

# Segmentação topical automática de atas de reunião

Ovídio José Francisco  
ovidiojf@gmail.com

## RESUMO

### Keywords

## 1. INTRODUÇÃO

Frequentemente atas de reunião tem 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.

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.

É útil em aplicações que trabalham com textos sem quebras de assunto, ou seja, não apresentam parágrafos, seções ou capítulos, como transcrições automáticas de áudio e grandes documentos que contêm assuntos não idênticos como atas de reunião e notícias.

O interesse por segmentação textual tem crescido em em aplicações voltadas a recuperação de informação e sumarização de textos. Essa técnica pode ser usada para aprimorar o acesso a informação quando essa é solicitada por um usuário por meio de uma consulta, onde é possível oferecer porções menores de texto mais relevante ao invés de exibir um documento maior que pode conter informações menos pertinente. A sumarização de texto também pode ser aprimorada ao processar segmentos separados por tópicos ao invés de documentos inteiros.

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.

## 2. TRABALHOS RELACIONADOS

Os principais algoritmos de segmentação textual baseiam-se na ideia de coesão léxica entre assuntos. Isto é, a mudança de tópicos é acompanhada de uma proporcional mudança de vocabulário. A partir disso, vários algoritmos foram propostos. Dessa forma, assumem o pressuposto que um segmento pode ser identificado e delimitado pela análise das palavras

que o compõe.

Uma vez que coesão léxica é pressuposto básico da maioria dos algoritmos, o cálculo da similaridade entre textos é fundamental. Uma medida de similaridade frequentemente utilizada é a *cosine*, a qual pode ser vista na equação 1, sendo  $f_{x,j}$  a frequência da palavra  $j$  na sentença  $x$  e  $f_{y,j}$  sendo a frequência da palavra  $j$  na sentença  $y$ .

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

Entre os trabalhos mais influentes podemos citar o *Text-Tiling* [4] proposto por Hearst. Ela propõe um algoritmo baseado em janelas deslizantes, onde para cada candidato a limite, analisa-se o texto circundante. Um limite ou quebra se segmento é identificado quando a similaridade entres os blocos apresenta uma queda considerável.

Ela propõe um algoritmo baseado em janelas deslizante, para analisar blocos de texto adjacentes e identificar os limites com base nas similaridades dos blocos.

O algoritmo 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 contêm as frequências de suas palavras. Então, usa-se *cosini* (equação 1) para calcular a similaridade entre os blocos.

Finalmente, os limites são identificados sempre que a similaridade entre blocos adjacentes entre cada candidato ultrapassa um determinado *threshold*

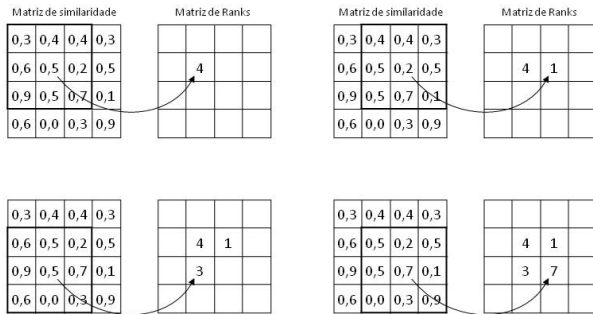
Apresenta baixa complexidade computacional, devido a simplicidade do algoritmo e baixa eficiência quando comparado a outros métodos mais sofisticados como mostrando em [3, 5].

Choi [3] apresenta um trabalho que usa *cosine*, a qual é exibida na equação 1, como medida similaridade e apresenta um esquema de ranking em seu algoritmo, o *C99*. Embora muitos dos melhores trabalho utilizarem 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 desproporcional no cálculo. 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 apresenta um esquema de ranking para contornar esse problema.

Cada valor na matriz de similaridade é substituído por seu ranking local. Onde ranking é o número de elementos vizinhos com similaridade menor, o qual é calculado com a equação 2. Um exemplo é mostrado na Figura 1 abaixo.



**Figure 1: Exemplo de construção de uma matriz de rank**

$$r(x, y) = \frac{\text{Numero de elementos com similaridade menor}}{\text{Numero de elementos examinados}} \quad (2)$$

Finalmente, na etapa de *clustering*, Choi utiliza um método baseado no algoritmo de maximização de Reynar [?] para identificar os limites entre os segmentos.

Semelhante a esse trabalho, outras abordagens foram propostas como ...

[1] faz uma adaptação do *TextTiling* ao contexto das conversas em reuniões com múltiplos participantes.

### 3. ADAPTAÇÃO ÀS ATAS DE REUNIÃO

Os algoritmos *TextTiling* e *C99* foram propostos para o inglês e sem domínio determinado, ou seja, a proposta inicial é trabalhar em qualquer texto nessa língua. A proposta desse trabalho é adaptá-los ao contexto das atas de reunião em português do Brasil. As subseções seguintes tratam das adaptações para esse nicho mais específico. A seção 4 mostra a análise dos algoritmos adaptados.

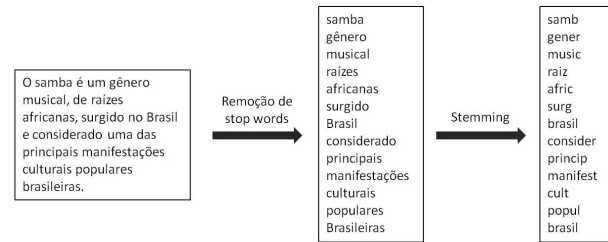
O vocabulário das reuniões, ainda que em tópicos diferentes, compartilham certo vocabulário pertencente ao ambiente onde as se deram as reuniões. Isso é um fator que diminui a o princípio da coesão léxica entre os segmentos. As atas de reunião costumam ter um estilo de escrita que deve ser levando em 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 expostas simples alterações para aumentar a eficiência dos algoritmos e encontra o melhor modelo para a tarefa de segmentar o texto das atas em tópicos.

#### 3.1 Préprocessamento

O texto a ser segmentado frequentemente é extraído de documentos em formatos como *pdf* ou de processadores de texto. Após a extração, esse pode passar por processos de transformação os quais serão apresentados a seguir.

A etapa de pre-processamento, em um documento contendo texto puro, acontece em dois passos principais. Primeiro elimina-se as palavras consideradas menos informativas, as quais são chamadas de *stop words*, para isso, utiliza-se uma lista contendo 438 palavras. Em seguida, remove-se os sufixos das palavras restantes, mantendo apenas o radical da palavra. A Figura 2 mostra a etapa de pré-processamento em uma sentença em português.



**Figure 2: Exemplo de pré-processamento**

Há ainda outros passos presentes nessa etapa como remoção de acentos, transformações de caixa, remoção de pontuação, os quais são relativamente simples e não requerem maiores detalhes.

##### 3.1.1 Remoção de ruídos

As atas frequentemente contém trechos que podem ser considerados pouco informativos e descartados durante o pré-processamento. Após a extração, cabeçalhos e roda-pés 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 extração, quanto a leitura do texto pelo usuário.

Também é comum o uso de numerais para marcação de páginas e linhas, da mesma forma, são pouco informativos e podem ser removidos.

Nesse trabalho, esses elementos são removidos por meio de heurísticas simples, uma vez que, o descarte não causa perda de informação e pode facilitar a identificação dos segmentos, pois melhora a coesão do texto. Outro benefício é manter os segmentos livres de textos que fogem do assunto.

#### 3.2 Identificação de candidatos

É preciso fornecer aos algoritmos os candidatos iniciais a limites de segmento. Aproveitando do estilo de escrita e baseando-se na pontuação do texto é possível indicar quebras de parágrafo, finais de sentenças ou mesmo palavras.

Ocorre que em atas de reunião é uma prática comum redigi-las de forma que o conteúdo discutido fica em parágrafo único, e quebras de parágrafo são usados para formatação de outros elementos como espaço para assinaturas. Indicar todo *token* como ponto candidato obriga a ajustá-lo de maneira a não segmentar uma ideia ou frase. Assim, nesse trabalho, os finais de sentença são considerado candidatos passíveis a limite entre segmentos.

Devido ao estilo de pontuação desses documentos, como encerrar sentenças usando um ";" e inserção de linhas extras, usa-se as regras abaixo para identificar os finais de sentenças.

---

**Algorithm 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 character for )}] "'
9 fim
```

---

## 4. AVALIAÇÃO

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, mantendo um conteúdo legível, 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 juizes, 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 tarefa secundária, analisar seu impacto na aplicação final.

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

Para quantificar a eficiência dos algoritmos, segue uma revisão das principais métricas aplicáveis.

O vocabulário das reuniões, ainda que em tópicos diferentes, compartilham certo vocabulário pertencente ao ambiente onde as se deram as reuniões. Isso é um fator que diminui a o princípio da coesão léxica entre os segmentos.

### 4.1 Medidas de Avaliação

As medidas de avaliação tradicionalmente utilizadas em *information retrieval* como precisão e revocação trazem alguns problemas na avaliação de segmentadores automáti-

cos. Conforme o algoritmo aponta mais segmentos no texto, tende a melhorar a revocação e ao mesmo tempo, reduzir a precisão, um problema que pode ser contornado usando F1 que faz uma combinação das duas levando em conta seus pesos, o que por outro lado é mais difícil de interpretar. Essas medidas falham ao não serem sensíveis a *near misses*, ou seja, quando um limite não coincide exatamente com o esperado, mas fica próximo [5].

A Figura 3 mostra um exemplo com duas segmentações hipotéticas e uma referência. 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.

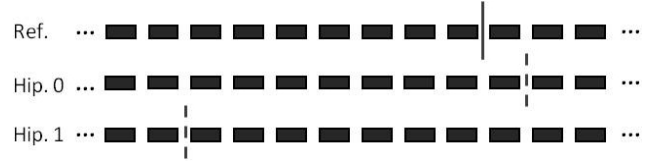


Figure 3: Exmplos de segmentação

#### 4.1.1 $P_k$

A fim de resolver o problema de *near misses*, Beeferman *et al.* [2] apresentam uma nova medida chama  $P_k$  que atribui valores parciais a *near misses*. 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 dividido pelo quantidade de segmentações analisadas. Esse valor serve como medida de dissimilaridade entre as segmentações e pode ser interpretada como a probabilidade de duas sentenças extraídas aleatoriamente pertencerem ao mesmo segmento.

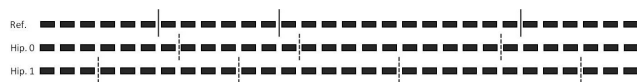
#### 4.1.2 *WindowDiff*

Pevzner [6] 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 e a quantidade de segmentos, entre outros problemas.

Como solução, propõem um novo método, o qual chamam de *WindowDiff* que traz duas diferenças principais: a dobra a 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 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  $P_k$ , consegue resolver seus principais problemas e mantém sua proposta inicial de sensibilidade a *near misses*, penalizando os menos que os falsos positivos puros.

## 5. ANÁLISE DOS RESULTADOS



**Figure 4: Exemplo de construção de uma matriz de rank**

## 6. CONCLUSÃO

## 7. REFERENCES

- [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. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210, 1999.
- [3] 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.
- [4] 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.
- [5] R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. pages 167–171, 2009. cited By 10.
- [6] 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.