

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* [3] 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 [2, 4].

Choi [2] 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 ???. Um exemplo é mostrado na Figura ?? abaixo.

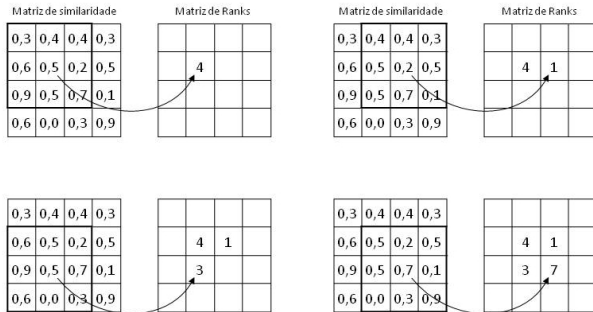


Figure 1: Exemplo de construção de uma matriz de rank

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

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

3.1 Préprocessamento

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 ?? mostra a análise dos algoritmos adaptados.

3.2 Identificação de sentenças

4. ANÁLISE DOS RESULTADOS

5. AVALIAÇÃO

Definir o que é um bom algoritmo de segmentação avaliação todos precisam de um gold text

1 - Concatenação 2 - Juízes concordam ou não 3 - Mediador na reunião 4 - Não avaliar o segmentador e sim o resultado da aplicação final.

De acordo com [?] há duas principais dificuldades na avaliação de segmentadores automáticos. A primeira é conseguir

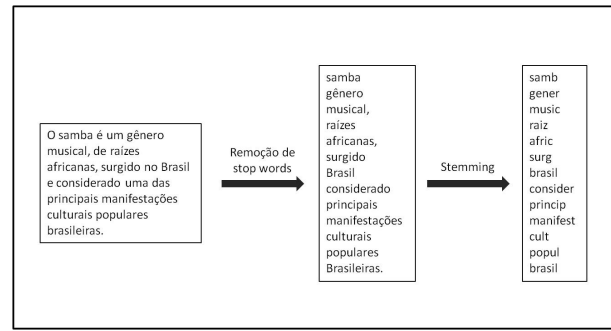


Figure 2: Exemplo de préprocessamento

A etapa de pré-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

3.1.1 Remoção de ruídos

passos menores 1 - heurística simples para remover cabeçalho e rodapé. 2 - remoção de numerais 3 - remoção de acentos, transformações de caixa, remoção de pontuação.

um referência confiável de texto segmentado, ou seja, uma segmentação ideal, já que juízes humanos costumam não concordar entre si, sobre onde os limites estão. 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 contornar essas dificuldades, algumas abordagens podem ser utilizadas. Algumas autores preferem detectar a segmentação em textos formados pela concatenação de documentos distintos, para que não haja diferenças subjetivas [?]. Há ainda outros que não avaliam o algoritmo diretamente, mas seu impacto na aplicação final[?, ?, ?]. Outras abordagens apenas atribuem um segmento cada quebra de parágrafo [?]

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.

5.1 Medidas de Avaliação

5.1.1 Pk

5.1.2 WindowDiff

No trabalho de [?], os autores apontam problemas na avaliação mais tradicional Pk, como a demasiada penalização dos falsos negativos e a desconsideração de *near misses*, quando um limite entre tópicos não casa exatamente com esperado mas fica próximo a ele.

A ideia é mover uma janela pelo texto e penalizar o algoritmo sempre que o número de limites (proposto pelo algoritmo) não coincidir com o número de limites (reais) para aquela janela de texto.

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] 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.
- [3] 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.
- [4] R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. pages 167–171, 2009. cited By 10.