Métodos de Otimização Combinatória Aplicados ao Problema de Compressão MultiFrases

Elvys Linhares Pontes, Thiago Gouveia da Silva, Andréa Carneiro Linhares, Juan-Manuel Torres-Moreno, Stéphane Huet

Métodos de Otimização Combinatória Aplicados ao Problema de Compressão MultiFrases

Elvys Linhares Pontes[1], Thiago Gouveia da Silva[1,2,3], Andréa Carneiro Linhares[5], Juan-Manuel Torres-Moreno[1,4], Stéphane Huet[1]

[1]LIA/CERI – Université d'Avignon et Pays de Vaucluse (UAPV), Avignon – France

[2]Inst. Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), PB – Brasil

[3]Instituto de Computação – Univ. Federal Fluminense (UFF), Niterói – RJ – Brasil

[4]École Polytechnique de Montréal, Montréal, Canada

[5]Universidade Federal do Ceará (UFC), Sobral – CE – Brasil

elvys.linhares-pontes@alumni.univ-avignon.com, thiago.gouveia@ifpb.edu.br

RESUMO

A Internet possibilitou o aumento considerável da quantidade de informação disponível. Nesse contexto, a leitura e o entendimento desse fluxo de informações tornaram-se tarefas dispendiosas. Ao longo dos últimos anos, com o intuito de ajudar a compreensão dos dados textuais, várias aplicações da área de Processamento de Linguagem Natural (PLN) baseando-se em métodos de Otimização Combinatória vem sendo implementadas. Contundo, para a Compressão MultiFrases (CMF), técnica que reduz o tamanho de uma frase sem remover as principais informações nela contidas, a inserção de métodos de otimização necessita de um maior estudo a fim de melhorar a performance da CMF. Este artigo descreve um método de CMF utilizando a Otimização Combinatória e a Teoria dos Grafos para gerar frases mais informativas mantendo a gramaticalidade das mesmas. Um experimento com 40 *clusters* de frases comprova que nosso sistema obteve uma ótima qualidade e foi melhor que o estado da arte.

PALAVRAS CHAVE. Otimização Combinatória, Compressão MultiFrases, Grafo de Palavras.

ABSTRACT

The Internet has led to a dramatic increase in the amount of available information. In this context, reading and understanding this flow of information have become costly tasks. In the last years, to assist people to understand textual data, various Natural Language Processing (NLP) applications based on Combinatorial Optimization have been devised. However, for Multi-Sentences Compression (MSC), method which reduces the sentence length without removing core information, the insertion of optimization methods requires further study to improve the performance of MSC. This article describes a method for MSC using Combinatorial Optimization and Graph Theory to generate more informative sentences while maintaining their grammaticality. An experiment led on a corpus of 40 clusters of sentences shows that our system has achieved a very good quality and is better than the state-of-the-art.

KEYWORDS. Combinatorial Optimization, Multi-Sentences Compression, Word Graph.

1. Introdução

O aumento da quantidade de dispositivos eletrônicos (*smartphones*, *tablets*, etc) e da Internet móvel tornaram o acesso à informação fácil e rápido. Através da Internet é possível ter acesso aos acontecimentos de todo o mundo a partir de diferentes *sites*, *blogs* e portais. Páginas como a Wikipédia e portais de notícias fornecem informações detalhadas sobre diversas temáticas, entretanto os textos são longos e possuem muitas informações irrelevantes. Uma solução para esse problema é a geração de resumos contendo as principais informações do documento e sem redundâncias (Linhares Pontes *et al.*, 2014). Vista a vasta quantidade e o fácil acesso às informações, é possível automatizar a análise e geração de resumos a partir da análise estatística, morfológica e sintática das frases (Torres-Moreno, 2014).

O Processamento da Linguagem Natural (PLN) concerne à aplicação de sistemas e técnicas de informática para analisar a linguagem humana. Dentre as diversas aplicações do PLN (tradução automática, compressão textual, etc.), a Sumarização Automática de Textos (SAT) consiste em resumir um ou mais textos automaticamente. O sistema sumarizador identifica os dados relevantes e cria um resumo a partir das principais informações (Linhares Pontes *et al.*, 2015). A CMF é um dos métodos utilizados na SAT para gerar resumos, que utiliza um conjunto de frases para gerar uma única frase de tamanho reduzido gramaticalmente correta e informativa (Filippova, 2010; Boudin e Morin, 2013).

Neste artigo, apresentamos um método baseado na Teoria dos Grafos e na Otimização Combinatória para modelar um documento como um Grafo de Palavras (GP) (Filippova, 2010) e gerar a CMF com uma melhor qualidade informativa.

A seção 2 descreve o problema e os trabalhos relacionados à CMF. Detalhamos a abordagem e a modelagem matemática nas seções 3 e 4, respectivamente. O corpus, as ferramentas utilizadas e os resultados obtidos são discutidos na seção 5. Finalmente, as conclusões e os comentários finais são expostos na seção 6.

2. Compressão MultiFrases

A Compressão MultiFrases (CMF) consiste em produzir uma frase de tamanho reduzido gramaticalmente correta a partir de um conjunto de frases oriundas de um documento, preservando-se as principais informações desse conjunto. Uma compressão pode ter diferentes valores de Taxa de Compressão (TC), entretanto quanto menor a TC maior será a redução das informações nele contidas. Seja D o documento analisado composto pelas frases $\{f_1, f_2, \ldots, f_n\}$ e $frase_{CMF}$ a compressão desse documento, a TC é definida por:

$$TC = \frac{||frase_{CMF}||}{\sum_{i=1}^{n} \frac{||f_i||}{n}}. \quad (1)$$

onde $||f_i||$ é o tamanho da frase f_i (quantidade de palavras). Dessa forma, os principais desafios da CMF são a seleção dos conteúdos informativos e a legibilidade da frase produzida.

Dentre as diversas abordagens feitas sobre a CMF, algumas baseiam-se em analisadores sintáticos para a produção de compressões gramaticais. Por exemplo, Barzilay e McKeown (2005) desenvolveram uma técnica de geração *text-to-text* em que cada frase

do texto é representada como uma árvore de dependência. De forma geral, essa técnica alinha e combina estas árvores para gerar a fusão das frases analisadas. Outra abordagem possível é descrita por Filippova (2010), que gerou compressões de frases de boa qualidade utilizando uma simples modelagem baseada na Teoria dos Grafos e uma lista de *stopwords*[1]. Boudin e Morin (2013) geraram a CMF mais informativas a partir da análise da relevância das frases geradas pelo método de Filippova.

Visto que os trabalhos apresentados utilizaram uma modelagem simples e obtiveram resultados de boa qualidade, este trabalho baseia-se na mesma modelagem utilizada por Filippova e métodos de otimização combinatória para aumentar a informatividade da CMF. As subseções 2.1 e 2.2 descrevem os métodos utilizados por Filippova (2010) e Boudin e Morin (2013), respectivamente.

2.1. Filippova

Filippova (2010) modelou um documento D composto por frases similares como um Grafo de Palavras (GP). O GP é um grafo direcionado $GP = (V, A)$, onde V é o conjunto de vértices (palavras) e A é o conjunto de arcos (relação de adjacência). Dessa forma, dado um documento D de frases similares $\{f_1, f_2, \ldots, f_n\}$, o GP é construído a partir da adição dessas frases no grafo. A Figura 1 ilustra o GP descrito por Filippova das seguintes frases:

a) George Solitário, a última tartaruga gigante Pinta Island do mundo, faleceu.
b) A tartaruga gigante conhecida como George Solitário morreu na segunda no Parque Nacional de Galapagos, Equador.
c) Ele tinha apenas cem anos de vida, mas a última tartaruga gigante Pinta conhecida, George Solitário, faleceu.
d) George Solitário, a última tartaruga gigante da sua espécie, morreu.

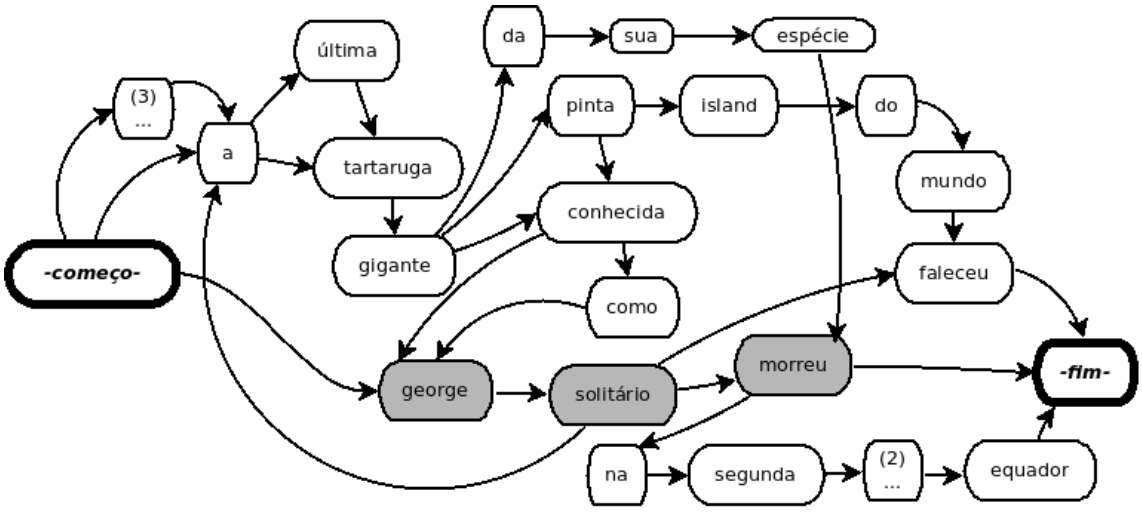

Figura 1. Grafo de palavras gerado a partir das frases a-d e um possível caminho representando a compressão (Filippova, 2010). Removemos as vírgulas das frases para facilitar a compreensibilidade do grafo.

Inicialmente, o GP é composto pela primeira frase (a) e pelos vértices *-começo-* e *-fim-*. Uma palavra é representada por um vértice existente somente se ela possuir a

[1] *Stopwords* são palavras comuns sem relevância informativa para uma frase. Ex: artigos, preposições, etc.

mesma forma minúscula, mesma *Part-Of-Speech* (POS)[2], e se não existir outra palavra dessa mesma frase que já tenha sido mapeada nesse vértice. Um novo vértice é criado caso não seja encontrado um vértice com suas características no GP. Dessa forma, cada frase representa um caminho simples entre os vértices -começo- e -fim-.

As frases são analisadas e adicionadas individualmente ao GP. Para cada frase analisada, as palavras são inseridas na seguinte ordem:

1. Palavras que não sejam *stopwords* e para os quais não existam nenhum candidato no grafo ou mapeamento não ambíguo;
2. Palavras que não sejam *stopwords* e para os quais existam vários candidatos possíveis no grafo ou que ocorram mais de uma vez na mesma frase;
3. *Stopwords*.

Nos grupos 2 e 3, o mapeamento das palavras é ambíguo, pois há mais de uma palavra no grafo que referencia a mesma palavra/POS. Nesse caso, as palavras predecessoras e posterioras são analisadas para verificar o contexto da palavra e escolher o mapeamento correto. Caso uma dessas palavras não possua o mesmo contexto das existentes no grafo, um novo vértice é criado para representá-la.

Tendo adicionado os vértices, os pesos dos arcos representam o nível de coesão entre as palavras de dois vértices a partir da frequência e da posição dessas palavras nas frases, conforme as Equações 2 e 3:

$$w(e_{i,j}) = \frac{\text{coesão}(e_{i,j})}{freq(i) \times freq(j)}, \tag{2}$$

$$\text{coesão}(e_{i,j}) = \frac{freq(i) + freq(j)}{\sum_{f \in D} dist(f,i,j)^{-1}}, \tag{3}$$

$$dist(f,i,j) = \begin{cases} pos(f,i) - pos(f,j) & \text{se } pos(f,i) < pos(f,j) \\ 0 & \text{caso contrário} \end{cases} \tag{4}$$

onde $freq(i)$ é a frequência da palavra mapeada no vértice i e a função $pos(f,i)$ retorna a posição da palavra i na frase f.

A partir do GP, o sistema calcula os 50 menores caminhos[3] que tenham no mínimo oito palavras e ao menos um verbo. Por fim, o sistema normaliza os *scores* (distâncias do caminhos) das frases geradas a partir do comprimento das mesmas e seleciona a frase com o menor *score* normalizado como a melhor CMF.

2.2. Boudin e Morin

Boudin e Morin (2013) (BM) propuseram um método para melhor avaliar a qualidade de uma frase e gerar compressões mais informativas a partir da abordagem descrita por Filippova (seção 2.1). BM utilizaram a mesma metodologia de Filippova para gerar os 200 menores caminhos, que tenham no mínimo oito palavras e ao menos um verbo, do

[2]POS é a classe gramatical de uma palavra numa frase.
[3]Ressaltando que cada caminho no GP representa uma frase.

GP. Ao invés de realizar uma simples normalização dos valores de cada frase como Filippova, BM mensuraram a relevância da frase gerada (caminho c no GP) a partir das *keyphrases*[4] e o comprimento das frases, conforme Equações 5 e 6:

$$score(c) = \frac{\sum_{i,j \in \text{caminho}(c)} w_{(i,j)}}{||c|| \times \sum_{k \in c} score_{kp}(k)}, \tag{5}$$

$$score_{kp}(k) = \frac{\sum_{w \in k} \text{TextRank}(w)}{||k|| + 1}, \tag{6}$$

onde $w_{(i,j)}$ é o score entre os vértices i e j descrito na Equação 2, o algoritmo TextRank (Mihalcea e Tarau, 2004) que calcula a relevância de uma palavra w no GP a partir das suas palavras predecessoras e posteriores e $score_{kp}(k)$ é a relevância da *keyphrase* k presente no caminho c. Por fim, a frase com o menor *score* é a escolhida para a compressão do texto.

3. Nova modelagem do problema

Os métodos de Filippova e BM calculam os menores caminhos do GP analisando somente o nível de coesão entre duas palavras vizinhas no texto. Após a geração dos caminhos, os *scores* de cada caminho são normalizados para escolher o "menor"deles. Entretanto, duas palavras possuindo uma forte coesão não significa que as mesmas possuam uma boa informatividade. Por mais que a normalização ou reanálise das frases seja eficiente, esses métodos estão sempre limitados às frases geradas pela análise do nível de coesão. Portanto, a geração de 50 ou 200 dos menores caminhos (frases) não garante a existência de uma frase com boa informatividade. Por isso, propomos um método para analisar concomitantemente a coesão e a relevância das palavras a fim de gerar uma compressão mais informativa de um documento.

O método aqui exposto visa calcular o caminho mais curto analisando a coesão das palavras e bonificando os caminhos que possuam palavras-chaves e *3-grams*[5] frequentes do texto. Inicialmente, utiliza-se a mesma abordagem de Filippova (seção 2.1) para modelar um documento D como um GP e calcular a coesão das palavras. Além de considerar a coesão, analisamos as palavras-chaves e os *3-grams* do documento para gerar uma CMF mais informativa. As palavras-chaves auxiliam a geração de caminhos com as principais informações descritas no texto. Como o documento D é composto por frases similares, consideramos que o documento possui somente uma temática. A *Latent Dirichlet Allocation* (LDA) é um método para analisar as frases de um texto e identificar o conjunto de palavras que representam as temáticas nele abordadas (Blei *et al.*, 2003). Configura-se o método LDA para identificar o conjunto de palavras que representa uma única temática do documento. Finalmente, esse conjunto de palavras constitui as palavras-chaves do documento D.

Uma outra consideração sobre o documento analisado é que a presença de uma palavra em diferentes frases aumenta sua relevância para a CMF (vale salientar que consideramos a relevância dos *stopwords* igual a zero). A partir da ponderação dos *2-grams*

[4] *Keyphrases* são as palavras que representam o conteúdo principal do texto.
[5] *3-gram* é formado por 3 palavras vizinhas.

(Equação 2), consideramos que a relevância de um *3-gram* é baseado na relevância dos dois *2-grams* que o formam, como descrito na Equação 7:

$$\text{3-gram}(i,j,k) = \frac{qt_3(i,j,k)}{\max_{a,b,c \in GP} qt_3(a,b,c)} \times \frac{w(e_{i,j}) + w(e_{j,k})}{2}, \tag{7}$$

onde $qt_3(i,j,k)$ é quantidade de *3-grams* composto pelas palavras dos vértices i, j e k no documento. Os *3-grams* auxiliam a geração de CMF com estruturas importantes para o texto e incrementam a qualidade gramatical das frases geradas.

O nosso sistema calcula os 50 menores caminhos do GP que possuam ao menos 8 palavras, baseado na coesão, palavras-chaves e *3-grams* (Equação 9). Contrariamente ao método de Filippova, as frases podem ter score negativo, pois reduzimos o valor do caminho composto por palavras-chaves e *3-grams*. Dessa forma, normalizamos os scores dos caminhos (frases) baseado na função exponencial para obter um score maior que zero, conforme a Equação 8:

$$score_{norm}(f) = \frac{e^{score_{opt}(f)}}{||f||}, \tag{8}$$

onde $score_{opt}(f)$ é o valor do caminho para gerar a frase f a partir da Equação 9. Finalmente, selecionamos a frase com menor score normalizado e contendo, ao menos, um verbo como a melhor compressão das frases do documento.

Para exemplificar nosso método, simplificamos sua análise e utilizamos o texto modelado na Figura 1. Nessa figura, existem diversos caminhos possíveis entre os vértices *-começo-* e *-fim-*. A partir das palavras-chaves "George", "gigante", "solitário" "tartaruga"e "última", nosso método gerou a compressão "a tartaruga gigante conhecida george solitário morreu". Dentre as 5 palavras-chaves analisadas, foi gerada uma compressão contendo 4 delas e com as principais informações das frases.

4. Modelo Matemático Proposto

Formalmente, o GP utilizado pode ser representado como segue: seja $GP = (V, A)$ um grafo orientado simples no qual V é o conjunto de vértices (palavras), A o conjunto de arcos (*2-grams*) e b_{ij} é o peso do arco $(i,j) \in A$ (coesão das palavras dos vértices i e j, Equação 2). Sem perda de generalidade, considere v_0 como o vértice -começo- e adicione um arco auxiliar do vértice -fim- para v_0. Adicionalmente, cada vértice possui uma cor indicando se o mesmo é uma palavra-chave. Denotamos K como o conjunto de cores em que cada palavra-chave do documento representa uma cor diferente. A cor 0 (não palavras-chaves) possui o custo $c_0 = 0$ e as palavras-chaves possuem o mesmo custo $c_k = 1$ (para $k > 0$ e $k \in K$). O conjunto T é composto pelos *3-grams* do documento com uma frequência maior que 1. Cada *3-gram* $t = (a,b,c) \in T$ possui o custo $d_t = \text{3-gram}(a,b,c)$ (Equação 7) normalizados entre 0 e 1.

Existem vários algoritmos com complexidade polinomial para encontrar o menor caminho em um grafo. Contudo, a restrição de que o caminho deve possuir um número mínimo $Pmin$ de vértices (o número mínimo de palavras da compressão) torna o problema NP-Hard. De fato, encontrar o menor caminho no GP descrito implica encontrar

um ciclo com início e fim em v_0, e caso $Pmin$ seja igual a $|V|$, o problema corresponde ao Problema do Caixeiro Viajante (PCV). Nesse caso, como o PCV é um caso especial do problema descrito, ele também será NP-Hard.

O modelo matemático proposto para resolução do problema apresentado define cinco grupos de variáveis de decisão:

- $x_{ij}, \forall (i,j) \in A$, indicando se o arco (i,j) faz parte da solução;
- $y_v, \forall v \in V$, indicando se o vértice (a palavra) v faz parte da solução;
- $z_t, \forall t \in T$, indicando se o *3-gram* t faz parte da solução;
- $w_k, \forall k \in K$, indicando que alguma palavra com a cor (palavra-chave) k foi utilizada na solução; e
- $u_v, \forall v \in V$, variáveis auxiliares para eliminação de sub-ciclos da solução.

O processo de encontrar as 50 melhores soluções se deu pela proibição das soluções encontradas e reexecução do modelo. Optamos por essa estratégia em virtude da simetria gerada pela técnica de eliminação de sub-ciclos que utilizamos. A formulação é apresentada nas expressões (9) a (22).

$$\text{Minimize } \left(\alpha \sum_{(i,j)\in A} b_{i,j} \cdot x_{i,j} - \beta \sum_{k \in K} c_k \cdot w_k - \gamma \sum_{t \in T} d_k \cdot z_t\right) \tag{9}$$

$$\text{s.a.} \quad \sum_{v \in V} y_v \geq Pmin, \tag{10}$$

$$\sum_{v \in V(k)} y_v \geq w_k, \qquad \forall k \in K, \tag{11}$$

$$2z_t \leq x_{ij} + x_{jl}, \qquad \forall t = (i,j,l) \in T, \tag{12}$$

$$\sum_{i \in \delta^-(v)} x_{iv} = y_v \qquad \forall v \in V, \tag{13}$$

$$\sum_{i \in \delta^+(v)} x_{vi} = y_v \qquad \forall v \in V, \tag{14}$$

$$y_0 = 1, \tag{15}$$

$$u_0 = 1, \tag{16}$$

$$u_i - u_j + 1 \leq M - M \cdot x_{ij} \qquad \forall (i,j) \in A, j \neq 0, \tag{17}$$

$$x_{ij} \in \{0,1\}, \qquad \forall (i,j) \in A, \tag{18}$$

$$z_l \in \{0,1\}, \qquad \forall t \in T, \tag{19}$$

$$y_v \in \{0,1\}, \qquad \forall v \in V, \tag{20}$$

$$w_k \in [0,1], \qquad \forall k \in K, \tag{21}$$

$$u_v \in [1,|V|], \qquad \forall v \in V. \tag{22}$$

A função objetiva do programa (9) maximiza a qualidade da compressão gerada. As variáveis α, β e γ controlam, respectivamente, a relevância da coesão, das palavras-chaves e dos *3-grams* na geração da compressão. A restrição (10) limita o número de vértices (palavras) utilizadas na solução. O conjunto de restrições (11) faz a correspondência entre as variáveis de cores (palavras-chaves) e de vértices (palavras), sendo $V(k)$ o conjunto de todos os vértices com a cor k (uma palavra-chave pode ser representada por mais de um vértice). O conjunto de restrições (12) faz a correspondência entre as variáveis de *3-grams* e de arcos (*2-grams*). As igualdades (13) e (14) obrigam que para cada palavra usada na solução exista um arco ativo interior (entrando) e um exterior (saindo), respectivamente. A igualdade (15) força que o vértice zero seja usado na solução. Por fim, as restrições (16) e (17) são responsáveis pela eliminação de sub-ciclos enquanto as expressões (18)-(22) definem o domínio das variáveis.

Como discutido em Pataki (2003), existem duas formas clássicas de evitar ciclos em problemas derivados do PCV. A primeira consiste na criação de um conjunto exponencial de cortes garantindo que para todo subconjunto de vértices $S \subset V$, $S \neq \emptyset$, haja exatamente $|S| - 1$ arcos ativos (mais detalhes em Lenstra *et al*. (1985)). A segunda, conhecida como formulação Miller–Tucker–Zemlin (MTZ) utiliza um conjunto auxiliar de variáveis, uma para cada vértice, de modo a evitar que um vértice seja visitado mais de uma vez no ciclo e um conjunto de arcos-restrições. Mais informações sobre a formulação MTZ podem ser obtidas em Öncan *et al*. (2009).

Neste trabalho, optamos por eliminar sub-ciclos utilizando o método MTZ, uma vez que sua implementação é mais simples. Para tal, utilizamos uma variável auxiliar u_v para cada vértice $v \in V$, e o conjunto de arcos-restrições definido em (17). Nesse grupo de restrições, M representa um número grande o suficiente, podendo ser utilizado o valor $M = |V|$.

5. Experimentos computacionais

O desempenho do sistema proposto foi analisado a partir de diversos valores dos parâmetros (β e γ) associados à função objetivo. Os testes foram realizados num computador com processador i5 2.6 GHz e 6 GB de memoria RAM no sistema operacional Ubuntu 14.04 de 64 bits. Os algoritmos foram implementados utilizando a linguagem de programação Python e as bibliotecas takahe[6] e gensim[7]. O modelo matemático foi implementado na linguagem C++ com a biblioteca Concert e o *solver* utilizado foi o CPLEX 12.6.

5.1. Corpus e ferramentas utilizadas

Para avaliar a qualidade dos sistemas, utilizamos o corpus publicado por Boudin e Morin (2013). Esse corpus contém 618 frases (média de 33 palavras por frase) divididas em 40 *clusters* de notícias em Francês extraídos do Google News[8]. A taxa de redundância de um corpus é obtida pela divisão da quantidade de palavras únicas pela quantidade de palavras de cada *cluster*. A taxa de redundância do corpus que utilizamos é 38,8%. Cada palavra do corpus é acompanhada por sua POS. Para cada *cluster*, há 3 frases comprimidas por profissionais. Dividimos o corpus em duas partes de 20 *clusters*. A primeira parte é utilizada como corpus de aprendizado e a outra parte como corpus de teste. As frases do

[6]Site: http://www.florianboudin.org/publications.html
[7]Site: https://radimrehurek.com/gensim/models/ldamodel.html
[8]Site: https://news.google.fr

corpus de aprendizado tem o tamanho médio de 34,1 palavras e as frases do corpus de teste tem um tamanho médio de 31,6 palavras.

As características mais importante da CMF são a informatividade e gramaticalidade das frases. A informatividade representa a porcentagem das principais informações transmitidas no texto. Como consideramos que as compressões de referência possuem as informações mais importantes, avaliamos a informatividade de uma compressão baseada nas informações em comum entre a mesma e as compressões de referência usando o sistema ROUGE (Lin, 2004). Utilizamos as métricas de cobertura ROUGE-1 e ROUGE-2, que analisam os *1-grams* e *2-grams*, respectivamente, das compressões de referências presentes nas compressões geradas pelos sistemas, para estimar a informatividade das compressões geradas.

Devido a complexidade da análise gramatical de uma frase, foi utilizado uma avaliação manual para estimar a qualidade das compressões propostas por nosso sistema. Como a avaliação humana é lenta, utilizamos essa técnica somente para o corpus de teste. Para o corpus de aprendizado, decidimos avaliar somente a qualidade informativa (coberturas ROUGE-1 e ROUGE-2) e a TC devido ser inviável a análise manual da quantidade de testes do nosso sistema.

5.2. Resultados

Nomeamos nosso sistema como GP+OPT e utilizamos os sistemas de Filippova e de BM como *baselines*. Testamos o GP+OPT utilizando 1, 3, 5, 7 e 9 palavras-chaves[9] (PC) obtidas a partir do método LDA. Como o GP+OPT utiliza como base o método de Filippova, tornamos fixo o $\alpha = 1.0$ (priorizando a coesão das compressões geradas) e variamos β e γ de tal forma que:

$$\beta + \gamma < 1.0;\ \beta, \gamma = 0.0, 0.1, ..., 0.8, 0.9. \tag{23}$$

Todos os sistemas geraram a compressão de um documento em tempo viável (menos de 6 segundos). Devido à grande quantidade de testes gerados para o corpus de aprendizado, selecionamos os resultados que generalizam o funcionamento do GP+OPT. A Tabela 1 descreve a qualidade e a TC das compressões. Essa tabela é dividida em 4 partes. A primeira descreve os resultados das *baselines* e as demais partes descrevem os resultados do nosso sistema. A primeira parte da tabela comprova que o pós-tratamento utilizado por BM (análise da relevância das *keyphrases*) é melhor que a simples normalização dos scores das frases realizada por Filippova. O aumento da relevância dos *3-grams* na nossa modelagem melhora a informatividade da compressão sem aumentar bruscamente a TC, pois os *3-grams* favorecem a utilização de *2-grams* frequentes no texto (segunda parte da Tabela 1). Além disso, os 3-grams podem melhorar a qualidade gramatical, pois eles adicionam conjuntos de palavras gramaticalmente corretos à compressão.

Apesar do aumento da relevância das palavras-chaves gerar compressões com uma maior TC, as compressões são mais informativas (a terceira parte da tabela) e proporcionam as melhores compressões (linhas em negrito da Tabela 1). Dentre os melhores resultados (última parte da Tabela), escolhemos a versão do nosso sistema com PC=9, β=0.8

[9]Visto que o texto é composto de frases similares sobre um mesmo tópico, consideramos que 9 palavras é a quantidade máxima de palavras-chaves para representar um tópico.

e γ=0.1, pois essa configuração prioriza as palavras-chaves e tenta adicionar *3-grams* às compressões.

Tabela 1. A métrica de cobertura do ROUGE e a TC das compressões do corpus de aprendizado. As linhas em negrito destacam os melhores resultados e a versão selecionada do nosso sistema é marcada por uma estrela.

Sistemas	ROUGE-1	ROUGE-2	TC
Filippova (2010)	0,58769	0,43063	51,9%
Boudin e Morin (2013)	0,62364	0,45467	55,8%
GP+OPT PC=9 β=0.2 γ=0.0	0,53249	0,37515	48,3%
GP+OPT PC=9 β=0.2 γ=0.2	0,55202	0,40276	50,0%
GP+OPT PC=9 β=0.2 γ=0.4	0,57806	0,42385	51,8%
GP+OPT PC=9 β=0.2 γ=0.6	0,58996	0,43265	54,1%
GP+OPT PC=9 β=0.0 γ=0.2	0,49858	0,36156	43,0%
GP+OPT PC=9 β=0.2 γ=0.2	0,55202	0,40276	50,0%
GP+OPT PC=9 β=0.4 γ=0.2	0,58039	0,42394	52,1%
GP+OPT PC=9 β=0.6 γ=0.2	0,60072	0,43884	54,3%
GP+OPT PC=9 β=0.2 γ=0.7	0,59956	0,44160	48,3%
GP+OPT PC=7 β=0.6 γ=0.3	0,59981	0,43467	48,3%
GP+OPT PC=9 β=0.6 γ=0.3	0,61707	0,45033	48,3%
GP+OPT PC=9 β=0.8 γ=0.0	**0,62874**	**0,46089**	56,6%
GP+OPT PC=9 β=0.8 γ=0.1*	**0,62874***	**0,46089***	56,6%
GP+OPT PC=9 β=0.9 γ=0.0	**0,62874**	**0,46089**	56,6%

Selecionado a melhor configuração do nosso sistema, validamos a qualidade dos sistemas utilizando o corpus de teste (Tabela 2). Similar aos resultados do corpus de aprendizado, o método de BM foi melhor que o método de Filippova para as métricas ROUGE-1 e ROUGE-2. GP+OPT obteve resultados bem superiores que as *baselines* comprovando que a análise da coesão juntamente com as palavras-chaves e *3-grams* auxiliam a geração de melhores compressões. Apesar dos valores da TC do nosso sistema terem sido maiores que os valores da TC das *baselines*[10], a TC do sistema GP+OPT ficou próxima da TC das compressões dos profissionais (TC = 59%).

Tabela 2. A métrica de cobertura do ROUGE e a TC das compressões do corpus de teste.

Sistemas	ROUGE-1	ROUGE-2	TC
Filippova (2010)	0,58455	0,43939	51,1%
Boudin e Morin (2013)	0,62116	0,45734	55,2%
GP+OPT PC=9 β=0.8 γ=0.1	**0,70009**	**0,50207**	65,1%

Com o intuito de melhor analisar a qualidade informativa e gramatical das compressões, 5 franceses avaliaram as compressões geradas por cada sistema e notificaram a qualidade gramatical e informativa para o corpus de teste (Tabela 3). Nosso sistema gerou estatisticamente compressões mais informativas que as *baselines*. Apesar da média da gramaticalidade do nosso sistema ter sido inferior a dos demais sistemas, não podemos confirmar qual sistema é estatisticamente melhor para a gramaticalidade devido ao fato dos intervalos de confiança da gramaticalidade dos sistemas se cruzarem. Portanto, nosso

[10] A diferença do tamanho médio das frases entre os sistemas GP+OPT e Filippova foi 3,7 palavras.

Tabela 3. Média e intervalo de confiança da avaliação manual da informatividade e gramaticalidade das compressões do corpus de teste. As notas possíveis para cada métrica são entre 0 e 5.

Sistemas	Gramaticalidade	Informatividade
Filippova (2010)	**4,2** $\pm$ 0,18	2,86 $\pm$ 0,32
Boudin e Morin (2013)	3,99 $\pm$ 0,21	3,31 $\pm$ 0,32
GP+OPT PC**=9** β**=0.8** γ**=0.1**	3,93 $\pm$ 0,22	**3,95** $\pm$ 0,23

sistema pode gerar compressões com qualidade gramatical igual às compressões geradas pelos métodos de Filippova ou de BM.

Desse modo, pode-se afirmar que o GP+OPT apresentou melhores resultados que as *baselines* gerando compressões mais informativas e com uma boa qualidade gramatical.

6. Considerações Finais e Proposta de Trabalhos Futuros

A CMF gera frases de boa qualidade sendo uma ferramenta interessante para a SAT. A análise concomitante da coesão, palavras-chaves e *3-grams* identificaram as informações principais do documento. Apesar do nosso sistema ter gerado compressões com uma TC maior que as *baselines*, a informatividade foi consideravelmente melhor. A análise manual dos franceses comprovou que nosso método gerou compressões mais informativas e mantendo uma boa qualidade gramatical.

Os próximos trabalhos visam criar um corpus similar ao de BM para o idioma Português e testar o desempenho do nosso sistema para diferentes idiomas. Além disso, pretende-se adaptar o sistema para escolher a relevância das palavras-chaves e dos *3-grams* baseados no tamanho e no vocabulário do documento. Finalmente, objetiva-se implementar diferentes métodos para a obtenção de palavras-chaves, a fim de avaliar o impacto de cada um na qualidade da geração da CMF.

Agradecimentos

Este trabalho foi financiado parcialmente pelo projeto europeu CHISTERA-AMIS ANR-15-CHR2-0001.

Referências

Barzilay, R. e McKeown, K. R. (2005), Sentence fusion for multidocument news summarization. *Computational Linguistics*, v. 31, n. 3, p. 297–328.

Blei, D. M., Ng, A. Y. e Jordan, M. I. (2003), Latent dirichlet allocation. *Journal Machine Learning Research*, v. 3, p. 993–1022.

Boudin, F. e Morin, E. Keyphrase extraction for n-best reranking in multi-sentence compression. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 298–305, Atlanta, Georgia. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1030`, 2013.

Filippova, K. Multi-sentence compression: Finding shortest paths in word graphs. Huang, C.-R. e Jurafsky, D. (Eds.), *COLING*, p. 322–330. Tsinghua University Press. URL `http://dblp.uni-trier.de/db/conf/coling/coling2010.html#Filippova10`, 2010.

Lenstra, J. K., Kan, A. R., Lawler, E. L. e Shmoys, D. *The traveling salesman problem: a guided tour of combinatorial optimization*. John Wiley & Sons, 1985.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. *ACL workshop on Text Summarization Branches Out*, 2004.

Linhares Pontes, E., Linhares, A. C. e Torres-Moreno, J.-M. Sasi: sumarizador automático de documentos baseado no problema do subconjunto independente de vértices. *Anais do Simpósio Brasileiro de Pesquisa Operacional*, 2014.

Linhares Pontes, E., Linhares, A. C. e Torres-Moreno, J.-M. Lia-rag: a system based on graphs and divergence of probabilities applied to speech-to-text summarization. *CCCS (Call Centre Conversation Summarization) Multiling 2015 Workshop*, 2015.

Mihalcea, R. e Tarau, P. TextRank: Bringing order into texts. *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

Öncan, T., Altınel, İ. K. e Laporte, G. (2009), A comparative analysis of several asymmetric traveling salesman problem formulations. *Computers & Operations Research*, v. 36, n. 3, p. 637–654.

Pataki, G. (2003), Teaching integer programming formulations using the traveling salesman problem. *SIAM review*, v. 45, n. 1, p. 116–123.

Torres-Moreno, J.-M. *Automatic Text Summarization*. John Wiley & Sons. ISBN 978-1-84821-668-6, 2014.