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 noticias.

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.

Os algoritmos avaliados 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. Nesse artigo, os principais serão analisados na perspectiva de atas de reunião.

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 entre os principais trabalhos relacionados a segmentação textual estão o *TextTiling* e o *C99*

2.0.1 TextTiling

O algoritmo TextTiling, proposto por [**?**]

2.0.2 C99

3. ANÁLISE DOS RESULTADOS

4. AVALIAÇÃO

avaliação todos precisam de um gold text

1 - Concatenação 2 - Juizes 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 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 atribuiem um segmento cada quebra de parágrafo [**?**]

4.1 Medidas de Avaliação

4.1.1 Pk

4.1.2 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.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

5. TEXTTILINGBR

Adaptações nos algoritmos originais para o contexto das atas

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] A. H. Chaibi, M. Naili, and S. Sammoud. Topic segmentation for textual document written in arabic language. *Procedia Computer Science*, 35:437 – 446, 2014.

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

[5] 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.

[6] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. pages 53–65, 1996. cited By 50.