Categorização cognitiva no processamento de notícias: Um estudo a partir da sumarização automática multi-documentos

Victor S. Bursztyn[1], Raul S. Ferreira[1], Ygor Canalli[1], Geraldo Xexéo[1 2]

[1] Programa de Engenharia de Sistemas e Computação/COPPE – Universidade Federal do Rio de Janeiro, Brazil
{vbursztyn, raulsf, canalli, xexeo}@cos.ufrj.br
[2] Departamento de Ciência da Computação/IM - Universidade Federal do Rio de Janeiro, Brazil

Abstract. In the past decades, researchers from several fields have been studying people's opinions on products. Deep love and hate connections have caught the attention of cognitive scientists, who have posed theories associating those sorts of judgements to different cognitive interactions with products features. A particular theory, published by cognitive scientist Donald Norman, explains in a widely accepted manner that people tend to interact with products through three cognitive levels. In this study, we embrace Norman's theory to explain how readers connect to news produced by web portals. Similarly, we group news subjects into three different collections, each with a prevailing cognitive attractiveness. Then, assuming three "cognitive categories" of news, we pose an experiment that shows how text features vary according to each category. Our feature-based multi-document news summarizer searches for the optimal weights configuration in each collection. By finding three different configurations with substantial gains over a grossly general configuration, we argue that cognitive categorization is meaningful to news processing.

Categories and Subject Descriptors: I.5.4 [**Applications**]: Text processing

Keywords: cognitive-based summarization, text mining

1. INTRODUÇÃO

Este trabalho introduz a categorização cognitiva como parte do processo de sumarização de notícias, medindo os ganhos reais com o uso dessa informação. Argumentamos que os três níveis cognitivos propostos por [Norman 2008], por definição, abarcam diferentes editorias numa única coleção. Assim, como três grupos de editorias, fazem parte da estrutura de produção e de oferta de notícias pelos portais digitais. Verificamos que as editorias começam a ser estudadas pontualmente, enquanto que a categorização cognitiva ainda é inexplorada. Ambas são informações à disposição das aplicações que processam notícias, com uma diferença significativa entre si: as editorias podem ser muito numerosas; já os níveis cognitivos oferecem uma organização mais concisa e de fácil administração.

Segundo [Radev et al. 2002] um sumário é um texto produzido a partir de uma ou mais fontes textuais, que carrega consigo informações relevantes sobre elas, com um tamanho substancialmente menor. Por sua vez, a extração é o procedimento de seleção de seções importantes no texto original. Um sumário extrativo, portanto, é feito de sentenças preservadas das fontes originais.

Para [Nenkova et al. 2011], a sumarização automática multi-documentos possui grande importância. Ela agrega valor ao reduzir o tempo de busca de uma informação, uma vez que os sumários são exibidos antes das fontes originais. Além do ganho de tempo, o autor cita estudos que demonstram a importância dos sumários para o entendimento de assuntos com muitas referências: quando disponíveis, verifica-se que os leitores produzem relatórios melhores.

Para validar nossa proposta, desenvolvemos um sistema de sumarização automática multi-documentos e mostramos que ele pode obter um resultado nitidamente melhor se os documentos forem previamente classificados de acordo com os níveis cognitivos [Norman 2008].

2. TRABALHOS RELACIONADOS

SUMMONS [McKeown and Radev 1995] é o primeiro exemplo, que se tem registro, de sistema sumarizador multi-documentos. A abordagem proposta buscava por repetições de conceitos nos textos, para então selecionar as sentenças relevantes. Outras abordagens vieram sendo propostas nos anos seguintes [Zhang 2002; Evans et al. 2005; Lebanon et al. 2007], como se pode verificar no levantamento de [Das and Martins 2007].

Em pesquisa mais recente, feita por [Tabassum and Oliveira 2015], constata-se a existência de abordagens mais comuns para a sumarização multi-documentos: a baseada em *features*; em ontologias; em cognição; em eventos; e no discurso. Frequentemente, sumarizadores podem ser híbridos. Na prática, este trabalho apresenta um sumarizador baseado em *features* textuais básicas, mas, fundamentalmente, também trata de um premissa cognitiva forte.

Em [Camargo et al. 2012], os autores realizam uma análise do corpus CSTNews[1] inteiramente focada na editoria "Mundo", contendo notícias sobre acidentes, desastres naturais e incidentes políticos. Já [Chen and Li 2013] propõe uma abordagem cognitiva que simula, especificamente, o processo humano de leitura, com a memória das palavras, suas associações e outros três aspectos cognitivos estimulados durante uma leitura. [Kumar et al. 2014] propõe um modelo de sumarização multi-documentos baseado em outras *features* textuais de discurso, onde, na mesma linha de [Camargo et al. 2012], foca no que chama de "domínio": notícias sobre desastres naturais.

Portanto, se por um lado existem, na literatura, referências ao uso da cognição na modelagem de sumarizadores, ainda não se usou da mesma ciência para entender como diferentes editorias ou domínios afetam o processamento de notícias. Também não se propôs uma categorização cognitiva para elas.

3. OS TRÊS NÍVEIS COGNITIVOS

A premissa fundamental para este estudo deriva do trabalho do professor Donald Norman[2]. Para isso, entre seus livros, destaca-se "Design Emocional: Por que Adoramos (ou Detestamos) os Objetos do Dia a Dia" [Norman 2008], onde Norman associa conceitos da psicologia cognitiva à apreciação ou repulsa causadas por produtos. Mais especificamente, o autor ensina a planejar produtos que atendam a todos os interesses de seus potenciais consumidores. Ele explica que existem três níveis cognitivos básicos na percepção de um objeto:

(1) O nível visceral diz respeito aos aspectos físicos dos produtos, bem como a seus efeitos mais instintivos. Objetos simétricos, com texturas suaves e cores harmoniosas tendem a nos relaxar, ao passo que objetos pontiagudos, ásperos e com combinações de cores mais agressivas podem, até mesmo, nos estressar. Em poucas palavras, os produtos se comunicam com os instintos humanos através desse primeiro nível cognitivo.
(2) O nível funcional diz respeito à capacidade que produtos têm para efetivamente solucionar problemas relevantes de nosso cotidiano. Objetos são percebidos por suas qualidades práticas, o que significa que também precisam ser fáceis de se usar. Quando produtos são planejados para serem vistos como soluções práticas e efetivas, buscam comunicar, justamente, com o nosso segundo nível cognitivo.
(3) O nível afetivo, por fim, diz respeito à capacidade que produtos têm para criar ou despertar relações afetivas. Segundo Norman, quando o carro Mini Cooper foi lançado, a primeira crítica especializada informava ao leitor que o carro, em si, dispunha de atributos que variavam de bons a

[1] http://www.icmc.usp.br/ taspardo/sucinto/cstnews.html
[2] professor emérito de ciência cognitiva na Universidade da Califórnia, professor de ciência da computação na Universidade de Northwestern. Maiores detalhes em: https://en.wikipedia.org/wiki/Don_Norman

medíocres. Porém, a mesma crítica concluía sugerindo que se relevasse eventuais defeitos, pois há muito tempo um carro não arrancara tantos sorrisos. Ao evocar o visual dos carros de brinquedo, os criadores do Mini Cooper conseguiram comunicar com o nível afetivo de muitos consumidores. Assim, o terceiro nível cognitivo também pode se tornar a pauta principal no planejamento de produtos.

Os três níveis do Design Emocional, uma vez que se baseiam em conceitos da psicologia cognitiva — por sua vez, focada na compreensão da mente humana —, podem permear outras atividades profissionais intensamente focadas em consumidores. Nesse sentido, é possível observar conexões entre os três níveis e os esforços de planejamento dos conteúdos em portais digitais. Analogamente, notícias podem ser produzidas visando três tipos de atração:

(1) Na intenção de alcançarem um novo clique do consumidor de notícias, elas podem ser feitas buscando comunicação com nossos instintos, mais que com os demais níveis. São exemplos as notícias dramáticas que ocorrem distantes de nosso universo cotidiano, de nossa realidade. Sabe-se, pelos jornalistas que as formatam, que a motivação do clique é um interesse visceral. A atração, neste caso, não é funcional nem afetiva.
(2) Outras notícias podem comunicar com o nível funcional de seu público, esclarecendo alguma problemática que o impacte diretamente. São exemplos as notícias sobre economia ou sobre sua comunidade local, que abordam o dia a dia do leitor, onde também cabem esclarecimentos dos detalhes mais úteis e intimamente conhecidos.
(3) Por fim, elas podem comunicar com o nível afetivo, quando discorrem sobre elementos geradores de afeto. O mais provável, nestes casos, é que os cliques sejam feitos devido ao universo afetivo de seus consumidores, o que leva os jornalistas a detalharem, ainda, outros aspectos do conteúdo. São exemplos as notícias sobre esportes, novelas, artistas, entre outras.

4. O PROBLEMA DE SUMARIZAÇÃO

Seja A uma coleção de a_i assuntos distintos, cada qual contendo α_j artigos, onde $\alpha_j \in a_i$. Sejam σ_k as sentenças que compõem cada artigo α_j. Seja também s_i o sumário extrativo associado a a_i, formado por sentenças pertencentes aos artigos α_j. Deseja-se utilizar a informação do tipo cognitivo de a_i de forma a tornar s_i o mais parecido possível com s_i^*, o sumário ideal do assunto.

O sumarizador construído para lidar com o problema se baseia na hipótese de que determinadas características das sentenças possuem uma importância maior de acordo com a categoria cognitiva do assunto. Inicialmente, coletamos diversas *features* $f_1, \ldots, f_r$ de cada sentença σ_k dos artigos α_j de um determinado assunto $a_i \in A$, de forma que cada *feature* é a medida de uma característica da sentença. Além disso, avaliamos o *valor agregado* $v(\sigma)$ de uma sentença σ qualquer da seguinte maneira:

$$v(\sigma) = \sum_{\rho=1}^{r} w_\rho \cdot f_\rho(\sigma),$$

onde $w_1, \ldots, w_r$ são pesos associados a cada *feature*.

Após isso, para cada assunto a_i, unimos as sentenças dos artigos α_j num único conjunto S_i, tal que

$$S_i = \{\sigma : \sigma \in a_i, v(\sigma) \geq t\},$$

onde t é um limiar arbitrário de corte. Iremos produzir um sumário extrativo s_i de um assunto a_i a partir da resolução do *problema-da-mochila booleana* sobre o conjunto de sentenças S_i. Para modelar o sumarizador como um problema-da-mochila booleana, consideramos cada sentença σ um item, que pode ou não ser colocado na *mochila* M, de forma que o *peso* $p(\sigma) = |\sigma|$, onde $|\sigma|$ é o comprimento de σ em caracteres, e seu valor $v(\sigma)$, conforme definido anteriormente. Adicionalmente, a capacidade

P da mochila M é dada pelo comprimento em caracteres de um *sumário ideal* s_i^*. Assim, para cada σ_l, as L sentenças de S_i, desejamos encontrar $X = \{x_l : x_l \in \{0, 1\}\}$, que resolva o problema

$$\max \sum_{l=1}^{L} x_l \cdot v(\sigma_l) \text{ sujeito a } \sum_{l=1}^{L} x_l \cdot p(\sigma_l) \leq P.$$

Com isso, temos que o sumário extrativo s_i do assunto a_i é dado por $s_i = \{\sigma_l : x_l = 1\}$. Em outras palavras, s_i é uma seleção ótima de sentenças, que maximizam o valor agregado do sumário produzido, de acordo com um limite de caracteres. Por fim, reordena-se as sentenças em s_i de acordo com sua ordem de ocorrência original, para assim, produzir o sumário extrativo multi-documentos.

Seja Γ a quantidade de tipos cognitivos, e $A_\gamma \subset A$ partições da coleção de assuntos, de forma que A_γ contém os n_γ assuntos do γ-ésimo tipo cognitivo. Com isso, extraímos empiricamente a configuração ótima de pesos $w_1, \ldots, w_r$ para cada A_γ, da seguinte maneira: Sejam $c_\delta \in C$ as possíveis configurações de peso, onde $c_\delta = \{w_1^{(\delta)}, \ldots, w_r^{(\delta)}\}$. Seja também $s_i(c_\delta)$ o sumário extrativo de um assunto $a_i \in A_\gamma$ obtido conforme descrito anteriormente, utilizando-se da configuração de pesos c_δ. Desejamos obter a configuração ótima de pesos c_γ^* que maximize a métrica F_1 em relação aos sumários ideais s_i^* dos assuntos pertencentes à partição A_γ, a qual é dada por

$$c_\gamma^* = \arg\max_{c_\delta \in C} \sum_{i=1}^{n_\gamma} F_1(s_i(c_\delta), s_i^*).$$

Nesta etapa, utilizamos o *Grid Search* [Staelin 2003] para selecionar a melhor configuração de pesos para uma determinada partição de A. A técnica do *Grid Search* consiste numa busca exaustiva sobre todas as combinações possíveis de parâmetros, onde para cada parâmetro é dada uma lista de valores admissíveis.

5. ANÁLISE EXPERIMENTAL DO IMPACTO DA CATEGORIZAÇÃO COGNITIVA

Com o objetivo de sustentar a hipótese levantada, desejamos mostrar que cada tipo cognitivo de notícias γ possui um c_γ^* consideravelmente distinto, ou seja, que a cada tipo cognitivo de notícias devem ser atribuídas importâncias distintas às características das sentenças para que se obtenha um sumário extrativo multi-documentos de melhor qualidade.

Para viabilizar o experimento e aproximar-se das respostas, desenvolvemos o "Dynamic and Extractive Summarizer towards Human Interests on News" (ou DESHIN[3]). Trata-se de uma solução código aberto e desenvolvida em Python para a geração de sumários extrativos a partir de um conjunto de notícias sobre um mesmo tema. Como será explicado mais adiante, os três níveis cognitivos fazem parte da estrutura do DESHIN desde sua concepção, uma vez que os temas são agrupados e organizados em três coleções, cada uma simbolizando um "tipo cognitivo".

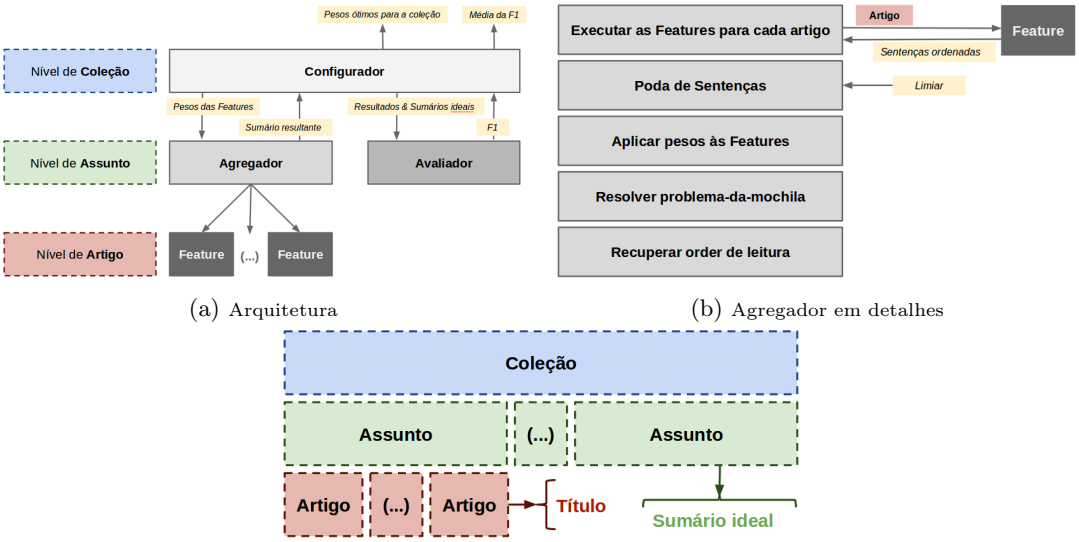
A proposta levantada foi modelada em um esquema conforme os módulos abaixo, os quais são ilustrados na Figura 1a. Os passos executados pelo Agregador para sumarizar um assunto, conforme a proposta apresentada anteriormente, são ilustrados na Figura 1b:

Agregador. responsável por sumarizar um assunto.

Avaliador. responsável por avaliar o sumário gerado pelo agregador em relação ao sumário ideal do assunto.

Configurador. responsável por realizar as chamadas do Agregador e Avaliador para cada coleção de tipo cognitivo, e realizar os ajustes ótimos de pesos de cada coleção de tipo cognitivo.

[3]Disponível em https://github.com/vbursztyn/DESHIN



(a) Arquitetura (b) Agregador em detalhes

(c) Modelo de dados

Figura 1: Esquema do DESHIN

Com o apoio do *Natural Language Toolkit* (NLTk) e da biblioteca *NLTk Trainer*, as *features* responsáveis pela modelagem dos sumários são:

f_B: A "*feature* base"tem nomeação distinta das demais por uma razão que será esclarecida na etapa de análise dos resultados. Trata-se da "*feature*"que calcula a similaridade média de cada sentença com a lista de títulos das notícias. A medida de similaridade adotada é o coseno. Antes do cálculo, para ambos os lados, são removidas palavras irrelevantes (*stopwords*) e sobre o que resta, reduzimos as palavras a seus radicais (*stemming*). Seja um assunto a formado pelos artigos $\alpha_1, \ldots, \alpha_n$, onde S é o conjunto de todas as sentenças dos artigos de a, e $\tau(\alpha_i)$ é o título do i-ésimo artigo do assunto a. Assim, para uma sentença $\sigma_l \in S$ qualquer, temos que $f_B(\sigma_l) = \frac{1}{n} \sum_{j=1}^{n} sim(\sigma_l, \tau(\alpha_j))$, onde $sim(a, b)$ é a similaridade do cosseno dos termos de duas sentenças a, b quaisquer.

f_1: Trata-se da *feature* que amplifica a importância das sentenças de acordo com suas localizações no texto. Nesta implementação, as pontuações decaem linearmente até o centro do texto e voltam a crescer, em igual proporção, até a última sentença. Como resultado, são priorizadas as sentenças iniciais e finais das notícias. Seja α um artigo formado pelas setenças $\sigma_1, \ldots, \sigma_m$. Assim, para uma sentença $\sigma_k \in \alpha$ qualquer, temos que $f_1(\sigma_k) = \left|k - \frac{m}{2}\right| \times 0.1$.

As três *features* seguintes se baseiam em um rotulador de *part-of-speech* treinado com o *NLTk Trainer*. O córpus usado para o treinamento é o Mac Morpho[4], muito compatível com o nosso experimento pois é baseado em notícias publicadas pelo jornal Folha de São Paulo, que, como visto, é uma das fontes abarcadas pelo CSTNews.

f_2: Trata-se da *feature* que amplifica sentenças que façam mais referências aos atores mais frequentes no texto. Isso é feito com a contagem de todos os sintagmas nominais presentes na notícia para, depois, pontuar as sentenças que contenham os mais frequentes. Seja α um artigo formado pelas sentenças σ_k, onde cada sentença é um conjunto de termos t_i. Seja também NPr o conjunto de nomes próprios de α, $score(t_i) = \#\{\text{ocorrências de } t_i \text{ em } NPr\}$ e $MF = \max_{t_i \in \alpha} score(t_i)$. Assim, para $\sigma_k \in \alpha$, temos que $f_2(\sigma_k) = \left(\sum_{t_i \in \sigma_k} score(t_i) \right) \cdot \frac{1}{MF}$.

[4]Disponível em https://sites.google.com/site/linguacorpus/acdc/mac-morpho

f_3: De forma oposta, esta *feature*, amplifica sentenças que façam mais referências aos atores menos frequentes no texto. Assim, para $\sigma_k \in \alpha$, baseado nas definições anteriores temos que $f_3(\sigma_k) = MF \cdot \left(\sum_{t_i \in \sigma_k} score(t_i) \right)^{-1}$.

f_4: Por fim, trata-se da *feature* que amplifica sentenças que contenham mais verbos e substantivos, assumindo serem indicadores de relevância para cada sentença. Seja α um artigo formado pelas sentenças σ_k, com $1 \leq k \leq n$, tal que σ_k é formada pelos termos t_i. Sejam Sb e Vr os conjuntos de substantivos e verbos de α, respectivamente. Assim, para $\sigma_k \in \alpha$, temos que $f_4(\sigma_k) = \#\{t_i : t_i \in \sigma_k, t_i \in Sb, t_i \in Vr\}/n$.

Todas as *features* retornam a lista de sentenças de uma notícia. A ordenação é decrescente, a partir das pontuações calculadas com o critério da *feature* em questão, normalizadas entre 0.0 e 1.0.

5.1 Base de dados

A base de dados usada foi o córpus multi-documentos CSTNews [Aleixo and Pardo 2008], criado pela USP, que reúne matérias sobre 32 assuntos, coletadas dos jornais Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Para cada tema, o córpus oferece diversas opções de sumários, além de uma série de meta-dados sobre as sentenças que compõem cada artigo. No âmbito deste trabalho, foram aproveitados apenas os sumários extrativos, validados por dois juízes humanos. Os meta-dados provenientes do córpus não foram aproveitados pelas features. A Figura 1c ilustra o modelo de dados, tal como estruturado pelo DESHIN.

Mais especificamente, aproveitamos 25 dos 32 assuntos, distribuídos entre as três coleções conforme a Tabela I. Não foram aproveitados os demais assuntos, pois a coleção para o tipo 3 já compreendia todas as opções presentes no córpus. Se os 32 fossem incluídos, haveria menos simetria entre as coleções.

Tipo 1 (visceral)	Tipo 2 (funcional)	Tipo 3 (afetivo)
C12_Mundo_EnchenteCoreia	C20_Politica_CPMF	C27_Esportes_GoleadaEquador
C32_Mundo_FalhaNuclear	C2_Politica_ReeleicaoLula	C28_Esportes_HeptaVolei
C23_Mundo_EnchenteReinoUnido	C16_Politica_Sanguessugas	C8_Esportes_LigaVolei
C18_Mundo_AtaqueVirginia	C11_Cotidiano_PCC	C24_Esportes_FabianaMue
C26_Mundo_FuracaoMexico	C6_Cotidiano_CanteiroObras	C25_Esportes_CopaAmerica
C14_Mundo_AcidenteTrens	C17_Politica_EleicaoAlckmim	C7_Ciencia_NovoPlaneta
C15_Mundo_ExplosaoMoscou	C30_Dinheiro_LucroItau	C19_Esportes_Maradona
C13_Mundo_SriLanka	C9_Politica_Desvio	C31_Esportes_Jade
C1_Mundo_AviaoCongo		

Tabela I: Coleções

5.2 Resultados

Adotou-se para a avaliação dos resultados a medida F_1, calculada a partir da precisão P e da revocação R dos sumários resultantes, frente aos respectivos sumários ideais fornecidos pelo córpus. Segundo [Yang and Liu 1999], a medida F_1 é dada por $F_1 = \frac{2*P*R}{(P+R)}$, sendo $P = A/(A+B)$ e $R = A/(A+C)$, em que A equivale à quantidade de sentenças corretamente selecionadas para o sumário resultante; B às que foram equivocadamente selecionadas para o sumário resultante (i.e. não estão no sumário ideal); e C às que deveriam ter sido selecionadas, mas não foram.

Os resultados da execução do modelo proposto sobre as coleções previamente descritas foram divididos em três partes. Na Tabela II, são destacados os resultados para o uso isolado da chamada

feature base (f_B), para criar uma referência usada *baseline*, por ser uma medida básica aceita como útil na sumarização.

f_B, f_1, f_2, f_3 e f_4	Coleção	F_1 Médio
1, 0, 0, 0, 0	tipo 1	0.155
1, 0, 0, 0, 0	tipo 2	0.160
1, 0, 0, 0, 0	tipo 3	0.256

Tabela II: Uso isolado da *feature* base

Na Tabela III, são destacados os resultados referentes às configurações de pesos vencedoras para cada coleção. São, dessa forma, os resultados ótimos com as features f_B, f_1, f_2, f_3 e f_4 ajustadas para cada coleção. Observa-se, na última coluna à direita, o quanto a existência de *features* adicionais a f_B, somada à existência da categorização cognitiva, agregam ao processo de sumarização. Em comparação à abordagem de f_B, o modelo proposto oferece ganhos de $13,60\%$ a $51,35\%$ na qualidade dos sumários resultantes. Também se observa a configuração ótima com todos os tipos numa única coleção.

f_B, f_1, f_2, f_3 e f_4	Coleção	F_1 Médio	Ganho frente à f_B
0.5, 0.5, 0.1, 0.1, 0.7	tipo 1	0.235	51.35%
0.5, 0.1, 0.1, 0.1, 0.7	tipo 2	0.189	17.50%
0.7, 0.1, 0.1, 0.1, 0.3	tipo 3	0.291	13.60%
0.7, 0.1, 0.1, 0.1, 0.3	todos	0.216	-

Tabela III: Configurações de pesos vencedoras para cada coleção

Na Tabela IV, simulamos os casos em que uma das configurações vencedoras é genericamente aplicada às três coleções. Nesse contexto, não se reconhece a categorização cognitiva. Ao invés, assume-se que um único modelo atenderá a um processo de sumarização universal. Observa-se, mais uma vez, que tal generalização causa perdas substanciais na qualidade dos sumários resultantes, frente ao que pode ser alcançado com o uso dessa informação para diferenciar as coleções (vide Tabela III).

f_B, f_1, f_2, f_3 e f_4	Coleção	F_1 Médio	Perda frente ao melhor ajuste
0.5, 0.5, 0.1, 0.1, 0.7	tipo 1	0.235	0
0.5, 0.5, 0.1, 0.1, 0.7	tipo 2	0.104	- 44.95%
0.5, 0.5, 0.1, 0.1, 0.7	tipo 3	0.195	- 33.05%
0.5, 0.1, 0.1, 0.1, 0.7	tipo 1	0.194	- 17.30%
0.5, 0.1, 0.1, 0.1, 0.7	tipo 2	0.189	0
0.5, 0.1, 0.1, 0.1, 0.7	tipo 3	0.221	- 23.90%
0.7, 0.1, 0.1, 0.1, 0.3	tipo 1	0.197	- 16.15%
0.7, 0.1, 0.1, 0.1, 0.3	tipo 2	0.164	- 12.95%
0.7, 0.1, 0.1, 0.1, 0.3	tipo 3	0.291	0

Tabela IV: Aplicação genérica de uma das configurações vencedoras

Nesse experimento, fomos então capazes de validar a ideia que o conhecimento das categorias cognitivas pode impactar positivamente no processo de sumarização multi-documentos.

6. CONCLUSÃO E TRABALHOS FUTUROS

Com foco na análise de nossa principal premissa, o DESHIN foi bem sucedido. Mesmo que em cima de *features* medianas, foram testados projetos equivalentes de sumarizadores, comparados entre si no aspecto mais relevante para este estudo. Como demonstrado na Tabela IV, o uso da categorização cognitiva garantiu ganhos que oscilam de 12.95% a 44.95%, configurando um ganho médio bastante significativo. Nesse sentido, um novo experimento relevante seria a reaplicação dessa premissa no

contexto do trabalho de [Castro Jorge and Pardo 2010], para tangibilizar esses ganhos na forma de um F_1 ainda mais alto.

O desempenho médio para as três coleções é de 0.238, 10.20% superior ao F_1 de 0.216 obtido com o *Grid Search* sobre os 25 assuntos numa única coleção. No entanto, é inferior ao estado-da-arte [Castro Jorge and Pardo 2010]. Uma explicação aparentemente natural está na qualidade das *features* implementadas pelo DESHIN, que podem ser melhores que a *feature* base, isolada, mas ainda estão distantes do estado-da-arte científico. Nota-se, por exemplo, que nenhuma das configurações vencedoras fez uso das *features* f_2 e f_3, já que, em todas, seus pesos estão reduzidos ao menor valor permitido (0.1). No mesmo córpus, [Castro Jorge and Pardo 2010] demonstram uma série de estratégias de sumarização baseadas em *features* mais sofisticadas (segundo a teoria *Cross-document Structure*, ou CST), alcançando F_1 superior a 0.5. Conclui-se, dessa forma, que há muito espaço para melhora com a exploração de *features* mais fortes.

Por fim, apontamos algumas direções que podem ser consideradas futuramente: a ampliação do estudo em sumarização multi-documentos, com um corpus maior, no idioma inglês e com *features* mais próximas do estado-da-arte; a inclusão no CSTNews ou criação de um *dataset* de notícias categorizadas entre os três tipos, para abrir novas possibilidades de pesquisa; o desenvolvimento de classificadores, com base no *dataset* criado; e, finalmente, o uso dos níveis cognitivos como mecanismos de filtragem da recomendação personalizada em portais digitais; entre outros possíveis caminhos de pesquisa.

REFERENCES

Aleixo, P. and Pardo, T. *CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multi-documento CST (cross-document structure theory.* ICMC-USP, 2008.

Camargo, R. T., Maziero, E. G., Pardo, T. A., and de Linguística Computacional, N. I. Corpus analysis of aspects in multi-document summaries–the case of news texts from "world" section. In *the OnlineProceedings of the 11 th Corpus Linguistics Symposium*, 2012.

Castro Jorge, M. L. d. R. and Pardo, T. A. S. Experiments with cst-based multidocument summarization. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing.* TextGraphs-5. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 74–82, 2010.

Chen, J. and Li, W. Cognitive-based multi-document summarization approach. In *Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on.* IEEE, pp. 214–217, 2013.

Das, D. and Martins, A. F. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU* vol. 4, pp. 192–195, 2007.

Evans, D. K., McKeown, K., and Klavans, J. L. Similarity-based multilingual multi-document summarization, 2005.

Kumar, Y. J., Salim, N., Abuobieda, A., and Albaham, A. T. Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing* vol. 21, pp. 265–279, 2014.

Lebanon, G., Mao, Y., and Dillon, J. V. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research* 8 (10): 2405–2441, 2007.

McKeown, K. and Radev, D. R. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, pp. 74–82, 1995.

Nenkova, A., Maskey, S., and Liu, Y. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011.* Association for Computational Linguistics, pp. 3, 2011.

Norman, D. A. *Design emocional: por que adoramos (ou detestamos) os objetos do dia-a-dia.* Rocco, 2008.

Radev, D. R., Hovy, E., and McKeown, K. Introduction to the special issue on summarization. *Computational linguistics* 28 (4): 399–408, 2002.

Staelin, C. Parameter selection for support vector machines. *Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1*, 2003.

Tabassum, S. and Oliveira, E. A review of recent progress in multi document summarization. In *Doctoral Symposium in Informatics Engineering*, 2015.

Yang, Y. and Liu, X. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, pp. 42–49, 1999.

Zhang, Z.; Blair-Goldensohn, S. R. D. Towards cst-enhanced summarization. in the proceedings of aaai conference, 2002.