



Departamento de Ciência da Computação
Universidade de Brasília

Estudo segmentação de textos

Matheus Stauffer Viana de Oliveira
`matheusvostauffer@gmail.com`

19 de agosto de 2020

Segmentação de texto

Segmentação de discurso

Abordagens

- Segmentação de textos

- Segmentação de discurso

References

- ▶ Segmentação de textos é a tarefa de extrair segmentos coerentes de texto.
- ▶ Esses segmentos podem ser categorizados como palavra, sentença, frase ou qualquer unidade de informação dependendo da tarefa de análise envolvida [2].

Example

Publicações oficiais tais como o Diário Oficial do Distrito Federal (DODF) são fontes de informação sobre todos os atos oficiais do governo. | Embora esses documentos sejam ricos em conhecimento, || analisar esses textos manualmente por especialistas é uma tarefa complexa e inviável considerando o crescente volume de documentos, || resultado da frequente quantidade de publicações no veículo de comunicação do Governo do Distrito Federal (GDF). |

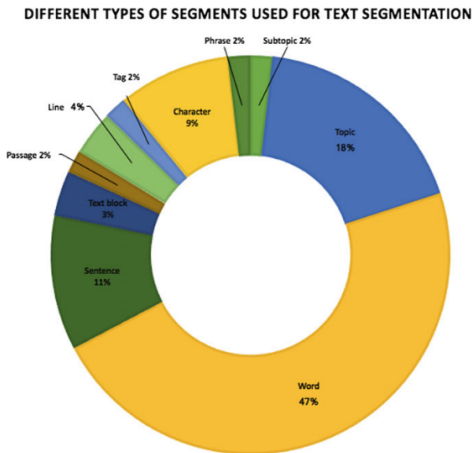


Figura: Tipos de segmentos [2]

- Segmentação: determinar as posições onde os (trechos/blocos/segmentos/tópicos/...) mudam em um stream de texto [1]

Example

Embora esses documentos sejam ricos em conhecimento, [M] analisar esses textos manualmente ... volume de documentos, [M] resultado da frequente ... do Distrito Federal (GDF).

[M] indica mudança de tópico

- Modelagem de tópicos: objetivo *semântico*.

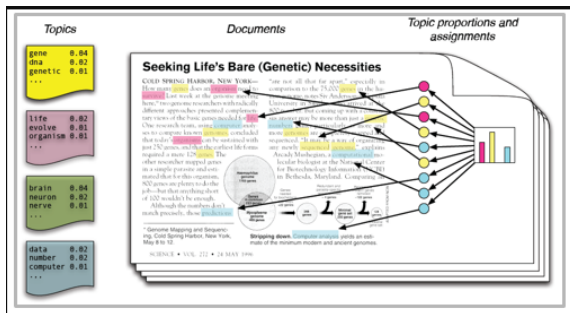


Figura: Exemplo de modelagem de tópicos ¹

¹<https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>

- Uma tarefa relacionada à segmentação de texto que quebra pedaços de texto em sub-elementos de uma sentença chamados Unidades Elementares de Discurso (UEDs ou EDUs, em inglês).
- EDUs são unidades minimais na análise de discurso de acordo com a Teoria da Estrutura Retórica (Mann e Thompson, 1988) [7].

Sentence 1:

Annuities are rarely a good idea at the age 35 || because of withdrawal restrictions

Sentence 2:

Wanted: || An investment || that's as simple and secure as a certificate of deposit || but offers a return || worth getting excited about.

Figure 2: Example discourse segmentations from the RST-DT dataset (Carlson et al., 2001). In the segmentations, the EDUs are separated by the || character.

Figura: Exemplo de EDUs [6]

- ▶ No alvorecer da pesquisa com segmentação de texto, predominavam abordagens não-supervisionadas, que primavam quantificar coesão léxica em segmentos pequenos de texto [6].
- ▶ Como é difícil definir e quantificar o que se entende por *coesão léxica*, em geral o termo era designado por/aproximado a ‘contar repetições de palavras’ [6].
- ▶ No entanto, esse tipo de abordagem sofre com dois principais problemas: é difícil de especializar para um domínio dado e não lida com questões relativas à multi-escala [6].
- ▶ Nesse sentido, a pesquisa mais recente tem se baseado em propostas supervisionadas, em particular abordagens de redes neurais [6].

- ▶ O problema de segmentação de discurso é, por natureza, mais propenso a encontrar resultados em propostas supervisionadas. Nesse sentido, um desafio é limitação de datasets focados na tarefa; em geral, as abordagens para o problema se pautam em anotadores e recursos externos para ajudar os modelos a generalizarem [6].
- ▶ Novas propostas recentes se pautam em modelos pré-treinados para obter representações de palavras ou sentenças, como o trabalho de (Li et. al. 2018) [8], que propõe dar a um modelo de sequence-to-sequence uma sequência de embeddings GloVe (Pennington et al., 2014) [9] como entrada para gerar as quebras de EDUs.

- ▶ O trabalho de (Lukasik et. al. 2020) [6] introduz novas arquiteturas de modelos baseadas em transformers e embeddings contextuais estilo-BERT para tarefas de segmentação de texto e discurso.
- ▶ Os resultados são promissores: os pesquisadores estabeleceram um novo estado-da-arte.



Regina Barzilay

Text Segmentation (2005)

<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-864-advanced-natural-language-processing-fall-2005/lecture-notes/lec13.pdf>



Irina Pak e Phoey Lee Teh

Text Segmentation Techniques - a critical review (2018)

https://www.researchgate.net/publication/321227216_Text_Segmentation_Techniques_A_Critical_Review





Hemant Misra, François Yvon, Olivier Cappé e Joemon Jose

Text segmentation: A topic modeling perspective (2019)

<https://hal.archives-ouvertes.fr/hal-01960703/document>

-  **Martin Riedl e Chris Biemann**
How Text Segmentation Algorithms Gain from Topic Models (2012)
<https://www.aclweb.org/anthology/N12-1064.pdf>
-  **Hanna M. Wallach**
Topic Modeling: Beyond Bag-of-Words (2006)
https://people.cs.umass.edu/~wallach/talks/beyond_bag-of-words.pdf
-  **Michal Lukasik, Boris Dadachev, Gonçalo Simões, Kishore Papineni**
Text Segmentation by Cross Segment Attention
<https://arxiv.org/pdf/2004.14535.pdf>
-  **William C Mann e Sandra A Thompson**
Rhetorical structure theory: Toward a functional theory of text organization (1988)
Journal for the Study of Discourse, 8(3):243–281.

-  **Jing Li, Aixin Sun, and Shafiq Joty.**
Segbot: A generic neural text segmentation model with pointer network (2018).
In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 4166–4172. International Joint Conferences on Artificial Intelligence Organization.
-  **Jeffrey Pennington, Richard Socher, and Christopher Manning.**
Glove: Global vectors for word representation (2014)
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.