Uma Comparação Sistemática de Diferentes Abordagens para a Sumarização Automática Extrativa de Textos em Português

A Comparison of Multiple Approaches for the Extractive Summarization of Portuguese Texts

Miguel Costa e Bruno Martins
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
{miguel.angelo.costa,bruno.g.martins}@tecnico.ulisboa.pt

Resumo

A sumarização automática consiste na tarefa de gerar automaticamente versões condensadas de textos fonte, apresentando-se como um dos problemas fundamentais nas áreas da Recuperação de Informação e do Processamento de Linguagem Natural. Neste artigo, considerando metodologias puramente extrativas, são comparadas diferentes abordagens na tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. Através da utilização da bancada ROUGE como forma de medir a qualidade dos sumários produzidos, são reportados resultados para dois domínios experimentais diferentes, respetivamente envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português. Os resultados obtidos demonstram que uma *baseline* simples, baseada na seleção da primeira frase, obtém melhores resultados na construção de títulos de notícias de forma extrativa, em termos de várias métricas ROUGE. No segundo domínio experimental, envolvendo a geração de sumários de notícias, o método que obteve melhores resultados foi o algoritmo LSA Squared, para as várias métricas ROUGE consideradas neste trabalho.

Palavras chave

Sumarização Automática, Avaliação Comparativa

Abstract

Automatic document summarization is the task of automatically generating condensed versions of source texts, presenting itself as one of the fundamental problems in the areas of Information Retrieval and Natural Language Processing. In this paper, different extractive approaches are compared in the task of summarizing individual documents corresponding to journalistic texts written in Portuguese. Through the use of the ROUGE package for measuring the quality of the produced summaries, we report on results for two different experimental domains, involving (i) the generation of headlines for news articles written in European Portuguese, and (ii) the generation of summaries for news articles written in Brazilian Portuguese. The results demonstrate that methods based on the selection of the first sentences have the best results when building extractive news headlines in terms of several ROUGE metrics. Regarding the generation of summaries with more than one sentence, the method that achieved the best results was the LSA Squared algorithm, for the various ROUGE metrics.

Keywords

Automatic Summarization, Comparative Evaluation

1 Introdução

A sumarização automática consiste na tarefa de gerar versões condensadas de textos fonte, apresentando-se como um dos problemas fundamentais nas áreas da Recuperação de Informação e do Processamento de Linguagem Natural (Luhn, 1958; Baxendale, 1958; Edmundson, 1969). Assim como na sumarização de textos feita por humanos, um bom sumário gerado automaticamente deve preservar as ideias principais dos textos fonte, articulando em torno destas ideias as informações centrais contidas nos documentos. Nos dias de hoje, com a crescente disponibilização de informação textual na Web, a demanda por sistemas de sumarização automática que sejam rápidos, fiáveis e robustos é maior do que nunca. Temos ainda que a produção automática de sumários tem inúmeras aplicações, sendo que estudos anteriores demonstraram já a sua utilidade a ajudar utilizadores em tarefas envolvendo a compreensão de informação em documentos textuais (i.e., sumários automáticos podem acelerar a tomada de decisões com base em informação textual (Mani et al., 2002)), ou como

This work is licensed under a
Creative Commons Attribution 3.0 License

forma de complementar outras aplicações e interfaces. No entanto, apesar da investigação nesta área se ter iniciado há já mais de sessenta anos, existe ainda um logo caminho a percorrer, e a tarefa está longe de estar resolvida. Argumentamos que através deste artigo outros investigadores na área podem agora ter uma ideia mais precisa de qual a performance relativa para vários dos métodos frequentemente usados na área.

Neste artigo, considerando uma metodologia extrativa para a sumarização automática de textos (i.e., os sumários são gerados pela justaposição de segmentos extraídos dos textos originais), são comparadas diferentes abordagens na tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. As abordagens sob comparação incluem métodos baseados em fatorização de matrizes (i.e., duas abordagens baseadas em decomposição em valores singulares, e duas abordagens baseadas em fatorização de matrizes não-negativas), métodos baseados em centralidade em grafos (i.e., adaptações do algoritmo PageRank propostas para a tarefa de sumarização automática), e *baselines* heurísticas correspondendo, por exemplo, à seleção das primeiras frases dos documentos e à seleção das frases com o maior número de palavras-chave (i.e., bi-gramas frequentes). Através da utilização da bancada ROUGE (Lin, 2004b) como forma de medir a qualidade dos sumários produzidos, são reportados resultados para dois domínios experimentais diferentes, nomeadamente:

(I) Na tarefa de gerar sumários curtos (i.e., contendo apenas uma frase) que possam ser usados como títulos de textos jornalísticos, avaliando os diferentes métodos através de uma coleção extensa de artigos escritos em Português Europeu, com cada um dos artigos associado ao título correspondente, tal como selecionado pelos editores de um portal de notícias *on-line*. Nestes testes verificou-se que o método mais simples, baseado na seleção da primeira frase de cada documento, obtém os melhores resultados.

(II) Na tarefa de gerar sumários para textos jornalísticos na variante Brasileira do Português, avaliando os diferentes métodos através de experiências com as coleções de documentos TeMário (Pardo & Rino, 2003; Maziero et al., 2007). Nestes testes verificou-se que uma abordagem baseada em decomposição de matrizes em valores singulares obtém os melhores resultados.

O restante conteúdo deste artigo encontra-se organizado da seguinte forma: A Secção 2 apresenta os principais conceitos e trabalhos anteriores na área. A Secção 3 descreve detalhadamente os vários métodos de sumarização automática que foram alvo do nosso estudo comparativo. A Secção 4 descreve a metodologia experimental considerada. A Secção 5 apresenta os resultados obtidos e uma breve análise desses resultados. Finalmente, a Secção 6 apresenta um breve resumo das principais conclusões, e descreve possíveis caminhos para trabalho futuro.

2 Trabalho Relacionado

A sumarização automática de documentos tem sido investigada ativamente na área do Processamento de Linguagem Natural (PLN) desde a segunda metade do século passado. Esta é atualmente uma área de investigação muito vasta, com inúmeros trabalhos publicados. Aos leitores deste artigo, sugere-se a consulta do trabalho de Nenkova et al. (2011) para uma visão geral sobre a área, ou a consulta do relatório técnico de Pardo (2008) para um resumo de trabalhos importantes focados no Português.

Radev et al. (2002) definiram um sumário como *um texto que é produzido a partir de um ou mais textos originais, que transmite a informação importante no(s) texto(s) original(ais), e que não é maior do que a metade do(s) texto(s) original(ais), normalmente tendo um tamanho significativamente menor*. Esta definição simples captura três aspetos importantes que caracterizam a investigação sobre o tema da sumarização automática de documentos, nomeadamente (i) os sumários podem ser produzidos a partir de um único documento ou de vários documentos, (ii) os sumários devem preservar a informação importante, e (iii) os sumários devem ser curtos. Dado que o fluxo e a densidade de informação, num determinado documento, é geralmente não-uniforme (i.e., algumas partes dos documentos são mais importantes do que outras), o grande desafio que se apresenta à sumarização automática consiste em discriminar as partes mais informativas de um documento.

O texto introdutório de Radev et al. (2002) também apresenta alguma terminologia importante na área da sumarização automática. Temos assim que *extração* se refere ao processo de identificação de segmentos importantes de um texto original, reproduzindo-os na íntegra aquando da geração de sumários. Por outro lado, *abstração* é um processo que tem como objetivo produzir conteúdos textuais novos que reflitam

os aspetos importantes de uma dada fonte textual. Um processo de *fusão* combina segmentos extraídos de forma coerente. Finalmente, um processo de *compressão* visa remover segmentos pouco importantes dos textos produzidos como sumários (Coster & Kauchak, 2011). Importa notar que, enquanto a sumarização automática extrativa se preocupa principalmente com o conteúdo dos sumários, baseando-se geralmente apenas na extração de frases, a sumarização automática abstrativa coloca uma forte ênfase na forma, com o objetivo de produzir sumários gramaticalmente corretos, o que geralmente requer técnicas avançadas de geração de linguagem natural (i.e., a sumarização abstrativa normalmente envolve a fusão da informação extraída, a compressão de frases, e a reformulação de frases (Knight & Marcu, 2002; Jing & McKeown, 2000; Almeida & Martins, 2013)). Embora um sumário abstrativo possa ser mais conciso, os sumários puramente extrativos são mais viáveis computacionalmente, e estas abordagens mais simples tornaram-se o padrão no campo da sumarização automática de textos.

Os primeiros trabalhos de investigação, abordando a sumarização automática extrativa de documentos textuais escritos em Inglês, propuseram métodos heurísticos para extrair as frases mais importantes de documentos individuais, usando combinações de atributos como a frequência de palavras ou de segmentos de texto (Luhn, 1958), a posição no texto (Baxendale, 1958), ou a ocorrência de segmentos-chave inferidos para os textos (Edmundson, 1969). Por exemplo na abordagem proposta por Luhn (1958), as palavras são inicialmente transformadas nos seus radicais (i.e., é efetuado um processo de *stemming*), e as *stop-words* comuns são removidas. Luhn compilou uma lista de palavras significativas (i.e., palavras associadas a conteúdos semânticos importantes), classificando-as por ordem decrescente de frequência, sendo que o índice de cada palavra nesta lista ordenada proporciona uma medida da sua importância. Ao nível de cada frase, um fator de importância é derivado das palavras que a constituem, refletindo o número de ocorrências de palavras significativas dentro da frase, e a distância linear entre elas (i.e., considerando a possível utilização de palavras não significativas nas frases, entre as ocorrências de palavras significativas). As várias frases de um documento são classificados por ordem do seu fator de importância, e as primeiras frases nesta ordenação são finalmente selecionadas para formar o sumário. Vários trabalhos posteriores consideraram ideias semelhantes às propostas nestes trabalhos seminais, concentrando-se na aplicação a outros idiomas ou a domínios de aplicação específicos, com especial foco no caso de textos jornalísticos. No caso particular da língua Portuguesa, a grande maioria dos trabalhos anteriores estudou a aplicação de abordagens heurísticas semelhantes às descritas atrás.

Na década de 1990, com a crescente popularização do uso de técnicas de aprendizagem automática em tarefas de PLN, surgiram uma série de publicações envolvendo abordagens estatísticas para a produção automática de sumários. Por exemplo Kupiec et al. (1995) descrevem um método derivado do trabalho de Edmundson (1969), que era capaz de aprender a partir de dados anotados manualmente. Uma função de classificação, baseada na abordagem naïve-Bayes, era usada para categorizar cada frase como interessante de considerar (i.e., de extrair) para o sumário ou não. As características usadas pela função de classificação eram muito semelhantes às do trabalho de Edmundson (1969), mas estes autores incluíram ainda o comprimento da frase e a presença de palavras em maiúsculas. A cada frase era atribuída uma pontuação de acordo com a probabilidade obtida pelo classificador naïve-Bayes, e as n frases melhor pontuadas formam o sumário.

Aone et al. (1999) também utilizaram um classificador naïve-Bayes, mas neste caso com um conjunto de características mais ricas (e.g., considerando características tais como pesos obtidos pela heurística *term-frequency* $\times$ *inverse document frequency* (TF-IDF) para cada um dos termos presentes nos documentos, como forma de tentar capturar conceitos-chave nos documentos a sumarizar). Neste trabalho, além de palavras individuais, os autores consideraram ainda o uso de colocações relevantes (i.e., bi-gramas de substantivos) calculadas estatisticamente, ou o uso de entidades mencionadas nos textos, como unidades de contagem. Os autores também utilizaram técnicas simples como forma de resolver alguns tipos de co-referências nos textos (e.g., associar os acrónimos dentro de um documento a um mesmo conceito único, como *EUA* a *Estados Unidos* ou como *IBM* a *International Business Machines*, por forma a aumentar a coesão nas representações). Sinónimos e variantes morfológicas também foram fundidas ao considerar os termos lexicais individuais, sendo os mesmos identificados através da WordNet (Miller, 1995).

Lin (1999) rompeu com o pressuposto de modelação em que as características usadas na classificação são independentes umas das outras, tentando modelar o problema da extração de frases importantes usando árvores de decisão e com um

amplo conjunto de características, em vez de usar um classificador naïve-Bayes. Osborne (2002), por sua vez, utilizou modelos de máxima entropia treinados por um método de gradiente descendente conjugado, considerando características como pares de palavras, com todas as palavras truncadas para um máximo de dez carateres, o comprimento das frases, a posição das frases no texto, e outras características simples tais como a ocorrência das frases dentro de secções de introdução ou de conclusão.

Conroy & O'Leary (2001) abordaram o problema da extração de frases desde documentos de texto através de modelos de Markov com variáveis ocultas (HMMs), numa tentativa de capturar dependências locais entre frases. Apenas três características representativas das frases foram utilizadas neste trabalho, nomeadamente a posição da frase no documento (i.e., informação embutida na própria estrutura de estados do HMM), o número de termos na frase, e a probabilidade de observar os termos da frase, dados os termos do documento. Além de HMMs, outros trabalhos anteriores basearam-se em modelos sequenciais mais sofisticados, por exemplo considerando o formalismo dos *Conditional Random Fields* (CRFs) e utilizando conjuntos de características mais ricos (Shen et al., 2007).

As propostas baseadas em aprendizagem automática apresentam geralmente a desvantagem de necessitarem de dados de treino sob a forma de frases extraídas desde documentos de texto (i.e., os exemplos de treino consistem de frases anotadas como interessantes ou não de pertencer a um sumário extrativo). Explorando a convenção de que as partes mais importantes de um texto jornalístico são geralmente colocadas nos parágrafos iniciais (i.e., são poucas as técnicas de sumarização automática extrativa que conseguem resultados significativamente melhores do que uma abordagem simplista baseada em extrair as primeiras frases), Svore et al. (2007) propuseram e avaliaram uma abordagem baseada em redes neuronais (i.e., baseada num algoritmo de *learning to rank* denominado RankNet), na qual se treina um modelo a partir de características (e.g., frequências de n-gramas) extraídas desde frases em textos jornalísticos e das suas respetivas posições nos textos. O modelo aprende a inferir qual a posição que seria mais adequada para cada frase, sendo que posteriormente as frases associadas às primeiras posições são as selecionadas para a elaboração do sumário. Os autores utilizaram também, nos seus modelos, características derivadas de recursos de informação externos, tais como artigos da Wikipédia ou históricos de pesquisas efetuadas no motor de busca de notícias da Microsoft, sob a conjetura de que frases de um documento que contenham palavras frequentemente usadas nas pesquisas do motor de busca, ou que contenham entidades correspondentes a artigos da Wikipédia, devem ser consideradas como interessantes para colocar nos sumários.

Como forma de contornar a necessidade de obtenção de dados de treino, vários trabalhos anteriores exploraram ainda técnicas não-supervisionadas, por exemplo baseadas em fatorização de matrizes. Intuitivamente, estes métodos tentam agrupar as frases de um documento em grupos/componentes que sejam coerentes entre si, escolhendo posteriormente as frases mais representativas de cada grupo, por forma a construir os sumários. Por exemplo Gong & Liu (2001) propuseram um método que utiliza análise semântica latente (LSA) para selecionar frases adequadas à construção de um sumário. Este método cria inicialmente uma matriz de *termos* × *frases*, onde cada coluna representa o vetor de frequências ponderadas dos termos de uma frase. De seguida, é aplicada uma decomposição em valores singulares (SVD) sob a matriz, por forma a derivar a estrutura semântica latente. A(s) frase(s) com maior(es) peso(s) no primeiro conceito latente (i.e., o conceito correspondente ao primeiro valor singular) são finalmente selecionadas para a formação do sumário (i.e., o método escolhe a(s) frase(s) mais informativa(s) do primeiro valor singular). Steinberger & Ježek (2004) propuseram uma abordagem alternativa também baseada na decomposição SVD, capaz de produzir resultados de melhor qualidade, onde se usam os vários valores singulares. Por outro lado, autores como Lee et al. (2009) ou como Mashechkin et al. (2011) propuseram a utilização de fatorização de matrizes não-negativas (NMF), decompondo a matriz de *termos* × *frases* em factores não negativos, por forma a extrair as frases com maior pontuação em cada um dos componentes latentes descobertos desta forma.

Outros autores ainda propuseram a utilização de métodos não-supervisionados baseados em grafos, como forma de capturar as frases mais centrais para a construção de sumários extrativos. Os métodos baseados em grafos começam normalmente pela construção de um grafo que represente o documento, ou coleções de documentos, a sumarizar. Nesta representação, cada nó do grafo é geralmente uma frase e, se a similaridade entre um par de frases estiver acima de um dado valor limiar, então existe uma aresta entre o par de frases. As frases centrais são selecionadas para formar os sumários através de um

processo de votação pelas suas frases vizinhas. Por exemplo Erkan & Radev (2004) propuseram um algoritmo chamado LexRank para calcular a importância de cada frase, com base no conceito de *eigenvector centrality* (i.e., uma noção de prestígio dos nós, semelhante à que se encontra associada ao algoritmo PageRank do Google (Page et al., 1999; Franceschet, 2011)). Outros métodos semelhantes foram propostos por Mihalcea (2004), por Mihalcea & Tarau (2005), ou por Wan & Yang (2008). Yeh et al. (2005) propuseram um método que combina as ideias de análise semântica latente (LSA) e centralidade em grafos. As representações semânticas das frases, obtidas por decomposição em valores singulares, são usadas para construir um grafo de relações entre as frases. Finalmente, é aplicada uma medida de significância dos nós do grafo, baseada no trabalho original de Salton et al. (1997), e são escolhidas as k frases mais conectadas no grafo, sendo as mesmas apresentadas de acordo com a ordem com que as frases surgem no documento original.

Alguns esforços anteriores focaram-se, por sua vez, no agregar dos resultados de vários métodos diferentes de sumarização automática. Thapar et al. (2006) apresentaram uma abordagem de meta-sumarização baseada em grafos, que compara grafos gerados com base em cada um dos sumários produzidos pelos métodos individuais, com um grafo que agrega os resultados dos diferentes métodos de sumarização. Wang & Li (2010) avaliaram sistematicamente diferentes métodos para a combinação dos resultados de sistemas de sumarização extrativa (i.e., diferentes esquemas de agregação de ordenações de frases), propondo depois um método de consenso ponderado para agregar os resultados de vários métodos.

Vários trabalhos anteriores abandonaram a sumarização automática de documentos individuais, em vez disso considerando múltiplas fontes de informação (i.e., sumarização multi-documento) que se podem sobrepor e complementar, ocasionalmente apresentando ainda contradições. Na sumarização multi-documento, as principais tarefas relacionam-se não só com identificar e lidar com redundância em documentos, mas também com o reconhecer conteúdos novos e com o garantir que o sumário final é coerente e completo. Técnicas extrativas foram aplicadas à sumarização automática multi-documento, fazendo por exemplo uso de medidas de similaridade entre pares de frases. As abordagens propostas variam essencialmente na forma como estas semelhanças são utilizadas: alguns trabalhos procuram identificar temas comuns através do agrupamento (i.e., *clustering*) de frases e, em seguida, selecionam uma frase para representar cada grupo (McKeown & Radev, 1995; Radev et al., 2000), enquanto outros métodos geram uma frase composta de elementos extraídos de cada grupo (Barzilay et al., 1999), e outros autores ainda estudaram abordagens dinâmicas que incluem cada passagem candidata apenas se a mesma for considerada nova no que diz respeito às passagens incluídas anteriormente, através do conceito da relevância marginal máxima (Carbonell & Goldstein, 1998). Alguns autores abordaram ainda a sumarização multi-documento em conjunto com a compressão de frases (Almeida & Martins, 2013), enquanto outros autores estudaram o problema da sumarização multi-documento em diferentes variantes multilingues (Litvak et al., 2010; Siddharthan & McKeown, 2005; Fung & Ngai, 2006).

Importa ainda referir que alguns métodos modernos para a sumarização automática, em particular no caso da sumarização multi-documento ou no caso de métodos que combinam a compressão de frases com a extração de frases relevantes, se baseiam no formalismo da programação linear inteira (ILP). Nestas abordagens, a tarefa de sumarização é vista como um problema de otimização combinatória, em que se pretende selecionar um conjunto de frases, até um tamanho máximo preestabelecido, que maximize a soma das pontuações de relevância, e que ao mesmo tempo minimize a redundância entre as frases selecionadas, de um conjunto de frases obtido através do processamento do(s) documento(s) e considerando diferentes taxas de compressão para as frases originais (Hirao et al., 2009; Almeida & Martins, 2013). No entanto, os métodos de sumarização considerados no nosso estudo comparativo baseiam-se apenas na seleção das frases melhor pontuadas em termos da sua relevância, não tendo sido considerada a combinação de pontuações de relevância com outras medidas (e.g., capturando a redundância entre as várias frases seleccionadas).

3 Métodos de Sumarização Automática Considerados no Estudo Comparativo

Neste artigo, foram realizadas experiências comparativas envolvendo diferentes métodos de sumarização automática extrativa. As abordagens sob comparação incluem métodos baseados em fatorização de matrizes (i.e., duas abordagens baseadas em decomposição em valores singulares, e duas baseadas em fatorização de matrizes não-negativas), métodos baseados em grafos

(i.e., adaptações do algoritmo PageRank propostas para a tarefa de sumarização automática), e *baselines* heurísticas, correspondendo à seleção das primeiras frases dos documentos, e à seleção das frases com maior relevância em termos da ocorrência de termos-chave, especificamente considerando bi-gramas frequentes.

Os vários métodos de sumarização automática em estudo partilham uma fase de pré-processamento comum, em que os textos a sumarizar são segmentados em palavras e em frases, através do uso dos mecanismos de segmentação de textos, comuns a várias línguas Indo-Europeias, disponíveis num pacote Python para PLN denominado nltk[1]. Em todos os métodos testados, temos que as diferentes frases dos documentos foram representadas através de conjuntos de palavras-chave, tendo sido removidas *stop-words* comuns, e tendo os textos sido convertidos por forma a se utilizarem apenas carateres minúsculos e sem acentos.

As seguintes subsecções apresentam em detalhe cada um dos métodos considerados, os quais permitem essencialmente pontuar as frases de um dado texto fonte, de acordo com a sua adequabilidade para pertencerem a um sumário. A produção dos sumários com base nos métodos descritos de seguida baseia-se na seleção da(s) frase(s) melhor pontuada(s), concatenando-as pela ordem na qual surgem no texto original. As abordagens de sumarização consideradas neste artigo são desta forma puramente extrativas, muito embora para trabalho futuro se considere também a integração de abordagens de compressão de frases (Yamangil & Nelken, 2008; Coster & Kauchak, 2011; Bach et al., 2011).

3.1 Decomposição em Valores Singulares

A análise semântica latente (LSA) é um método algébrico frequentemente utilizado na área do PLN para analisar as relações entre documentos e os termos neles contidos, através da produção de um conjunto de conceitos latentes relacionados com os documentos e os termos. Este método assume que as palavras semanticamente relacionadas tendem a coocorrer nos textos. No contexto da aplicação em sumarização automática, uma matriz esparsa A representando as ocorrências de termos por frases (i.e., uma matriz com m linhas que representam os termos únicos, e com n colunas que representam cada frase) é construída a partir de um documento original. De seguida, é calculada uma decomposição em valores singulares (SVD), tipicamente com o objetivo de reduzir o número de linhas (i.e., os termos correlacionados são agrupados em conceitos, capturando assim fenómenos como a sinonímia entre termos), preservando a estrutura de similaridade entre as colunas (i.e., entre as frases). A decomposição SVD é tipicamente calculada com base num algoritmo que envolve duas etapas, o qual começa por reduzir a matriz A à sua forma bi-diagonal (i.e., a uma matriz com entradas diferentes de zero apenas na diagonal principal e nos valores que se encontram acima ou abaixo), e que de seguida calcula a decomposição SVD da matriz bi-diagonal através de um método iterativo.

Resumidamente, temos que a decomposição SVD fatoriza a matriz A em três matrizes, U, D e V^T, de tal forma que $A = UDV^T$. A matriz $U = [u_{ij}]$, de dimensão $m \times n$, é uma matriz unitária cujas colunas são vetores ortonormais, denominados os vetores singulares à esquerda. Por sua vez $D = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$ é uma matriz diagonal $n \times n$, cujos elementos diagonais são valores singulares não negativos, ordenados de forma decrescente. Finalmente, temos que $V^T = [v_{ji}]$ é uma matriz ortogonal de dimensão $n \times n$, cujas colunas são chamadas os vetores singulares à direita.

No contexto da aplicação em sumarização automática, um algoritmo simples pode ser usado para selecionar a(s) melhor(es) frase(s), com base na decomposição SVD e usando a matriz de valores singulares à direita V^T. Cada frase i é representada pelo vetor coluna $\psi_j = [v_{j1}v_{j2}, \ldots, v_{jn}]^T$ da matriz V^T e, com uma abordagem simples, basta-nos selecionar o primeiro vetor singular direito da matriz V^T e, em seguida, a(s) frase(s) que têm o maior valor índice no vetor. Seguidamente, se necessário, efetua-se o mesmo processo para o segundo vetor singular direito da matriz, até se chegar ao número desejado de frases selecionadas para a construção do sumário (Gong & Liu, 2001). Neste trabalho denominamos esta abordagem de LSA Classic.

Uma outra abordagem para selecionar a(s) melhor(es) frase(s) a partir da SVD foi proposta por Steinberger & Ježek (2004). Em vez de selecionar a(s) frase(s) de topo do primeiro vetor singular, a(s) frases são selecionadas com base num peso global obtido de todos os vetores singulares. Para cada vetor coluna da matriz V^T (i.e., para cada frase j), vamos calcular a raiz quadrada do quadrado dos seus componentes multiplicados pelo quadrado dos valores singulares correspondentes na matriz D, desta forma favorecendo os valores do índice na matriz V^T que correspondem aos maiores valores singulares. Em seguida, escolhem-se as frases com o maior peso com-

[1] http://www.nltk.org/

binado em todos os componentes importantes. De uma forma resumida, temos que de acordo com este método escolhemos a(s) frase(s) j que tenha(m) o(s) valor(es) de relevância mais elevado(s), tal como produzidos por:

$$s_j = \sqrt{\sum_{i=1}^{n} v_{j,i}^2 \times \alpha_i^2} \quad (1)$$

Na equação, s_j é o peso do vetor de termos j no modificado espaço de vetores latente e n é o número de dimensões do novo espaço. Denominamos esta abordagem por LSA Squared.

Nos testes aqui reportados, para efetuar a fatorização SVD das matrizes de termos por frases, foi utilizada a implementação de um pacote Python denominado scikit-learn[2].

3.2 Fatorização Não Negativa

A fatorização de matrizes não-negativas (NMF) é um método de decomposição de matrizes recentemente desenvolvido, o qual impõe a restrição de que as entradas e os factores resultantes devem ser não-negativos, ou seja, todos os elementos das matrizes resultantes da decomposição têm de ser iguais ou maiores que zero. Na proposta original de Lee & Seung (1999) a fatorização em matrizes não-negativas decompõe uma matriz A com m linhas e n colunas, correspondendo à representação das n frases contendo m termos, em duas matrizes não-negativas W e H, de forma a que $A_{m\times n} \approx W_{m\times r} \times H_{r\times n}$, onde $W_{m\times r}$ é uma matriz não-negativa de características semânticas (i.e., a matriz dos termos) e onde $H_{r\times n}$ é uma matriz não negativa de variáveis semânticas (i.e., a matriz das frases). Um dos algoritmos mais populares para encontrar decomposições NMF é baseado numa regra de atualização multiplicativa, que atualiza iterativamente as matrizes W e H até obter a convergência de uma função objetivo do tipo $J = ||A - WH||^2$ sob um determinado valor limiar predefinido, ou até o algoritmo exceder um determinado número de passos. As seguintes regras de atualização são utilizadas em cada passo do algoritmo iterativo:

$$H_{\alpha,\mu} \leftarrow H_{\alpha,\mu} \times \frac{(W^T A)_{\alpha,\mu}}{(W^T W H)_{\alpha,\mu}} \quad (2)$$

$$W_{i,\alpha} \leftarrow W_{i,\alpha} \times \frac{(AH^T)_{i,\alpha}}{(WHH^T)_{i,\alpha}} \quad (3)$$

Depois de encontrar a decomposição NMF, podemos calcular uma medida genérica de relevância para cada frase, de acordo com a proposta original de Lee et al. (2009), a qual corresponde à seguinte equação:

$$\mathrm{GR}_j = \sum_{i=1}^{r} \left(H_{i,j} \times \frac{\sum_{q=1}^{n} H_{i,q}}{\sum_{p=1}^{r}\sum_{q=1}^{n} H_{p,q}} \right) \quad (4)$$

A(s) frase(s) com valor(es) mais elevado(s) em termos da medida de relevância são finalmente selecionadas para a formação do sumário. Denominamos esta abordagem, neste trabalho, pela sigla NMF GR.

Outra abordagem baseada em NMF corresponde à proposta de Mashechkin et al. (2011), onde também se pretende encontrar uma medida de relevância para cada frase, mas de uma forma mais abrangente, utilizando a seguinte equação:

$$\mathrm{ExtR}_j = \sum_{i=1}^{k} \left(\|W_i\|^2 \times \|H_i\| \times H_{i,j} \right) \quad (5)$$

Na equação $\|W_i\|^2$ é o quadrado da norma Euclideana do vetor W_i e $\|H_i\|$ é a norma Euclideana dos tópicos do vetor H_i. A(s) frase(s) com valor(es) mais elevado(s) em termos desta medida de relevância são finalmente selecionadas para a formação do sumário. Denominamos esta abordagem por NMF ExtR.

Nos testes aqui reportados, para efetuar a fatorização NMF das matrizes, foi utilizada a implementação do pacote scikit-learn.

3.3 Centralidade em Grafos

Além de abordagens baseadas em fatorização de matrizes, foram também feitos testes com métodos baseados em grafos para a sumarização automática de textos, seguindo as ideias originalmente apresentadas por Mihalcea (2004). Para efetuar os testes em que se usam algoritmos baseados em grafos foi utilizado um pacote Python denominado Networkx[3]. Por forma a suportar a aplicação de algoritmos de ordenação baseados em grafos, em tarefas de sumarização automática, começamos por construir um grafo que represente o documento a sumarizar, interligando as suas frases através de relações de similaridade que capturem a sobreposição de conteúdos textuais. Estas relações entre pares de frases podem ser vistas como um processo de recomendação, em que uma frase que aborda alguns conceitos de um texto dá ao leitor uma recomendação, no sentido de ele se poder referir a outras frases no

[2] http://scikit-learn.org/

[3] http://networkx.github.io/

texto que abordem os mesmos conceitos (Mihalcea, 2004). Nas nossas experiências, foram testadas as seguintes métricas de similaridade diferentes, aplicando posteriormente uma variante do algoritmo PageRank (Page et al., 1999), para grafos pesados e não-direcionados:

- O coeficiente de similaridade de Jaccard entre as frases, com base em termos individuais. Este coeficiente mede a similaridade entre dois conjuntos de termos A e B, sendo definido como o tamanho da intersecção dos conjuntos dividido pelo tamanho da sua união:

$$\text{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

- A similaridade do cosseno entre as frases, com base em termos individuais. São calculadas as mesmas medidas de pesagem, utilizadas nos testes envolvendo decomposição de matrizes, para os termos em cada uma das frases, o que dá origem a uma matriz de *termos* × *frases*. Seguidamente é calculado o cosseno do ângulo θ formado entre os vetores que representem pares de frases A e B, de acordo com a seguinte equação:

$$\text{sim}(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (7)$$

Em cada uma das abordagens anteriores, o cálculo da similaridade entre pares de frases resulta num grafo denso (i.e., todos os nós encontram-se interligados entre si), em que existem pesos associados a cada aresta não-direcionada, indicando a força das associações entre os vários pares de frases no texto. Sob este grafo, é executada uma versão ligeiramente modificada do algoritmo PageRank (PR) (Page et al., 1999), por forma a obter uma ordenação dos nós do grafo de acordo com o seu prestígio/importância. Esta versão adaptada permite contabilizar pesos tal como associados às arestas, correspondendo à seguinte equação, onde $\text{Ln}(V_i)$ corresponde ao conjunto de nós ligados no grafo ao nó V_i, e onde $w_{j,i}$ corresponde ao peso da aresta que liga os nós j e i.

$$\text{PR}(V_i) = \frac{(1-d)}{N} + d \times \sum_{V_j \in \text{Ln(V}_\text{i})} \frac{w_{j,i}}{\sum_{V_k \in \text{Ln}(V_j)} w_{j,k}} \text{PR}(V_j) \quad (8)$$

Da equação acima, é possível verificar que cada nó V_i do grafo terá uma pontuação que depende de um fator inicial $(1-d)$, uniformemente associado cada um dos N nós e que normalmente é escolhido com base no valor $d = 0.85$. A pontuação de cada nó depende também das pontuações associadas aos nós que lhe estão diretamente associados. O cálculo do PageRank, normalmente efetuado de forma iterativa, produz, como resultado, uma distribuição de probabilidade sobre os nós do grafo, em que cada nó terá uma probabilidade correspondente à sua importância, tal como derivada das associações para com todos os restantes nós do grafo. Depois do algoritmo PageRank ser executado sob o grafo, as frases (i.e., os nós do grafo) são ordenadas pela sua pontuação, e a(s) frase(s) de topo são selecionadas para inclusão no sumário, do documento. No nosso estudo denominamos esta abordagem por TextRank.

Foi também testado o uso de uma probabilidade inicial não-uniforme para cada frase. O grafo é construído de forma semelhante ao caso do TextRank, ou seja, é criado um grafo não-direcionado, interligando cada frase do documento a sumarizar através de relações de similaridade. Uma distribuição de probabilidades inicial, sob cada um dos N nós, é calculada através da seguinte fórmula, a qual substitui a primeira parte da Equação 8:

$$\text{PR}(V_i) = \frac{\frac{1}{\text{POS}(V_i)^\alpha}}{\sum_{j=1}^{N} \frac{1}{\text{POS}(V_j)^\alpha}} \times (1-d) + d \times \ldots \quad (9)$$

Nesta equação o parâmetro α foi usado com o valor de 0.85, tendo este valor sido selecionado por ter obtido os melhores resultados num conjunto inicial de experiências. A função $\text{POS}(V_i)$ refere-se a posição de cada frase V_i no documento, dando-se desta forma um peso superior às primeiras frases. Neste trabalho denominamos esta abordagem por TextRank Init.

3.4 Abordagens Heurísticas

Foram ainda efetuados testes com dois métodos *baseline* inspirados em heurísticas simples que foram propostas em alguns dos trabalhos seminais na área da sumarização automática.

O primeiro destes métodos *baseline* pretende demonstrar a convenção de que as partes mais importantes de um texto jornalístico são geralmente colocadas nos parágrafos iniciais, tendo para isso sido comparadas duas variantes distintas desta ideia. A primeira variante consiste em selecionar a(s) primeira(s) frases de cada documento, e a segunda variante consiste em selecionar a(s) frase(s) de forma aleatória. O segundo método *baseline* procura selecionar a(s) frase(s) onde ocorrem mais vezes os conceitos chave do documento. São primeiro extraídos os n conceitos chave (i.e., os n bi-gramas de palavras com maior número de ocorrências) mais importantes do documento e, para cada frase, são somadas as pontuações associadas aos conceitos chave que nelas ocorram. Por fim, são selecionadas as frases com maior pontuação agregada.

4 Avaliação Experimental

A avaliação comparativa dos métodos de sumarização automática, descritos na secção anterior, foi feita com base em dois domínios experimentais diferentes, envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português.

No primeiro caso, foram usados 40.000 textos jornalísticos publicados originalmente no portal *sapo.pt*, associados aos respetivos títulos tal como selecionados pelos editores do portal. A tarefa a resolver consiste em gerar sumários curtos (i.e., de uma frase apenas) com base no texto das notícias, que se apresentem como bons títulos para os artigos (i.e., que sejam muito semelhantes aos títulos escolhidos pelos editores). Para a criação do corpus do *sapo.pt* apenas foram selecionadas notícias que tivessem no mínimo sete frases, número selecionado após o cálculo da média do número de frases nos textos de todo o corpus a que tivemos acesso. Após a seleção das notícias, o corpus do *sapo.pt* contém em média doze frases por documento e uma média de cento e quarenta e quatro palavras por notícia.

No segundo caso, foram usados os textos jornalísticos associados ao TeMário (sigla para *TExtos com suMÁRIOs*), um conjunto de dados criado originalmente em 2003 e posteriormente revisto em 2006, sendo que a tarefa a resolver consiste em gerar sumários para artigos jornalísticos individuais, semelhantes aos sumários criados por peritos humanos. O TeMário 2006 é um corpus de 150 textos jornalísticos na variante Brasileira do Português associados aos seus respetivos sumários (Maziero et al., 2007), construído para complementar o corpus TeMário original (Pardo & Rino, 2003), o qual contém 100 textos e sumários da mesma natureza. Ambos os corpora tiveram os seus resumos produzidos por especialistas humanos. As várias notícias associadas aos textos dos corpora TeMário contêm, em média, dezasseis frases e uma média de cento e trinta e cinco palavras por notícia, após a remoção de *stop-words*. Desta forma, no caso dos testes com estes dados, cada um dos métodos em estudo foi usado para selecionar trinta por cento do número de frases dos textos originais, para a construção dos sumários.

Em termos das métricas consideradas para a avaliação, temos que Lin (2004b) introduziu um conjunto de métricas denominadas *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), que deste então se tornaram a norma em termos de métricas para a avaliação automática de sistemas de sumarização. Estas métricas encontram-se implementadas num pacote de software[4] disponível livremente para a avaliação de sistemas de sumarização, o qual foi usado no contexto das nossas experiências.

Consideremos um conjunto de sumários de referência $R = \{r_1, \ldots, r_m\}$, e um sumário s gerado automaticamente. Consideremos ainda um vetor binário $\Phi_n(d)$ que representa os n-gramas contidos num documento d, onde o i-ésimo componente ϕ_n^i tem o valor 1 se o i-ésimo n-grama se encontra contido em d, e 0 caso contrário. A métrica ROUGE-N apresenta-se como uma estatística que captura a cobertura dos n-gramas, sendo calculada da seguinte forma:

$$\text{ROUGE-N}(s) = \frac{\sum_{r \in R} \langle \Phi_n(d), \Phi_n(s) \rangle}{\sum_{r \in R} \langle \Phi_n(d), \Phi_n(r) \rangle} \quad (10)$$

Na fórmula acima, $\langle ., . \rangle$ representa a definição usual para o produto interno de vetores. A métrica ROUGE-N, tal como definida acima, pode ser usada em cenários de avaliação em que existam múltiplos sumários de referência, muito embora a mesma também possa tomar apenas o sumário de referência mais similar para com o sumário gerado automaticamente:

$$\text{ROUGE-N}_{multi}(s) = \max_{r \in R} \frac{\langle \Phi_n(d), \Phi_n(s) \rangle}{\langle \Phi_n(d), \Phi_n(r) \rangle} \quad (11)$$

Uma outra métrica proposta por Lin (2004b) aplica o conceito da sub-sequência comum mais longa (LCS). O racional por detrás desta ideia prende-se com o facto de que quanto maior for a

[4] http://www.berouge.com/

LCS entre duas frases de sumários, maior a similaridade entre elas. Consideremos um conjunto de frases de referência $r_1, \ldots, r_u$ para os documentos em R, e consideremos um sumário candidato s (i.e., s corresponde à concatenação de frases extraídas por um sumarizador extrativo). A métrica ROUGE-L é definida como uma média harmónica calculada com base na LCS, tal como apresentada na seguinte equação:

$$\text{ROUGE-L}(s) = \frac{(1+\beta^2) \times \text{R}_{\text{LCS}}(s) \times \text{P}_{\text{LCS}}(s)}{\text{R}_{\text{LCS}}(s) + \beta^2 \times \text{P}_{\text{LCS}}(s)} \quad (12)$$

Na fórmula acima,

$$\text{R}_{\text{LCS}}(s) = \frac{\sum_{i=1}^{u} \text{LCS}(r_i, s)}{\sum_{i=1}^{u} |r_i|}$$

e

$$\text{P}_{\text{LCS}}(s) = \frac{\sum_{i=1}^{u} \text{LCS}(r_i, s)}{|s|},$$

sendo que $|x|$ denota o tamanho de uma frase x, e $\text{LCS}(x, y)$ denota o tamanho da sub-sequência comum mais longa entre as frases x e y. O parâmetro real β controla o balanceamento entre os componentes $\text{R}_{\text{LCS}}(s)$ e $\text{P}_{\text{LCS}}(s)$, tomando normalmente o valor de 1 (neste caso, a medida ROUGE-L corresponde exatamente a uma média harmónica). A função $\text{LCS}(x, y)$ pode ser calculada através de uma abordagem de programação dinâmica.

Uma outra métrica também introduzida por Lin (2004b) é a ROUGE-S, a qual pode ser vista como uma versão com intervalos da métrica ROUGE-N, com $N = 2$ (i.e., uma métrica baseada em *skip bi-grams*). Consideremos um vetor binário $\Psi_2(d)$ indexado por pares ordenados de palavras, onde o componente $\psi_2^i(d)$ toma o valor 1 caso o i-ésimo par seja uma sub-sequência de palavras existente em d, e 0 caso contrário. A métrica ROUGE-S pode ser calculada como:

$$\text{ROUGE-S}(s) = \frac{(1+\beta^2) \times \text{R}_{\text{S}}(s) \times \text{P}_{\text{S}}(s)}{\text{R}_{\text{S}}(s) + \beta^2 \times \text{P}_{\text{S}}(s)} \quad (13)$$

Na fórmula acima, temos que o parâmetro

$$\text{R}_{\text{S}}(s) = \frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(r_i) \rangle},$$

enquanto que o parâmetro

$$\text{P}_{\text{S}}(s) = \frac{\sum_{i=1}^{u} \langle \Psi_2(r_i), \Psi_2(s) \rangle}{\langle \Psi_2(s), \Psi_2(s) \rangle}.$$

Nos nossos testes, foi usada uma versão estendida do ROUGE-S denominada ROUGE-SU, com o número de *skip bi-grams* = 4 e que considera também a sequência de uni-gramas. A métrica ROUGE-SU pode ser obtida através da métrica ROUGE-S, ao adicionar marcadores de início e fim nas frases candidatas e de referência.

As várias versões da medida ROUGE foram avaliadas no passado, medindo a correlação para com avaliações produzidas por peritos humanos (Lin, 2004a,b). A variante ROUGE-2 apresenta-se como a melhor de entre as várias variantes da ROUGE-N, e as medidas ROUGE-L e ROUGE-SU todas apresentaram bons resultados. Para as experiências efetuadas no contexto deste artigo, são apresentados resultados sob todas estas métricas, considerando $N = 1, 2$, para o caso da métrica ROUGE-N por serem estes os valores mais usados na área, e considerando *skip bi-grams* de tamanho 4, para o caso da métrica ROUGE-SU.

Para as abordagens baseadas em decomposição de matrizes ou baseadas em grafos, testadas neste estudo, compararam-se diferentes métodos de pesagem dos termos contidos nos documentos, tal como especificados na Tabela 1.

Quanto às abordagens baseadas no algoritmo TextRank, foram testados diversos valores relativamente ao parâmetro correspondente ao número de iterações (i.e., 100, 200, 300, 400, 500). No entanto, não se obtiveram alterações significativas nos resultados, tendo sido selecionado o número mínimo de iterações testado (i.e., 100) para a apresentação dos resultados neste estudo. Nas Figuras 1 e 2 apresentamos os resultados para cada um dos métodos baseados em grafos que foram considerados, para ambos os corpora.

Para o caso dos métodos baseados nas ocorrências de bi-gramas frequentes, testaram-se diferentes números de bi-gramas k para a construção da lista de k bi-gramas mais importantes, aquando da geração dos sumários. Os resultados são apresentados na Figura 3.

No caso dos métodos baseados na decomposição SVD, foram considerados os diferentes pesos para os termos, mencionados anteriormente e descritos na Tabela 1. As Tabelas 2 e 3 mostram os resultados dos testes efetuados com as diferentes implementações para a seleção das melhores frases após a decomposição da matriz em valores singulares, destacando os melhores resultados para cada métrica ROUGE.

No caso dos métodos baseados na decomposição de matrizes não negativas, foram consideradas apenas 10 iterações para efetuar a decomposição da matriz. Foram testadas diver-

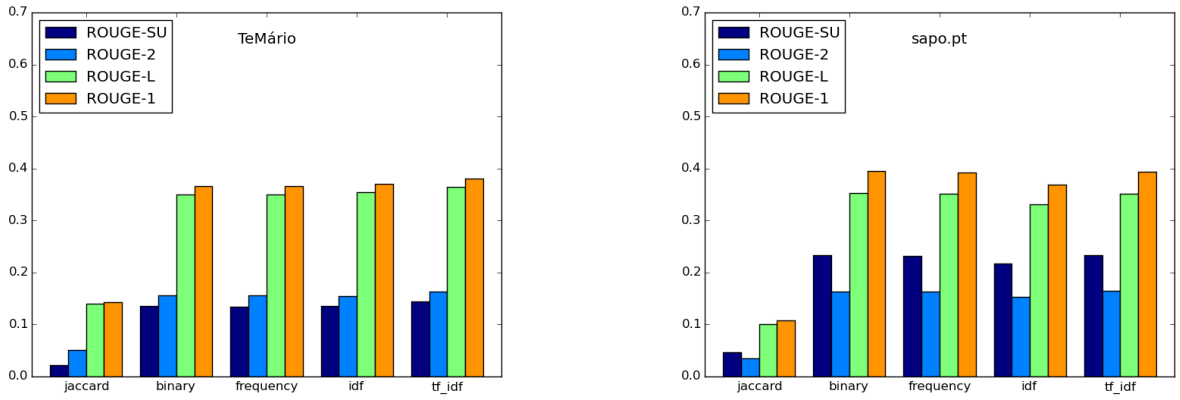

Figura 1: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com probabilidades iniciais uniformes.

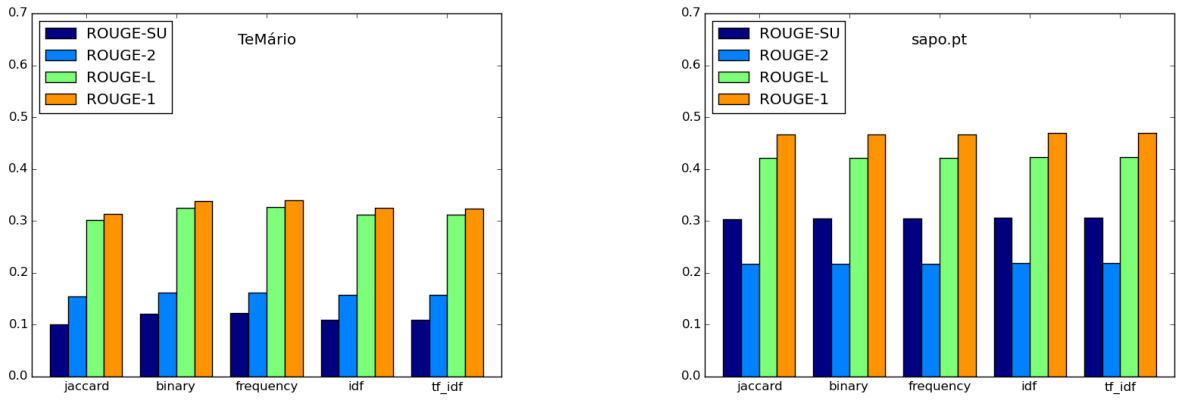

Figura 2: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com probabilidades iniciais não-uniformes.

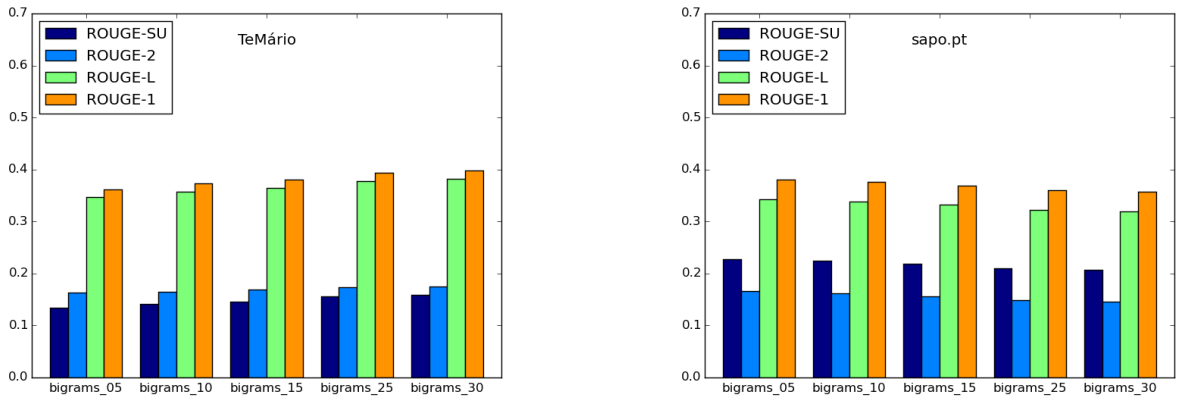

Figura 3: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando a abordagem baseada no número de bi-gramas da lista de k mais frequentes.

Peso	Fórmula
Peso Binário (BN)	$\text{BN}(j,i) = 0\|1$
Frequência (TF)	$\text{TF}(j,i) =$ frequência do termo i no documento j
Número Inverso de Frases (IDF)	$\text{IDF}(j,i) = \log(\text{nº de frases / nº de frases com o termo } i)$
TF-IDF	$\text{TF-IDF}(j,i) = \text{TF}(j,i) \times \text{IDF}(j,i)$

Tabela 1: Pesos usados nas abordagens baseadas em decomposição de matrizes ou baseadas em grafos.

	LSA Classic				**LSA Squared**			
	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	0.3695	**0.4249**	0.4237	0.3719	0.2266	0.2821	**0.2882**	0.2325
ROUGE-2	0.1608	**0.1941**	0.1938	0.1634	0.0651	0.0918	**0.1009**	0.0716
ROUGE-L	0.3319	**0.3820**	0.3814	0.3341	0.2008	0.2487	**0.2565**	0.2072
ROUGE-SU	0.2295	**0.2739**	0.2732	0.2323	0.1033	0.1400	**0.1499**	0.1103

Tabela 2: Resultados para os diferentes esquemas de pesagem de termos, nos métodos baseados em decomposição de matrizes em valores singulares, para o corpus *sapo.pt.*

	LSA Classic				**LSA Squared**			
	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	0.3865	**0.4116**	0.4076	0.3860	0.4117	**0.4369**	0.4352	0.4093
ROUGE-2	0.1491	**0.1658**	0.1633	0.1486	0.1604	**0.1768**	0.1759	0.1581
ROUGE-L	0.3692	**0.3942**	0.3903	0.3700	0.3930	**0.4179**	0.4166	0.3909
ROUGE-SU	0.1422	**0.1629**	0.1612	0.1424	0.1595	**0.1832**	0.1810	0.1582

Tabela 3: Resultados para os diferentes esquemas de pesagem de termos, nos métodos baseados em decomposição de matrizes em valores singulares, para o corpus TeMário.

sas dimensionalidades para o número de tópicos usados na decomposição da matriz, tendo sido testadas as dimensões de 25%, 50%, 75% e 100% do tamanho da matriz inicial. Os resultados estão exibidos nas Tabelas 4 e 5. Quando os valores entre os tópicos são iguais, selecionámos como melhor implementação o algoritmo que continha o número de tópicos mais pequeno.

5 Análise dos Resultados

A Tabela 6 mostra os resultados obtidos após a execução dos diferentes tipos de algoritmos nos dois corpora testados. Os melhores resultados com todos os tipos de algoritmos estão assim em destaque na Tabela 6.

Os resultados obtidos demonstram que os métodos baseados na seleção da primeira frase de uma notícia obtêm melhores resultados na construção de títulos de forma extrativa, em termos das várias métricas ROUGE testadas, conforme referenciado na Tabela 6. Na Figura 4 apresentam-se as distribuições para os valores da similaridade de Jaccard entre as frases que se encontram nas primeiras n posições dos documentos do corpus *sapo.pt*, e a frase que constitui o título do documento respetivo. Como se pode ver, as frases nas primeiras posições são claramente mais semelhantes para com os títulos, justificando-se desta forma os bons resultados obtidos através desta *baseline* muito simples.

Para a geração de sumários de notícias, a abordagem que apresentou melhores resultados foi a implementação do algoritmo LSA Squared utilizando um peso binário para os termos, em relação a todas as métricas ROUGE.

Efetuando uma análise aos algoritmos que utilizaram SVD, na implementação LSA Classic obteve-se os melhores resultados utilizando uma representação binária de cada termo nos documentos. Quanto à implementação LSA Squared, os resultados foram diferentes para ambos os corpus testados, tendo sido obtidos melhores resultados no corpus do *sapo.pt* com a contabilização da frequência de cada palavra nos respetivos documentos. Quanto ao corpus TeMário, os melhores resultados foram obtidos utilizando uma representação binária das palavras nos respetivos documentos, para ambos os algoritmos. Comparando ambas as implementações, a implementação LSA Classic obteve melhores resultados na geração de títulos de notícias, enquanto a implementação LSA Squared obteve melhores resultados na geração de sumários com mais do que uma frase.

		NMF GR				**NMF ExtR**			
ROUGE	%Frases	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	25%	0.2230	**0.2987**	0.2965	0.2264	0.2286	**0.3060**	0.3056	0.2338
	50%	0.2091	0.2679	0.2671	0.2264	0.2259	0.2876	0.2919	0.2301
	75%	0.1934	0.2507	0.2535	0.1934	0.2238	0.2799	0.2853	0.2289
	100%	0.1805	0.2286	0.2339	0.1814	0.2226	0.2720	0.2783	0.2264
ROUGE-2	25%	0.0660	0.1037	**0.1064**	0.0701	0.0661	0.1048	**0.1111**	0.0722
	50%	0.0641	0.0912	0.0926	0.0650	0.0661	0.0960	0.1037	0.0713
	75%	0.0601	0.0842	0.0872	0.0603	0.0645	0.0920	0.1004	0.0703
	100%	0.0567	0.0765	0.0800	0.0572	0.0652	0.0877	0.0964	0.0696
ROUGE-L	25%	0.1982	**0.2645**	0.2642	0.2021	0.2026	0.2693	**0.2721**	0.2084
	50%	0.1883	0.2391	0.2396	0.1886	0.2009	0.2543	0.2596	0.2055
	75%	0.1742	0.2246	0.2275	0.1740	0.1988	0.2471	0.2543	0.2043
	100%	0.1635	0.2051	0.2105	0.1645	0.1978	0.2406	0.2481	0.2020
ROUGE-SU	25%	0.1029	0.1552	**0.1566**	0.1072	0.1047	0.1578	**0.1629**	0.1113
	50%	0.0978	0.1364	0.1373	0.0985	0.1043	0.1455	0.1527	0.1097
	75%	0.0896	0.1254	0.1288	0.0896	0.1025	0.1397	0.1482	0.1084
	100%	0.0834	0.1119	0.1171	0.0841	0.1022	0.1348	0.1433	0.1073

Tabela 4: Resultados para as várias medidas ROUGE quando considerando diferentes números de variáveis latentes, nos métodos utilizando a decomposição de matrizes não negativas e sobre o corpus *sapo.pt.*

		NMF GR				**NMF ExtR**			
ROUGE	%Frases	IDF	Bin.	TF	TF-IDF	IDF	Bin.	TF	TF-IDF
ROUGE-1	25%	0.3492	0.3534	0.3520	0.3496	0.3590	0.3636	0.3649	0.3606
	50%	0.3576	**0.3686**	0.3679	0.3566	0.4044	0.3927	0.3906	0.4024
	75%	0.3317	0.3579	0.3572	0.3372	**0.4049**	0.3825	0.3833	0.4029
	100%	0.3333	0.3519	0.3481	0.3334	0.3992	0.3772	0.3734	0.3997
ROUGE-2	25%	0.1312	0.1392	0.1373	0.1310	0.1346	0.1446	0.1453	0.1350
	50%	0.1473	**0.1493**	0.1473	0.1309	0.1573	**0.1595**	0.1566	0.1560
	75%	0.1225	0.1419	0.1419	0.1259	0.1559	0.1549	0.1550	0.1548
	100%	0.1255	0.1397	0.1367	0.1247	0.1535	0.1519	0.1492	0.1531
ROUGE-L	25%	0.3330	0.3374	0.3358	0.3338	0.3419	0.3467	0.3480	0.3438
	50%	0.3418	**0.3526**	0.3512	0.3407	0.3864	0.3757	0.3735	0.3846
	75%	0.3179	0.3423	0.3420	0.3233	**0.3871**	0.3659	0.3670	0.3852
	100%	0.3183	0.3359	0.3322	0.3188	0.3814	0.3605	0.3563	0.3816
ROUGE-SU	25%	0.1163	0.1204	0.1199	0.1166	0.1224	0.1279	0.1284	0.1236
	50%	0.1220	**0.1333**	0.1324	0.1209	**0.1546**	0.1495	0.1481	0.1538
	75%	0.1066	0.1243	0.1246	0.1103	0.1546	0.1440	0.1445	0.1531
	100%	0.1081	0.1210	0.1186	0.1079	0.1509	0.1395	0.1367	0.1509

Tabela 5: Resultados para as várias medidas ROUGE quando considerando diferentes números de variáveis latentes, nos métodos utilizando a decomposição de matrizes não negativas e sobre o corpus TeMário.

	Textos Jornalísticos PT-PT				**Textos PT-BR no Corpus TeMário**			
	R-1	R-2	R-L	R-SU	R-1	R-2	R-L	R-SU
TextRank	0.3955	0.1646	0.3528	0.2339	0.3807	0.1635	0.3645	0.1443
TextRank Init	0.4699	0.2185	0.4236	0.3058	0.3400	0.1620	0.3269	0.1222
LSA Classic	0.4249	0.1941	0.3820	0.2739	0.4116	0.1658	0.3942	0.1629
LSA Squared	0.2882	0.1941	0.2565	0.1499	**0.4369**	**0.1768**	**0.4179**	**0.1832**
NMF ExtR	0.3060	0.1111	0.2721	0.1629	0.4049	0.1595	0.3871	0.1546
NMF GR	0.2987	0.1064	0.2645	0.1566	0.3686	0.1493	0.3526	0.1333
Primeiras Frase(s)	**0.4701**	**0.2186**	**0.4238**	**0.3060**	0.3307	0.1620	0.3183	0.1107
Frase(s) Aleatória(s)	0.2893	0.1062	0.2770	0.0819	0.1828	0.0586	0.1656	0.0856
Bi-gramas	0.3802	0.1667	0.3420	0.2282	0.3980	0.1757	0.3815	0.1586

Tabela 6: Resultados obtidos pelos vários métodos e para ambos os corpora.

Quanto aos algoritmos baseados em decomposição de matrizes não negativas, nomeadamente a implementação NMF-GR, os resultados também foram distintos para cada um dos corpora. No corpus *sapo.pt*, as melhores implementações, envolvem a utilização de 25% do número máximo de variáveis latentes a selecionar, utilizando representações binárias das palavras nas frases e a contabilização da frequência de cada palavra nos respetivos documentos, para diferentes medidas ROUGE.

Quanto ao corpus TeMário, os melhores resultados foram obtidos usando uma representação binária, com 50% quanto ao número de variáveis latentes a usar na decomposição da matriz. Na implementação NMF-Extr, os resultados também foram diferentes para ambos os corpora testados, tendo-se obtido melhores resultados, para quase todas as medidas ROUGE, com a utilização da frequência das palavras nos respetivos documentos, com 25% no número de variáveis latentes, no corpus *sapo.pt*. No corpus do TeMário, os melhores resultados foram obtidos usando uma representação baseada no número inverso de frases, e com 75% no número de variáveis latentes a selecionar, embora não para todas as métricas ROUGE. Comparando ambas as implementações, a implementação NMF-Extr obteve melhores resultados em ambos os corpora.

Os testes efetuados com as duas variantes do algoritmo TextRank também demonstraram resultados diferentes para cada um dos corpora. Quanto à seleção de títulos de notícias, os melhores resultados foram obtidos utilizando a métrica IDF para todas as métricas ROUGE, e para seleção das melhores frases os melhores resultados foram obtidos utilizando a métrica TF para todas as métricas ROUGE. A implementação do algoritmo TextRank com probabilidades iniciais não-uniformes obteve piores resultados no corpus do TeMário e obteve melhores resultados no corpus do *sapo.pt*, tendo resultados aproximados com a abordagem de seleção da primeira frase.

Num teste separado, tentámos verificar qual a influência que o parâmetro α da Equação 9, o qual controla o decaimento da importância dada às frases que ocorrem nas posições iniciais do documento a sumarizar, tem sobre os resultados finais. Os dois gráficos da Figura 5 ilustram os resultados obtidos sobre as coleções TeMário e *sapo.pt*. Os melhores resultados correspondem aos valores $\alpha = 0.85$ no caso do *sapo.pt*, e $\alpha = 0.25$ no caso do TeMário, embora as variações nos resultados sejam muito pequenas.

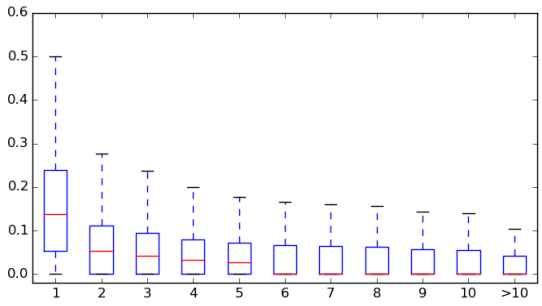

Figura 4: Distribuições para os valores da similaridade de Jaccard entre as frases do corpus *sapo.pt* que se encontram nas primeiras n posições, para com os títulos dos respetivos documentos.

Para os testes efetuados com os diferentes números de bi-gramas no método *baseline*, também se obtiveram resultados diferentes em ambos os corpora. Para a seleção dos títulos de notícias, a abordagem que obteve melhores resultados seleciona as frases que continham bigramas dos cinco bi-gramas mais frequentes, enquanto para a seleção das melhores frases para um sumário, a abordagem que obteve melhores resultados baseia-se na seleção das frases que continham o maior número de bi-gramas possíveis dentro dos trinta bi-gramas mais frequentes.

6 Conclusões e Trabalho Futuro

Este artigo apresentou uma comparação sistemática de diferentes abordagens extrativas para a tarefa de sumarizar documentos individuais correspondendo a textos jornalísticos escritos em Português. Através da utilização da bancada ROUGE como forma de medir a qualidade dos sumários produzidos, foram reportados resultados para dois domínios experimentais diferentes, envolvendo (i) a geração de títulos para textos jornalísticos escritos na variante Europeia do Português, e (ii) a geração de sumários com base em artigos jornalísticos escritos na variante Brasileira do Português. Os resultados obtidos demonstram que métodos heurísticos simples, baseados na seleção da primeira frase de uma notícia, obtêm melhores resultados na construção de títulos de forma extrativa, em termos de várias métricas ROUGE. Para a geração de sumários mais longos do que uma frase, o método que obteve melhores resultados foi o método LSA Squared, baseado na decomposição SVD de uma matriz de termos por frases.

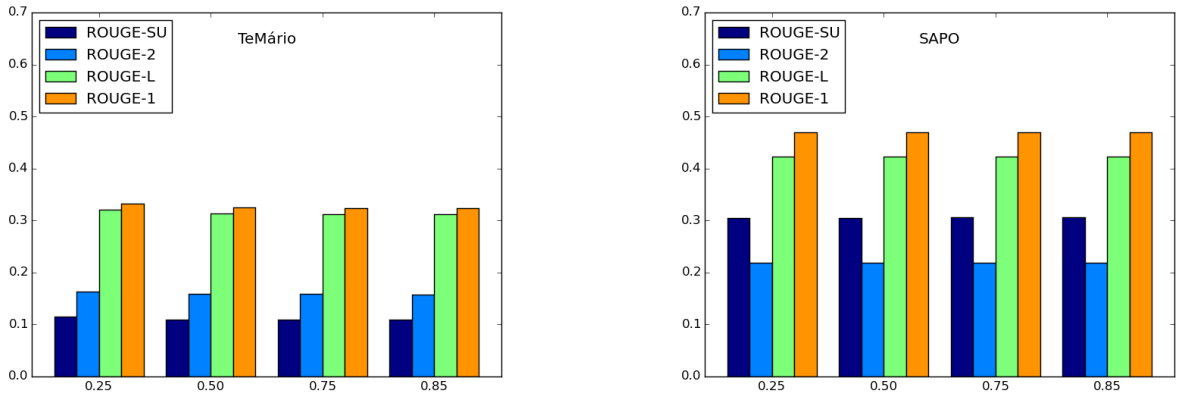

Figura 5: Resultados para os corpora TeMário (figura à esquerda) e *sapo.pt* (figura à direita), para as várias métricas ROUGE e quando usando o algoritmo TextRank com representações baseadas em TF-IDF, com probabilidades iniciais não-uniformes e com diferentes valores para o parâmetro α.

Para trabalho futuro, planeamos dar continuidade ao trabalho apresentado neste artigo, particularmente focando no problema da geração de títulos para artigos jornalísticos, ambicionando a integração de um módulo de sumarização automática, com estas características, num sistema de recomendação de notícias. Nesta aplicação em concreto, pretende-se abordar a geração de sumários curtos que não só capturem os aspetos mais importantes dos artigos, mas que também sejam personalizados em função dos interesses individuais dos utilizadores do sistema de recomendação, e que possam aumentar o rácio de cliques nas notícias apresentadas. Este é um problema muito importante no contexto de portais de notícias on-line, tais como o do serviço *sapo.pt*.

Pensamos que um método puramente extrativo terá sempre muitas limitações na aplicação concreta à geração de títulos para artigos jornalísticos e, como tal, planeamos integrar métodos de resolução de anáforas e de co-referências nas etapas de pré-processamento, por forma a poder enriquecer as frases antes de um processo de seleção para a formação de sumários. Planeamos também efetuar testes com outros métodos baseados em grafos e em adaptações do algoritmo PageRank, na tarefa de sumarização. Em particular, pensamos testar representações dos textos baseadas em grafos bi-partidos, em que os nós correspondentes a frases se interliguem com nós representando diferentes tipos de conceitos (e.g., termos individuais, entidades mencionadas, tópicos latentes, etc.) extraídos dos documentos textuais a sumarizar.

No passado, autores como Banko et al. (2000), Dorr et al. (2003) ou Alfonseca et al. (2013) abordaram já o desenvolvimento de abordagens específicas para a geração de títulos de artigos jornalísticos, indo além da sumarização extractiva. Nós pretendemos combinar abordagens para a sumarização extractiva e para a compressão de frases (e.g., consultar os trabalhos da autoria de Berg-Kirkpatrick et al. (2011) e de Almeida & Martins (2013) como exemplos recentes de abordagens deste tipo), abordando desta forma a geração de bons títulos para os artigos jornalísticos a apresentar num portal *on-line*.

Referências

Alfonseca, Enrique, Daniele Pighin & Guillermo Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1243–1253.

Almeida, Miguel B. & André F. T Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. Em *Proceedings of the Annual Meeting of the Association for Computer Linguistics*, 196–206.

Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky & Bjornar Larsen. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. Em Inderjeet Mani & Mark T. Maybury (eds.), *Advances in Automatic Text Summarization*, MIT Press.

Bach, Nguyen, Qin Gao, Stephan Vogel & Alex Waibel. 2011. TriS: A statistical sentence simplifier with log-linear models and margin-based discriminative training. Em *Proceedings of the International Joint Conference on Natural Language Processing*, 474–482.

Banko, Michele, Vibhu O. Mittal & Michael J. Witbrock. 2000. Headline generation based on statistical translation. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 318–325.

Barzilay, Regina, Kathleen R McKeown & Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. Em *Proceedings of the Annual Meeting of the Association for Computer Linguistics*, 550–557!

Baxendale, Phyllis B. 1958. Machine-made index for technical literature: An experiment. *IBM Journal of Research and Development* 2(4).

Berg-Kirkpatrick, Taylor, Dan Gillick & Dan Klein. 2011. Jointly learning to extract and compress. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 481–490.

Carbonell, Jaime & Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336.

Conroy, John M & Dianne P O'Leary. 2001. Text summarization via Hidden Markov Models. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 406–407.

Coster, William & David Kauchak. 2011. Learning to simplify sentences using Wikipedia. Em *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 1–9.

Dorr, Bonnie, David Zajic & Richard Schwartz. 2003. Hedge trimmer: a parse-and-trim approach to headline generation. Em *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 1–8.

Edmundson, Harold P. 1969. New methods in automatic extracting. *Journal of the ACM* 16(2).

Erkan, Günes & Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22(1).

Franceschet, Massimo. 2011. PageRank: standing on the shoulders of giants. *Communications of the ACM* 54(6).

Fung, Pascale & Grace Ngai. 2006. One story, one flow: Hidden Markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing* 3(2).

Gong, Yihong & Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19–25.

Hirao, Tsutomu, Jun Suzuki & Hideki Isozaki. 2009. Automatic summarization as a combinatorial optimization problem. *Transactions of the Japanese Society for Artificial Intelligence* 24(2).

Jing, Hongyan & Kathleen R McKeown. 2000. Cut and paste based text summarization. Em *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics Conference*, 178–185.

Knight, Kevin & Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence* 139(1).

Kupiec, Julian, Jan Pedersen & Francine Chen. 1995. A trainable document summarizer. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68–73.

Lee, Daniel D. & H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755).

Lee, Ju-Hong, Sun Park, Chan-Min Ahn & Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management* 45(1).

Lin, Chin-Yew. 1999. Training a selection function for extraction. Em *Proceedings of the International Conference on Information and Knowledge Management*, 55–62.

Lin, Chin-Yew. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? Em *Proceedings of the NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization*, s.p.

Lin, Chin-Yew. 2004b. ROUGE: a package for automatic evaluation of summaries. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 25–26.

Litvak, Marina, Mark Last & Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 927–936.

Luhn, Hans P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2).

Mani, Inderjeet, Gary Klein, David House, Lynette Hirschman, Therese Firmin & Beth Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering* 8(01). 43–68.

Mashechkin, Igor, Mikhail Petrovskiy, D.S. Popov & Dmitry V. Tsarev. 2011. Automatic text summarization using latent semantic analysis. *Programming and Computer Software* 37(6).

Maziero, Erick Galani, Vinícius Rodrigues Uzêda, Tiago Salgueiro Pardo & Maria das Graças Volpe Nunes. 2007. TeMário 2006: Estendendo o córpus TeMário. Relatório Técnico. NILC-TR-07-06 Núcleo Interinstitucional de Linguística Computacional.

McKeown, Kathleen & Dragomir R Radev. 1995. Generating summaries of multiple news articles. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74–82.

Mihalcea, Rada. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Em *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, s.p.

Mihalcea, Rada & Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization. Em *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing*, 19–24.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11).

Nenkova, Ani, Sameer Maskey & Yang Liu. 2011. Automatic summarization. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, s.p.

Osborne, Miles. 2002. Using maximum entropy for sentence extraction. Em *Proceedings of the Workshop on Automatic Summarization*, 1–8.

Page, Lawrence, Sergey Brin, Rajeev Motwani & Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Relatório Técnico. 1999-66 Stanford InfoLab.

Pardo, Thiago Alexandre Salgueiro. 2008. Sumarização automática: Principais conceitos e sistemas para o português brasileiro. Relatório Técnico. NILC-TR-08-04 Núcleo Interinstitucional de Linguística Computacional.

Pardo, Thiago Alexandre Salgueiro & Lucia Helena Machado Rino. 2003. TeMário: Um corpus para sumarização automática de textos. Relatório Técnico. NILC-TR-03-09 Núcleo Interinstitucional de Linguística Computacional.

Radev, Dragomir R, Eduard Hovy & Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics* 28(4).

Radev, Dragomir R, Hongyan Jing & Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. Em *Proceedings of the Workshop on Automatic Summarization*, 21–30.

Salton, Gerard, Amit Singhal, Mandar Mitra & Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing & Management* 33(2).

Shen, Dou, Jian-Tao Sun, Hua Li, Qiang Yang & Zheng Chen. 2007. Document summarization using conditional random fields. Em *Proceedings of the International Joint Conference on Artifical Intelligence*, 2862–2867.

Siddharthan, Advaith & Kathleen McKeown. 2005. Improving multilingual summarization: using redundancy in the input to correct MT errors. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 33–40.

Steinberger, Josef & Karel Ježek. 2004. Text summarization and singular value decomposition. Em *Proceedings of the International Conference on Advances in Information Systems*, 245–254.

Svore, Krysta Marie, Lucy Vanderwende & Christopher J.C. Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 448–457.

Thapar, Vishal, Ahmed A Mohamed & Sanguthevar Rajasekaran. 2006. Consensus text

summarizer based on meta-search algorithms. Em *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, 403–407.

Wan, Xiaojun & Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299–306.

Wang, Dingding & Tao Li. 2010. Many are better than one: Improving multi-document summarization via weighted consensus. Em *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 809–810.

Yamangil, Elif & Rani Nelken. 2008. Mining Wikipedia revision histories for improving sentence compression. Em *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 137–140.

Yeh, Jen-Yuan, Hao-Ren Ke, Wei-Pang Yang & I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management* 41(1). 75–95.