

Segmentação Textual Automática de Atas de Reunião

Resumo—A tarefa de segmentação textual consiste em dividir um texto em porções com significado relativamente independente, de maneira que cada segmento contenha um assunto. Muitas pesquisas avaliam técnicas de segmentação textual usando textos longos com a concatenação de notícias de vários assuntos ou com a transcrição de discursos ou reuniões com múltiplos participantes. Além disso, a maior parte das pesquisas é realizada considerando o idioma inglês. Esse artigo tem como objeto de estudo atas de reunião escritas em português, as quais, além do idioma menos estudado sob essas técnicas, possuem estilo de redação mais formal e sucinto, o que desfavorece técnicas baseadas em frequência de palavras. A segmentação desses documentos facilita a sua organização e acesso, oferecendo vantagens no processo de recuperação de informação em relação à busca por palavras-chave, pois é possível retornar porções de texto menores que contenham assuntos relevantes à intenção do usuário, ao invés de destacar as palavras-chave em um texto longo que pode conter informações menos pertinentes. Este artigo concentra-se nos principais algoritmos da literatura adaptando-os ao contexto das atas de reunião em português, bem como encontrar uma técnica que ofereça segmentos coesos e contribua para aprimorar sistemas de recuperação de informação para que respondam melhor à buscas do usuário. Para a avaliação experimental, um conjunto de atas do departamento de pós-graduação da UFSCar-Sorocaba foi manualmente segmentado por participantes das reuniões. Então, as divisões manuais foram comparadas com as divisões geradas automaticamente a fim de encontrar uma configuração que melhor retorne segmentos coesos de atas de reunião. Ao final, chegou-se a uma técnica capaz de segmentar as atas com performance similar a outros trabalhos da literatura.

I. INTRODUÇÃO

Reuniões são tarefas presentes em atividades corporativas, ambientes de gestão e nas organizações de um modo geral. O conteúdo das reuniões é frequentemente registrado em texto na forma de atas para fins de documentação e posterior consulta. A organização e consulta manual desses arquivos torna-se uma tarefa custosa, especialmente considerando o seu crescimento em uma instituição.

Devido a fatores como a não estruturação e volume dos textos, a localização de assunto é uma tarefa custosa. Usualmente, o que se faz são buscas manuais guiadas pela memória ou com uso de ferramentas computacionais baseadas em localização de palavras-chave.

Buscas por palavras-chave normalmente exigem a inserção de termos exatos que devem necessariamente estar contidos no trecho onde está o assunto. Então, apresentam um documento com as palavras buscadas em destaque, mantendo o texto longo. Assim, dificultando o ranqueamento por relevância. Além disso, não evita que textos próximos sejam retornados.

As atas são documentos textuais e portanto constituem dados não estruturados. Assim, um sistema que responde a consultas do usuário ao conteúdo das atas, retornando trechos de textos relevantes à sua intenção, é um desafio que envolve a compreensão de seu conteúdo [5]. Uma vez que a ata registra a sucessão de assuntos discutidos na reunião, há interesse em um sistema que aponte trechos de uma ata que tratam de um assunto específico. Tal sistema tem duas principais tarefas: descobrir quando há uma mudança de assunto e quais são esses assuntos [1]. Este trabalho tem foco principal na primeira tarefa, detecção de mudança de assuntos, que pode ser atendida pela segmentação automática de textos.

A tarefa de segmentação textual consiste dividir um texto em partes que contenham um significado relativamente independente. Em outras palavras, é identificar as posições onde há uma mudança significativa de tópicos.

A segmentação de textos é útil em aplicações que trabalham com textos sem indicações de quebras de assunto, ou seja, não apresentam seções ou capítulos, como transcrições automáticas de áudio, vídeos e grandes documentos que contêm vários assuntos como atas de reunião e notícias.

O interesse por segmentação textual tem crescido em em aplicações voltadas a recuperação de informação [17] e sumarização de textos [3] [4]. Essa técnica pode ser usada para aprimorar o acesso a informação quando essa é solicitada por meio de uma consulta, onde é possível oferecer porções menores de texto mais relevantes ao invés de exibir um documento maior que pode conter informações menos pertinentes. Além disso, encontrar pontos onde o texto muda de assunto, pode ser útil como etapa de pré-processamento em aplicações voltadas ao entendimento do texto, principalmente em textos longos. A navegação pelo documento pode ser aprimorada, em especial na utilização por usuários com deficiência visual, os quais utilizam sintetizadores de texto como ferramenta de acessibilidade [8]. A sumarização de texto pode ser aprimorada ao processar segmentos separados por tópicos ao invés de documentos inteiros.

Frequentemente atas de reunião têm a característica de apresentar um texto com poucas quebras de parágrafo e sem marcações de estrutura, como capítulos, seções ou quaisquer indicações sobre o tema do texto. As atas de reunião diferem dos textos comumente estudados em outros trabalhos em alguns pontos. O estilo contribui para a escrita de textos sucintos com poucos detalhes, pois o ambiente dá preferência a textos curtos. Segundo Choi [9], o segmentador tem a acurácia reduzida em segmentos curtos (em torno de 3 a 5 sentenças).

Há também um maior foco no idioma inglês, presente

na maioria dos artigos publicados. Embora haja abordagens voltadas para outros idiomas, falta ainda uma maior atenção na literatura sobre língua portuguesa e a documentos com características próprias como as atas de reunião. Diferenças de performance podem ser vistas no mesmo algoritmo quando aplicando em documentos de diferentes idiomas, onde o inglês apresenta um taxa de erro significativamente menor que outro como o alemão e o espanhol [13]. Assim, esse trabalho trata da adaptação e avaliação de algoritmos tradicionais ao contexto de documentos em português do Brasil, com ênfase especial nas atas de reuniões.

O restante deste artigo está dividido da seguinte forma. Na Seção II o processo de segmentação textual é melhor descrito, bem como os principais algoritmos da literatura. Na Seção III são apresentados trabalhos recentes que desenvolveram as técnicas existentes em contextos específicos. Na Seção IV as atas são descritas em suas principais características. Na Seção V é apresentada a proposta desse trabalho voltada às atas de reunião. Na Seção VI é detalhado a avaliação dos experimentos e seus resultados reportados na Seção VI-C. Por fim, na Seção VII são apresentadas as considerações finais e trabalhos futuros.

II. REFERENCIAL TEÓRICO

Um documento textual, sobre tudo quando longo, é frequentemente uma sucessão de assuntos. A segmentação textual ou segmentação topical é a tarefa de dividir um texto mantendo em cada parte um tópico com seu significado completo.

O texto deve ser previamente dividido em unidades de informação, que podem ser palavras, paragrafos e mais usualmente em sentenças. Essas unidades são processadas pelos algoritmos como a menor parte de um segmento, ou seja uma unidade não pode dividir, mas deve pertencer integralmente a um segmento. Por exemplo, algoritmos baseados em janelas deslizantes, avançam uma unidade a cada passo. Da mesma forma, algoritmos baseados em matrizes de similaridade, devem calcular a similaridade entre essas unidades. Usualmente usa-se sentenças como unidades de informação, ou seja, as janelas deslizam uma sentença a cada passo e matrizes de similaridades armazenam as similaridades entre as sentenças do texto. Assim, cada ponto entre duas unidades de informação é considerado um candidato a limite entre segmentos. Dessa forma, um segmento pode ser visto como uma sucessão de unidades de informação que compartilham o mesmo assunto.

Trabalhos anteriores se apoiam na ideia de que a mudança de tópicos em um texto é acompanhada de uma proporcional mudança de vocabulário, essa ideia, chamada de coesão léxica, sugere que a distribuição das palavras é um forte indicador da estrutura do texto. A partir disso, vários algoritmos foram propostos baseados na ideia de que um segmento pode ser identificado e delimitado pela análise das palavras que o compõe [10] [3].

Uma vez que a coesão léxica é pressuposto básico da maioria dos algoritmos, o cálculo da similaridade entre unidades de informação (comumente sentenças) é fundamental. Uma medida de similaridade frequentemente utilizada é o cosseno,

apresentada na Equação 1, onde $f_{x,j}$ a frequência da palavra j na sentença x e $f_{y,j}$ sendo a frequência da palavra j na sentença y .

$$Sim(x, y) = \frac{\sum_j f_{x,j} \times f_{y,j}}{\sqrt{\sum_j f_{x,j}^2 \times \sum_j f_{y,j}^2}} \quad (1)$$

A. Principais algoritmos

Entre os principais trabalhos da literatura podemos citar o *TextTiling* [11] e o *C99* [8]. O *TextTiling* é um algoritmo baseado em janelas deslizantes, onde para cada candidato a limite, analisa-se o texto circundante. Um limite ou quebra de segmento é identificado quando sempre que a similaridade cai abaixo de um limiar.

O *TextTiling* recebe uma lista de candidatos a limite, usualmente finais de parágrafo ou finais de sentenças. Para cada posição candidata são construídos 2 blocos, um contendo sentenças que a precedem e outro com as que a sucedem. O tamanho desses blocos é um parâmetro a ser fornecido ao algoritmo e determina o tamanho mínimo de um segmento. Em seguida, os blocos de texto são representados por vetores que contém as frequências de suas palavras. Então, usa-se cosseno (Equação 1) para calcular a similaridade entre os blocos adjacentes a cada candidato e identifica-se uma transição entre tópicos pelos vales na curva de dissimilaridade.

O *TextTiling* possui menor complexidade computacional. Por outro lado, algoritmos mais complexos apresentam acurácia relativamente superior como apresentado em [8], [13], [14].

Choi [8] apresenta um algoritmo baseado em ranking, o *C99*. Embora muitos trabalhos utilizem matrizes de similaridades, o autor traz observações. Ele aponta que para pequenos segmentos, o cálculo de suas similaridades não é confiável, pois uma ocorrência adicional de uma palavra causa um impacto que pode alterar significativamente o cálculo da similaridade. Além disso, o estilo da escrita pode não ser constante em todo o texto. Choi sugere que, por exemplo, textos iniciais dedicados a introdução costumam apresentar menor coesão do que trechos dedicados a um tópico específico. Portanto, comparar a similaridade entre trechos de diferentes regiões não é apropriado. Devido a isso, as similaridades não podem ser comparadas em valores absolutos. Então, o autor contorna esse problema fazendo uso de *rankings* de similaridade para encontrar os segmentos de texto.

Inicialmente é construída uma matriz que contém as similaridades de todas as unidades de texto. Em seguida, cada valor na matriz de similaridade é substituído por seu ranking local. Onde para cada elemento da matriz, seu *ranking* é o número de elementos vizinhos com valor de similaridade menor que o seu.

Então, para cada elemento delimita-se uma região da matriz, denominada máscara, para comparar com seus vizinhos. Na Figura 1 é apresentado um quadro de dimensões 3 x 3 destacado na matriz de similaridades, que contém os valores $\{0, 3; 0, 4; 0, 4; 0, 6; 0, 5; 0, 2; 0, 9; 0, 5; 0, 7\}$, onde cada elemento da matrix é a similaridade entre duas unidades de

informação. Tomando como exemplo o elemento com valor 0,5, a mesma posição na matriz de *ranks* terá o valor 4, pois esse é o número de vizinhos com valores inferiores a 0,5 dentro do quadro analisado na matriz de similaridades.

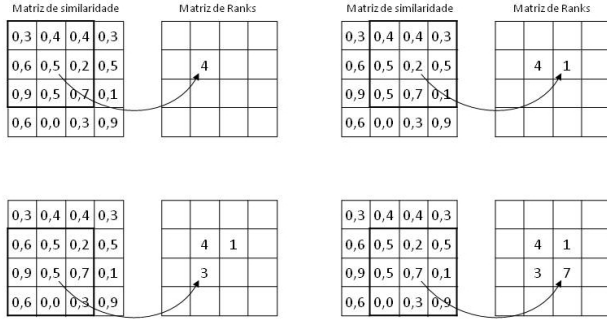


Figura 1: Exemplo de construção de uma matriz de rank. [8]

Finalmente, utiliza um método de divisão por *clustering* baseado no algoritmo de maximização de Reynar [16] para identificar os limites entre os segmentos.

B. Medidas de Avaliação

As medidas de avaliação tradicionais como acurácia e precisão são usadas em recuperação de informação e classificação automática para medir o desempenho de modelos de classificação e predição. Essas medidas são baseadas na comparação dos valores produzidos por uma hipótese com os valores reais. Uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo. Na Tabela ...

	Predição Positiva	Predição Negativa
Positivo real	VP (Verdadeiro Positivo)	FN (Falso Negativo)
Negativo real	FP (Falso Positivo)	VN (Verdadeiro Negativo)

Tabela I: Matriz de confusão.

As medidas de avaliação tradicionais baseiam se na contagem de acertos. Quando um fim de segmento extraído automaticamente coincide com o limite de referência, têm-se um acerto.

Essas medidas de avaliação tentam computar os erros do algoritmo, isto é falsos positivos e falsos negativos, a fim de calcular sua eficiência. Um falso positivo é um limite identificado pelo algoritmo que não corresponde a nenhum limite na segmentação de referência. Um falso negativo é quando o algoritmo não identifica um limite existente na segmentação de referência.

Nesse sentido, a precisão, que é a proporção de limites corretamente identificados pelo algoritmo, e a revocação, que é a proporção de limites verdadeiros que foram identificados pelo algoritmo, trazem alguns problemas na avaliação de segmentadores automáticos.

Conforme o algoritmo aponta mais segmentos no texto, este tende a melhorar a revocação e ao mesmo tempo, reduzir a precisão. Esse problema de avaliação pode ser contornado

utilizando a medida F^1 que é uma média harmônica entre precisão e revocação onde ambas tem a o mesmo peso.

Além dessas medidas, que consideram apenas se um segmento foi corretamente definido, pode-se também considerar a distância entre o segmento extraído automaticamente e o segmento de referência [13].

Na Figura 2 é apresentado um exemplo com duas segmentações extraídas automaticamente e uma referência. Na Figura 3, em ambos os casos não há nenhum verdadeiro positivo, o que implica em zero para os valores de precisão, acurácia, e revocação, embora a segunda hipótese possa ser considerada superior à primeira se levado em conta a proximidade dos limites.

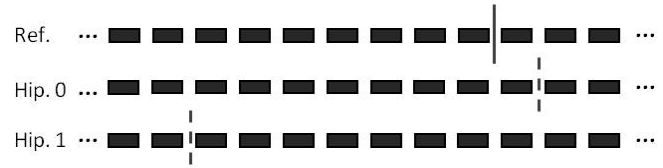


Figura 2: Exemplos de *near missing* e falso positivo puro. Os blocos indicam uma unidade de informação e as linha verticais representam os limites entre segmentos de texto representando um tópico do texto.

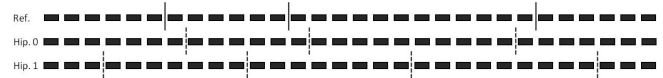


Figura 3: Exemplo de duas segmentações hipotéticas em comparação a uma ideal.

Considerando o conceito de *near misses*, algumas soluções foram propostas. As medidas de avaliação mais utilizadas são:

- 1) Em [2] é apresentada uma medida denominada P_k , a qual atribui valores parciais a *near misses*, ou seja, limites sempre receberão um peso proporcional à sua proximidade, desde que dentro de um janela de tamanho k . Para isso, esse método move uma janela de tamanho k e a cada posição verifica se o início e o final da janela estão ou não dentro do mesmo segmento e penaliza o algoritmo em caso de discrepância. Ou seja, dado duas palavras de distancia k , uma discrepância é computada quando o algoritmo e a referência não concordam se as palavras estão ou não no mesmo segmento. O valor de k é calculado como a metade da média dos comprimentos dos segmentos reais. Como resultado, é retornado a contagem de discrepâncias dividida pelo quantidade de segmentações analisadas. P_k é uma medida de dissimilaridade entre as segmentações e pode ser interpretada como a probabilidade de duas sentenças extraídas aleatoriamente pertencerem ao mesmo segmento.
- 2) *WindowDiff*. Pevzner [15] aponta problemas na avaliação mais tradicional, P_k [2]. Eles apontam que esse método penaliza demasiadamente os falsos negativos em relação

aos falsos positivos e a *near misses*, além disso, desconsidera o tamanho dos segmentos. Como solução, propõem um método, que traz duas diferenças principais: a dobra da penalidade para os falsos positivos a fim de diminuir o problema da subestimação dessa medida e, diferente de P_k , ao mover a janela pelo texto, penaliza-se o algoritmo sempre que o número de limites proposto pelo algoritmo não coincidir com o número de limites esperados para aquela janela de texto.

Com isso, demonstram em seu trabalho que, em relação a P_k , *WindowDiff* consegue resolver seus principais problemas e mantém sua proposta inicial de sensibilidade a *near misses*, penalizando-os menos que os falsos positivos puros.

III. TRABALHOS RELACIONADOS

Semelhante a este trabalho, outras abordagens foram propostas no sentido de propor segmentadores para outros idiomas além do inglês bem como alviá-los em contextos específicos como discursos e conversas em reuniões.

Diferenças de performance podem ser vistas no mesmo algoritmo quando aplicando em documentos de diferentes idiomas. Se aplicado ao inglês, apresenta um taxa de erro significativamente menor do que quando aplicado a outro idioma como o alemão e o espanhol [13].

A fim de adaptar os algoritmos *TextTiling* e *C99* ao idioma árabe, foi utilizado como *corpus* a concatenação de textos extraídos de notícias de diferentes países como Tunísia, Egito e Algeria, que tratam de assuntos distintos como política, esporte, cultura, história, tecnologia e artes, e que trazem particularidades nos estilos de escrita e até diferenças entre dialetos locais.

Foram propostos ajustes na etapa de préprocessamento onde verificou-se que diferenças no dialeto de cada país devem ser consideradas no processo de segmentação e que a adaptação depende da escolha de um *stemmer* adequado à cada dialeto [7].

Outra aplicação para essas técnicas é a segmentação de conversas em reuniões com múltiplos participantes onde se estuda os textos extraídos dos discursos, ou seja, o texto a ser segmentado é uma transcrição das falas dos participantes durante a reunião. Nesse sentido, como parte do projeto *CALOMA*¹, foi apresentado um segmentador baseado no *TextTiling*, que foi aplicado em reuniões com múltiplos participantes. Os autores apoiaram-se em elementos da fala como pausas, trocas de falantes e entonação para encontrar melhores segmentos [10]. O *corpus* utilizado contém a transcrição da fala dos participantes durante as reuniões que foram conduzidas por um mediador que propunha os tópicos e anotava o tempo onde os participantes mudavam de assunto [1] [18].

Ainda no contexto de reuniões com múltiplos participantes temos a segmentação funcional dos discursos, onde outros aspectos podem ser analisados, como a participação dos presentes na reunião. Nesse sentido, estuda-se a contribuição

das pessoas as quais podem ser categorizadas, por exemplo, em diálogos, discussões e monólogos. Sugere-se que alguns comportamentos podem dar pistas de mudança de discurso, como quando um participante toma a palavra por um tempo prolongado [5].

A presença de *pistas* pode ser um indicativo que ajuda a aprimorar a detecção de limites entre segmentos, uma vez que alguns elementos são frequentes na transição de tópicos. Essas *pistas* podem ser palavras como “Ok”, “continuando” e frases como “Boa noite”, “Dando prosseguimento” ou pausas prolongadas, ajudam a indicar uma mudança de tópicos. Essas *pistas* podem ser detectadas por meio de algoritmos de aprendizado de máquina ou anotadas manualmente [19] [10] [2].

Quanto a eficácia dos algoritmos, observou-se que aqueles que apresentam melhor performance o fazem ao custo de maior complexidade computacional, que se deve muitas vezes à construção de matrizes de similaridade entre todas as sentenças como em [8], nesses casos pode-se diminuir o tempo computacional evitando calcular a similaridade entre trechos muito distantes no texto.

Nesse sentido, foi apresentada uma abordagem que otimiza o processo ao computar as médias das similaridades entre sentenças de cada bloco, a qual chamou de *inner similarity* e em seguida usou esses valores para calcular as médias das similaridades entre todos os blocos a qual chamou de *outer similarity* [13]. Dessa forma não é criada uma matriz que contem as similaridades de todas as sentenças, mas apenas daquelas mais próximas. Os autores reportaram uma eficiência superior e uma eficácia comparável aos algoritmos mais complexos.

Para a obtenção do *corpus*, bem como segmentações de referência, muitos trabalhos utilizam a concatenação de textos na avaliação, os quais foram extraídos de coleções de documentos como *TDI*² e *WSJ*³. Nesses casos, tem-se facilmente os segmentos de referência onde cada limite de segmento separa os documentos originais. Outros avaliam seus trabalhos utilizando coleções de transcrições de áudios gravados durante conversas de reuniões entre pessoas como *ICSI* [12] e *AMI* [6]. Nesses casos, os segmentos de referência são dados por anotações feitas durante a reunião, onde registrou-se o tempo em que os participantes trocaram de assunto.

IV. CONJUNTO DE DOCUMENTOS

Para a condução desse trabalho é necessário que hajam documentos a serem segmentados automaticamente e que se conheça uma segmentação de referência para compará-los e avaliar a proposta. A seguir, são apresentadas as características particulares das atas em relação a outros tipos de textos e a segmentação de referência a ser utilizada.

A. Características das atas

As atas de reunião diferem dos textos comumente estudados em outros trabalhos em alguns pontos. O estilo contribui para a escrita de textos sucintos com poucos detalhes, pois

¹<http://www.ai.sri.com/project/CALO>

²<https://catalog.ldc.upenn.edu/LDC98T25>

³<https://catalog.ldc.upenn.edu/ldc2000t43>

o ambiente dá preferência a textos curtos. Segundo Choi [9], o segmentador tem a acurácia reduzida em segmentos curtos (em torno de 3 a 5 sentenças).

Para evitar um texto monótono à leitura, ao redigir o documento tem-se o cuidado de não repetir ideias e palavras em favor da elegância do documento. Tal característica enfraquece a coesão léxica e portanto o cálculo da similaridade é prejudicado. Por exemplo, duas sentenças diferem se uma contiver a palavra *computadores* e na seguinte *equipamentos*, mesmo que se refiram à mesma ideia. Além disso, o documento compartilha um certo vocabulário próprio do ambiente onde os assuntos são discutidos e com isso nota-se que os segmentos, embora tratem de assuntos diferentes, são semelhantes em vocabulário.

A presença de ruídos como cabeçalhos, rodapés e numeração de páginas e linhas prejudicam tanto similaridade entre sentenças como a apresentação final ao usuário. Porém, esses ruídos podem ser reduzidos ou eliminados como mostrado na Subseção V-A, sobre pré-processamento.

A maioria dos trabalhos enquadram-se em duas categorias no que se refere ao *corpus* estudado. A primeira utiliza a concatenação de textos de diferentes fontes e assuntos. Ocorre que essas características favorecem o processo de segmentação uma vez que os documentos produzidos pela concatenação contém tópicos obtidos de domínios distintos que compartilham pouco vocabulário. Por outro lado, essas características não estão presentes no contexto das atas, onde um documento é redigido por um mesmo autor, e todos tópicos foram produzidos no mesmo domínio, e dessa forma, têm estilo de escrita e vocabulário similares.

A segunda estuda a segmentação automática de textos obtidos da transcrição de áudios como de vídeos e de gravações de reuniões com grupos de pessoas. A transcrição da fala de pessoas, difere desse trabalho, pois este possui características típicas de um estilo formal, por exemplo, o texto sucinto, onde o autor se limita a relatar apenas o essencial de uma discussão e omite detalhes da discussão, poupando o leitor de diálogos repetitivos ou falas durante a transição de tópicos, o que resulta em segmentos mais curtos os quais desfavorecem algoritmos baseados em coesão léxica [8].

B. Segmentação de Referência

Para que se possa avaliar um segmentador automático de textos é preciso uma referência, isto é, um texto com os limites entre os segmentos conhecidos. Essa referência, deve ser confiável, sendo uma segmentação legítima que é capaz de dividir o texto em porções relativamente independentes, ou seja, uma segmentação ideal.

Entre as abordagens mais comuns para se conseguir essas referências, encontramos:

- A concatenação aleatória de documentos distintos, onde o ponto entre o final de um texto e o início do seguinte é um limite entre eles.
- A segmentação manual dos documentos, nesse caso, pessoas capacitadas, também chamadas de juízes, ou mesmo

o autor do texto, são consultadas e indicam manualmente onde há uma quebra de segmento.

- Em transcrição de conversas faladas em reuniões com múltiplos participantes, um mediador é responsável por encerrar um assunto e iniciar um novo, nesse caso o mediador anota manualmente o tempo onde há uma transição de tópico.

Em aplicações onde a segmentação é uma tarefa secundária e quando essas abordagens são custosas ou não se aplicam, é possível, ao invés de avaliar o segmentador, analisar seu impacto na aplicação final.

V. PROPOSTA: SEGMENTAÇÃO LINEAR AUTOMÁTICA DE ATAS DE REUNIÃO

Os algoritmos *TextTiling* e *C99* foram propostos para o inglês, independentemente de domínio, ou seja, a proposta inicial dos autores é trabalhar em qualquer texto nessa língua. Assim, propõe-se adaptá-los ao contexto das atas de reunião em português do Brasil, ou seja, em uma língua diferente e dentro de um contexto específico. As subseções seguintes tratam das adaptações para esse nicho mais específico. A Seção VI mostra a análise dos algoritmos adaptados.

O vocabulário das reuniões, ainda que em tópicos diferentes, compartilham certo vocabulário pertencente ao ambiente onde se deram as reuniões. Isso é um fator que diminui a coesão léxica entre os segmentos. As atas de reunião costumam ter um estilo de escrita que deve ser levado em conta na adaptação do algoritmos, como a identificação de finais de sentença na ausência de quebras de parágrafo, inserção de linhas que não separam assuntos, utilização de pontuação para transição de tópicos, e cabeçalhos e numerais ruidosos.

Nas subseções a seguir serão apresentados o pré-processamento e a identificação de segmentos candidatos considerados para a segmentação de atas.

A. Pré-processamento

O texto a ser segmentado frequentemente é extraído de documentos em formatos como *pdf*, *doc*, *docx* ou *odt*. Após a extração, uma ata tem em média 906 *tokens*, incluindo elementos menos informativos, a fim selecionar as palavras mais significativas, o texto passa por processos de transformação os quais serão apresentados a seguir.

- 1) Remoção de Cabeçalhos e rodapés: As atas frequentemente contém trechos que podem ser considerados pouco informativos e descartados durante o pré-processamento, como cabeçalhos e rodapés que se misturam aos tópicos tratados na reunião, podendo ser inseridos no meio de um tópico e criando uma quebra que prejudica tanto o algoritmo de segmentação, quanto a leitura do texto pelo usuário.
- 2) Identificação de finais sentenças: Cada final de sentença é identificado e marcado com uma *string* especial, esse processo é melhor descrito na Subseção V-B.
- 3) Limpeza: Todos os caracteres são convertidos em caixa baixa e elimina-se a acentuação, sinais de pontuação e todos os *tokens* menores que três caracteres.

- 4) Remoção de numerais: Nas atas, é comum o uso de numerais para marcação de páginas e linhas, da mesma forma, são pouco informativos e podem ser descartados.
- 5) Remoção de *Stop Words*: Remove-se as palavras consideradas menos informativas, para isso, utiliza-se uma lista contendo 438 palavras.
- 6) *Stemming*: Extrai-se o radical de cada palavra, para isso, as letras são convertidas em caixa baixa e aplica-se o algoritmo *Orengo*⁴ para remoção de sufixos.

Após cada passo, os *tokens* menos significativos são removidos e, ao final, são extraídas em média 441 palavras de cada ata, conforme mostrado na Tabela II.

Processo	Média de Tokens
Extração do texto	906
Remoção de Cabeçalhos Rodapés	813
Identificação de finais de sentença	813
Limpeza	535
Remoção de Numerais	526
Remoção de <i>Stop Words</i>	441
<i>Stemming</i>	441

Tabela II: Quantidade média de *tokens* extraídos após cada passo do pré-processamento.

A Figura 4 mostra a etapa de pré-processamento em uma sentença em português.

B. Identificação de candidatos

É preciso fornecer aos algoritmos os candidatos iniciais a limites de segmento. Para isso, é necessário escolher qual será a unidade de informação mínima que constitui um segmento. Baseando-se no estilo de escrita e considerando as pontuações de um texto, é possível, em alguns casos, indicar quebras de parágrafo, finais de sentenças ou mesmo palavras como elementos que encerram um segmento.

Ocorre que em atas de reunião é uma prática comum redigir as de forma que o conteúdo discutido fica em parágrafo único, além disso, as quebras de parágrafo são usados para formatação de outros elementos como espaço para assinaturas. Também não é conveniente indicar todo ponto entre *token* como candidato pois obrigaria a ajustar posteriormente os segmentos de maneira a não quebrar uma ideia ou frase. Assim, neste trabalho, os finais de sentença são considerados unidades de informação e portanto, passíveis a limite entre segmentos.

Devido ao estilo de pontuação desses documentos, como encerrar sentenças usando um “;” e inserção de linhas extras, foram usadas as regras apresentadas no Algoritmo 1 para identificar os finais de sentenças.

VI. AVALIAÇÃO EXPERIMENTAL

De acordo com [15] há duas principais dificuldades na avaliação de segmentadores automáticos. A primeira é conseguir uma referência, já que juízes humanos costumam não concordar entre si, sobre onde os limites estão e outras abordagens podem não se aplicar ao contexto. A segunda é

Algoritmo 1: Identificação de finais de sentença

Entrada: Texto

Saída : Texto com identificações de finais de sentença

```

1 para todo token, marcá-lo como final de sentença se:
2   Terminar com um !
3   Terminar com um . e não for uma abreviação
4   Terminar em . ? ; e:
5     For seguido de uma quebra de parágrafo ou
       tabulação
6     O próximo token iniciar com ( { [ " '
7     O próximo token iniciar com letra maiúscula
8     O penúltimo caracter for ) } ] " '
9 fim

```

que tipos diferentes de erros devem ter pesos diferentes de acordo com a aplicação. Há casos onde certa imprecisão é tolerável e outras, como a segmentação de notícias, onde a precisão é mais importante.

Para fins de avaliação desse trabalho, um bom método de segmentação é aquele cujo resultado melhor se aproxima do ideal, sem a obrigatoriedade de estar perfeitamente alinhado com tal. Ou seja, visto o contexto das atas de reunião, e a subjetividade da tarefa, não é necessário que os limites entre os segmentos (real e hipótese) sejam idênticos, mas que se assemelhem em localização e quantidade.

As próximas subseções mostram o conjunto de atas e a segmentação usada como referência em seguida são apresentadas as métricas de avaliação utilizadas neste trabalho e os testes realizados para avaliar os métodos.

A. Bases

A fim de obter um conjunto de documentos segmentados que possam servir como referência na avaliação, seis atas de reunião foram coletadas junto ao departamento de pós-graduação da UFSCar-Sorocaba⁵. Os documentos foram oferecidos à profissionais que participam de reuniões desse departamento e por meio de um *software* segmentaram o texto das atas conforme o julgamento de cada um. Os segmentos gerados manualmente foram comparados à segmentação automática conforme os critérios descritos a seguir.

B. Configuração experimental

As implementações dos algoritmos permitem ao usuário a configuração de seus parâmetros. O *TextTiling* permite ajustarmos dois parâmetros, sendo, o tamanho da janela (distância entre a primeira e a última sentença) para o qual atribuiu-se os valores 20, 40 e 60. Para o segundo parâmetro, o passo (distância que a janela desliza), atribuiu-se os valores 3, 6, 9 e 12. Gerando ao final 20 modelos.

O *C99* permite ajustarmos três parâmetros, sendo, a quantidade segmentos desejados, o qual é calculado como uma proporção dos candidatos a limite. Para isso atribuiu-se as proporções 0,2; 0,4; 0,6; 0,8; 1,0. O segundo parâmetro, o

⁴<http://www.inf.ufrgs.br/viviane/rspl/>

⁵http://www.pgccs.net/?page_id=1150

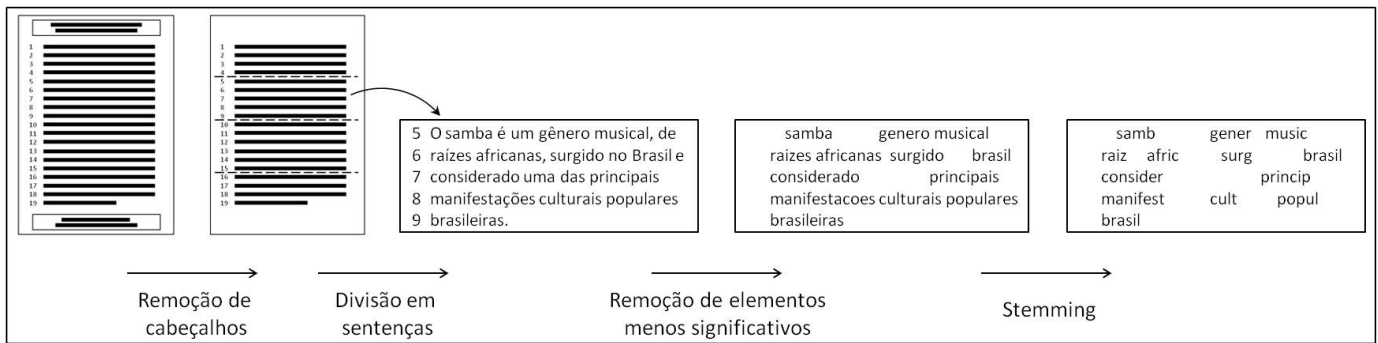


Figura 4: Exemplo de pré-processamento.

tamanho do quadro utilizado para gerar a matriz de ranking, atribuiu-se os valores 9 e 11, sendo 11 o valor padrão da apresentado pelo autor. Permite ainda, definirmos se as sentenças serão representados por vetores contendo a frequência ou o peso de cada termo, onde ambas as representações foram utilizadas. Considerando todos os parâmetros, foram gerados 20 modelos para o algoritmo C99.

C. Resultados



Em Construção

VII. CONCLUSÃO

As atas de reunião, objeto de estudo desse artigo, apresentam características peculiares em relação à discursos em reuniões e textos em geral. Características como segmentos curtos e coesão mais fraca devida ao estilo que evita repetição de palavras e ideias em benefício da leitura por humanos, dificulta o processamento por computadores.

Os algoritmos *TextTiling* e *C99* foram testados em um conjunto de atas reais coletadas do departamento de computação da UFSCar-Sorocaba. Por meio da análise dos dados chegou-se a um modelo cujos segmentos melhor se aproximaram as amostras de participantes das reuniões. Obteve-se resultados comparáveis aos vistos em discursos longos, porém um pouco inferiores, o que pode ser justificado pelo estilo peculiar de escrita das atas.

A segmentação de atas de reunião pode ajudar na organização, busca e compreensão dos conteúdos nelas contidos. Também outros domínios e aplicações diferentes podem se beneficiar dos resultados apresentados, como aplicações voltadas a resgate de informação, sumarização e acessibilidade. Assim, espera-se que outros trabalhos possam aproveitar deste.

Em trabalhos futuros, serão investigadas técnicas para descrever os segmentos e com isso aprimorar o acesso ao conteúdo das atas de reunião.

REFERÊNCIAS

- [1] S. Banerjee and A. Rudnicky. A texttiling based approach to topic boundary detection in meetings. volume 1, pages 57–60, 2006. cited By 3.
- [2] D. Beferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210, 1999.
- [3] B. K. Boguraev and M. S. Neff. Discourse segmentation in aid of document summarization. In *In Proceedings of Hawaii Int. Conf. on System Sciences (HICSS-33), Minitrack on Digital Documents Understanding, IEEE*, 2000.
- [4] B. K. Boguraev and M. S. Neff. Lexical cohesion, discourse segmentation and document summarization. In *In RIAO-2000, Content-Based Multimedia Information Access*, 2000.
- [5] M. H. Bokaei, H. Sameti, and Y. Liu. Linear discourse segmentation of multi-party meetings based on local and global information. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(11):1879–1891, Nov. 2015.
- [6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction, MLMI'05*, pages 28–39, Berlin, Heidelberg, 2006. Springer-Verlag.
- [7] A. H. Chaibi, M. Naili, and S. Sammoud. Topic segmentation for textual document written in arabic language. *Procedia Computer Science*, 35:437 – 446, 2014.
- [8] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000*, pages 26–33, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
- [9] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *In Proceedings of EMNLP*, pages 109–117, 2001.
- [10] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03*, pages 562–569, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [11] M. A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94*, pages 9–16, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.
- [12] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 364–367, 2003.
- [13] R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. pages 167–171, 2009. cited By 10.

- [14] H. Misra, F. Yvon, J. M. Jose, and O. Cappe. Text segmentation via topic modeling: An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1553–1556, New York, NY, USA, 2009. ACM.
- [15] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002. cited By 154.
- [16] J. C. Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, Philadelphia, PA, USA, 1998. AAI9829978.
- [17] J. C. Reynar. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 357–364, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.
- [18] G. Tur, A. Stolcke, L. Voss, S. Peters, D. Hakkani-Tür, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang. The calo meeting assistant system. *Trans. Audio, Speech and Lang. Proc.*, 18(6):1601–1611, Aug. 2010.
- [19] P. yun Hsueh, J. D. Moore, and S. Renals. Automatic segmentation of multiparty dialogue. In *EACL*, 2006.