takeshi Kobayakawa, Tadashi Kumano, and Naoto Kato. Syntax-driven sentence revision for broadcast news summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, UCNLG+Sum '09, pages 39--47, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Available from: http://dl.acm.org/citation.cfm?id=1708155.1708163. 14

[vHR13] Hans van Halteren and Margit Rem. Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century dutch charters. *Language Resources and Evaluation*, 47(4):1233--1259, 2013. Available from: http://dblp.uni-trier.de/db/journals/lre/lre47.html#HalterenR13. 17

[Wan10] Xiaojun Wan. Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1137--1145, Beijing, China, August 2010. Coling 2010 Organizing Committee. Available from: http://www.aclweb.org/anthology/C10-1128. 51

[Wik04] Wikipedia. Plagiarism --- Wikipedia, the free encyclopedia, 2004. Online; accessed 22-July-2004. Available from: http://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350. 18

[Xia00] Fei Xia. The segmentation guidelines for the penn chinese treebank (3.0), 2000. 14

[yah14] Yahoo!, 2014. Available from: `https://search.yahoo.com/`. 12

[yL04] Chin yew Lin. Rouge: a package for automatic evaluation of summaries. pages 25--26, 2004. 46, 47

[YsCyKQ07] Shiren Ye, Tat seng Chua, Min yen Kan, and Long Qiu. Document concept lattice for text understanding and summarization, 2007. 12

[ZW00] Klaus Zechner and Alex Waibel. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING*, pages 968--974. Morgan Kaufmann, 2000. Available from: `http://dblp.uni-trier.de/db/conf/coling/coling2000.html#ZechnerW00`. 11

Glossário

Anáfora	Em Linguística, é o processo de repetição de informação através de uma unidade linguística cuja referência representa semanticamente um sintagma que surge anteriormente.
Corpus	Conjunto de documentos ou informações relativas a uma disciplina, tema ou domínio. Em Linguística, denota um conjunto finito de enunciados representativos de uma determinada estrutura.
Hiperonímia	Relação entre palavras em que uma mais genérica (heperónimo), 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.
Holonímia	Relação de inclusão semântica entre duas unidades lexicais; uma denota um todo (holónimo) sem impor obrigatoriamente as suas prioridades semânticas à outra, considerada sua parte (merónimo).
HULTIGLIB	Ferramenta para processamento de texto em Java.
Meronímia	Relação de inclusão semântica entre duas unidades lexicais; uma denotando a parte (merónimo) e criando uma relação de dependência ao implicar a referência a um todo (holónimo), relativo a essa parte.
Morphadorner	Ferramenta para dividir o texto em tópicos em Java
N-grama	No campo da linguística computacional, representa uma sequência contígua de *n* elementos textuais, sendo eles caracteres, palavras, fonemas ou sílabas.
Polissemia	Qualidade de uma palavra que pode apresentar diferentes significados, dependendo dos usos linguísticos em que possa aparecer.
SentiWordNet	Recurso lexical para mineração da opinião (objectividade)
Sinonímia	Relação entre duas ou mais palavras que possuem proximidade semântica, podendo ser usadas no mesmo contexto sem que haja alteração de significado do enunciado em que ocorrem.
Stop Words	Palavras com frequência elevada que não transmitem informação
Textualidade	Conjunto de características que fazem com que um conjunto de frases seja considerado como um text., e não um amontoado de frases e palavras.