Análise de Aspectos de Sumários Multidocumento e sua Correlação com a Informatividade

1. Introdução

Aplicações computacionais capazes de tratar adequadamente a incrível quantidade de informações disponíveis atualmente, principalmente na web, têm se tornado cada vez mais necessárias. A sumarização automática multidocumento é um exemplo de aplicação, na qual, a partir de um conjunto de textos que versam sobre um mesmo assunto, produz-se um único sumário. Os sumários produzidos podem ser de propósito geral ou com foco em aspectos pré-definidos de acordo com a categoria dos textos, como é o caso da sumarização guiada. As categorias definidas pela TAC (*Text Analysis Conference*) de 2010 para a sumarização guiada são *accidents and natural disasters*, *attacks (criminal/terrorist)*, *health and safety*, *endangered and resources* e *trials*. Owczarzak e Dang (2011) definem os seguintes aspectos para a categoria *trials*: *who, who investigating*, *why*, *charges, plead* e *sentence*.

A sumarização guiada serve a dois propósitos: primeiro, cria-se um alvo mais focado para os sistemas de sumarização automática, neutralizando a variação humana e apontando os tipos concretos de informação exigidos pelo leitor e, segundo, fornece uma ferramenta de diagnóstico detalhado para analisar o conteúdo dos sumários automáticos (Owczarzak e Dang, 2011). Para alcançar tais objetivos, os autores sugerem verificar a similaridade entre os sumários manuais e automáticos, para identificar as deficiências e pontos positivos dos sumarizadores.

Para Zhang et al. (2011), um sumário que contenha todos os aspectos de acordo com dada categoria é considerado ideal, na medida em que aborda a necessidade do usuário, e é semanticamente bem estruturado e coerente. Os autores discutem dois problemas: 1) como encontrar os aspectos já que esses são unidades de conteúdo escondidas na superfície do texto e, 2) como usar os aspectos para produzir sumários coerentes. Em busca de uma solução para esses problemas, eles desenvolveram um sumarizador extrativo que seleciona sentença a partir do reconhecimento dos aspectos. Comparado ao melhor sumarizador da TAC de 2011, os sumários guiados por aspectos se mostraram mais informativos, de acordo com os valores fornecidos pela ROUGE (Lin, 2004).

Acredita-se que um sumário automático que contém o maior número de aspectos presentes no sumário manual tem boa informatividade. A partir disso, torna-se

interessante investigar se os sumários automáticos multidocumentos mantêm os mesmos aspectos presentes nos sumários manuais. Neste artigo, apresenta-se uma avaliação de sumários automáticos multidocumento versus sumários manuais, utilizando um conjunto de aspectos pré-definido e a medida ROUGE. Trata-se de uma abordagem baseada em córpus que visa investigar a informatividade dos sumários automáticos. O artigo está organizado da seguinte forma: na Seção 2, descrevem-se os experimentos juntamente com os resultados obtidos, e, a Seção 3, contém algumas considerações finais.

2. Experimentos e Resultados

Para realizar avaliação de sumários automáticos versus sumários manuais, foi utilizado o córpus CSTNews. O córpus é composto de 50 conjuntos de textos jornalísticos em português do Brasil, agrupados por assunto; cada grupo tem sumários multidocumento manuais e automáticos. Do córpus, selecionou-se quatro clusters de notícias policiais e para cada sumário manual e automático, quatro linguistas computacionais anotaram as sentenças a partir de uma lista pré-determinada de aspectos definida na TAC 2010. Como a categoria de textos policiais não estava prevista explicitamente nas categorias da TAC 2010, os linguistas computacionais foram orientados a criar novos aspectos quando achassem necessário. O Quadro 1 lista os 12 aspectos encontrados nos sumários com uma breve definição, elaborada pelos próprios anotadores. Os aspectos com um asterisco foram criados pelos anotadores.

Aspecto	**Descrição**	**Aspecto**	**Descrição**
What	O que aconteceu	*Who affected*	Quem foi afetado de forma negativa. Não deve conter danos explícitos
Who	Pessoa ou entidade que é o foco do fato principal	*Whataffected**	O que foi afetado de forma negativa. Não deve conter danos explícitos
When	Data, hora ou outros marcadores temporais	*Perpetrator*	Indivíduos ou grupos responsáveis
Where	Localização física	*Damage*	Danos explícitos a alguém ou edificações
Why	Razões que causaram o fato principal	*Importance*	Importância do fato principal
How	Como aconteceu o fato principal	*History**	Contexto histórico sobre um fato

Quadro1– Aspectos encontrados nos sumários

O Quadro2 apresenta um exemplo de sumário anotado. Os aspectos estão em letras maiúsculas depois das sentenças a que eles se referem. Percebe-se, pelo exemplo, que uma sentença pode receber mais de um aspecto.

[Terminou a rebelião de presos no Centro de Custódia de Presos de Justiça (CCPJ), em São Luís, no começo da tarde desta quarta-feira (17).] WHAT/WHERE/WHEN [O motim começou durante a festa do Dia das Crianças.]HISTORY [Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.]HOW/WHO-AFFECTED [Alguns menores saíram desmaiados e foram conduzidos para o atendimento médico.]DAMAGES [Quatro pessoas teriam ficado feridas.]DAMAGES

Quadro2 – Exemplo de sumário anotado

As Tabelas 1 e 2 apresentam as frequências dos aspectos por cluster nos sumários manuais e nos sumários automáticos, respectivamente. Os sumários automáticos foram produzidos pelo sistema CSTSumm (Jorge e Pardo, 2010), que seleciona as sentenças mais importantes com base no modelo de relacionamento multidocumento CST - *Cross-document Structure Theory* (Radev, 2000).

Tabela 1 - Frequência de aspectos nos sumários manuais

Anotação de Aspectos - Sumários Manuais												
	What	**Where**	**When**	**Who-affected**	**What-Affected**	**Who**	**Why**	**How**	**Damage**	**Perpetrator**	**History**	**Importance**
C11	1	1	1	0	7	0	0	3	2	1	0	0
C37	1	1	1	1	0	0	0	1	2	0	1	0
C39	1	1	1	0	0	1	1	0	0	0	0	1
C45	1	1	1	0	0	1	1	0	0	2	0	0

Tabela 2 - Frequência de aspectos nos sumários automáticos

Anotação de Aspectos - Sumários Automáticos (CSTSumm)												
	What	**Where**	**When**	**Who-affected**	**What-Affected**	**Who**	**Why**	**How**	**Damage**	**Perpetrator**	**History**	**Importance**
C11	2	2	2	0	2	0	0	2	1	1	0	0
C37	1	1	1	1	0	0	0	1	0	0	0	0
C39	2	2	2	0	0	2	2	0	0	0	0	0
C45	1	1	2	0	0	1	0	0	0	5	1	0

Analisando os resultados das Tabelas 1 e 2, pode-se observar que os aspectos *What*, *Where* e *When* ocorreram em todos os pares de sumários (manual e automático), pois trazem as informações básicas e primordiais das notícias apresentadas. No entanto, em alguns sumários automáticos esses aspectos apareceram mais de uma vez (valores sombreados na Tabela 2), devido à redundância de informações presente em tais sumários. Os aspectos ocorreram sempre na primeira sentença de cada sumário, com

exceção de um único par em que *Where* ocorreu no final do texto. A redundância existente nos sumários automáticos se deve ao fato de que os sistemas de sumarização ainda coletam informações repetidas, principalmente se elas aparecem na forma de paráfrase.

Os aspectos *What-affected* e *Who-affected* foram encontrados em apenas 2 pares de sumários. Um desses pares (C11) tratava sobre uma série de ataques criminosos ocorridos em São Paulo e apresentava relatos de diversos prédios públicos, agências bancárias e bases policiais que teriam sido afetadas (*What-affected*). Vale notar que, das 7 ocorrências de *What-affected* encontradas no sumário humano, apenas 2 foram observadas também no sumário automático. O outro par de sumários (C37), por sua vez, tratava do término de uma rebelião de presos em São Luís (MA) e apontava 30 reféns que foram feridos durante o motim (*Who-affected*), sendo assim, nota-se que a aparição destes aspectos depende muito do tipo de ocorrência noticiada.

Já o aspecto *Who,* aparece quando o sumário cita especificamente um órgão ou indivíduo, como no caso do sumário em que aparece o Ministério Público e a Polícia Federal. *Why* e *How* são aspectos que apresentam informações adicionais que podem ou não aparecer nos sumários, bem como *History* e *Importance* que são consideradas informações mais genéricas.

Com relação ao aspecto *Damages*, nota-se que apesar dos sumários tratarem de notícias do âmbito policial, nem sempre acontecem danos ou ocorrências explicitamente. Por último, o aspecto *Perpetrators* aparece dependendo do tipo de ocorrência e mais especificamente quando uma investigação policial foi concluída ou caminha para uma fase final.

A análise de aspectos foi comparada com a avaliação ROUGE desses sumários, cujos valores são apresentados na Tabela 3. Além disso, também foi observado o número de aspectos em comum entre sumários manuais e automáticos. A Tabela 4 apresenta o número de aspectos presentes nos sumários manuais e quantos desses ocorrem nos sumários automáticos (excluindo a redundância).

Tabela 3 - Avaliação ROUGE

Automáticos x Manuais			
	Recall	**Precision**	**F-Measure**
C11	0.58772	0.58772	0.58772
C37	0.58491	0.45588	0.51240
C39	0.72414	0.51852	0.60432
C45	0.66379	0.53103	0.59003

Tabela 4 - Número de aspectos em cada cluster

	Manual	Automático
C11	16	9
C37	8	5
C39	6	5
C45	7	6

A cobertura (*recall*) indica o quanto do sumário manual é mantido no sumário automático. Pelas Tabelas 3 e 4, pode-se observar que os sumários com maior cobertura em relação a medida ROUGE são os que mantêm o maior número de aspectos em relação ao sumário manual. Neste caso, tais sumários pertencem aos clusters C39 e C45, pois estes contêm quase todos os aspectos dos seus respectivos sumários manuais. Por outro lado, os sumários dos clusters C11 e C37 apresentaram baixa cobertura e possuem bem menos aspectos dos seus sumários manuais.

3. Conclusões

Este artigo apresentou uma avaliação de sumários automáticos multidocumento versus sumários manuais. A avaliação objetivou verificar se os aspectos pré-definidos pela TAC 2010 estavam sendo reproduzidos nos sumários automáticos. Apesar de a amostra avaliada ter sido pequena, pode-se dizer, como resultado preliminar, que quanto maior a cobertura dos sumários automáticos, mais fiéis eles são em relação aos aspectos presentes dos sumários humanos. Os aspectos funcionam como guias para alcançar a informatividade esperada nos sumários automáticos. Como trabalhos futuros, pretende-se estender a avaliação para outros sumários do córpus CSTNews.

Referências

Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese.In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 1-18.October 26, Cuiabá-MT, Brazil.

Jorge, M.L.C. e Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82. July 16, Uppsala/Sweden.

Lin, C. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In the *Proceedings of the Workshop on Text Summarization Branches Out* (WAS 2004), Barcelona, Spain.

Owczarzak, K. e Dang, H. (2011). Who wrote What Where: Analyzing the content of human and automatic summaries. In the *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pp. 25-32.Junho, Portland, Oregon.

Zhang, R., Li, W., Gao, D. (2011). Generating Coherent Summaries with Textual Aspects.In *Proceedingsof AAAI 2012*.