Ciclo de Encontros: 4 x PLN

Sumarização automática - 2ª ed

Paula Figueira Cardoso
Faculdade de Computação - UFPA

Jackson Wilke da Cruz Souza
Inst. de Ciência, Tec. e Inovação - UFBA

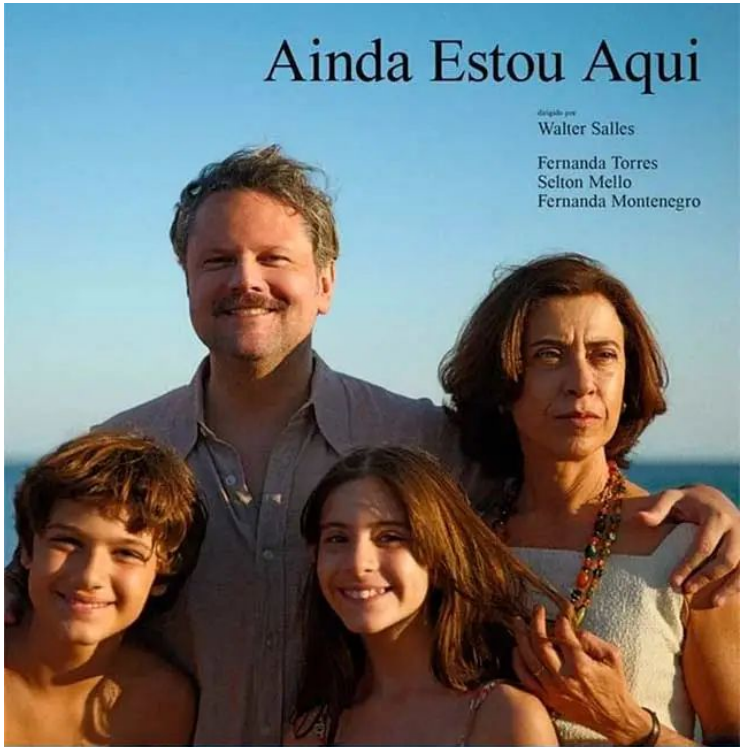
Crysttian Paixão
Inst. de Ciência, Tec. e Inovação - UFBA

12/02/2025

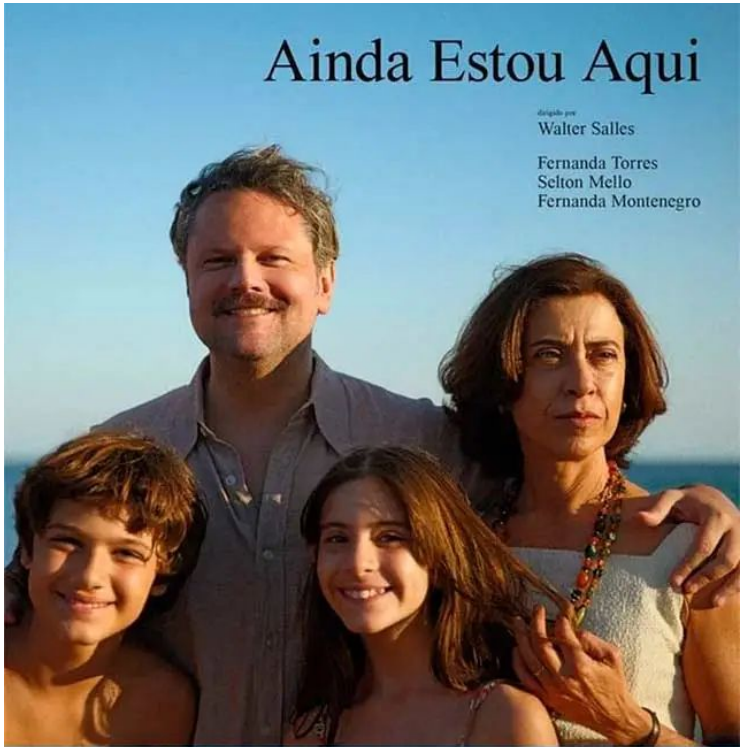
Sumário

Vamos do começo

O que é sumarizar?



O que é sumarizar?



Quantas horas tem o filme "Ainda Estou Aqui"?

Duração: **135 min.**

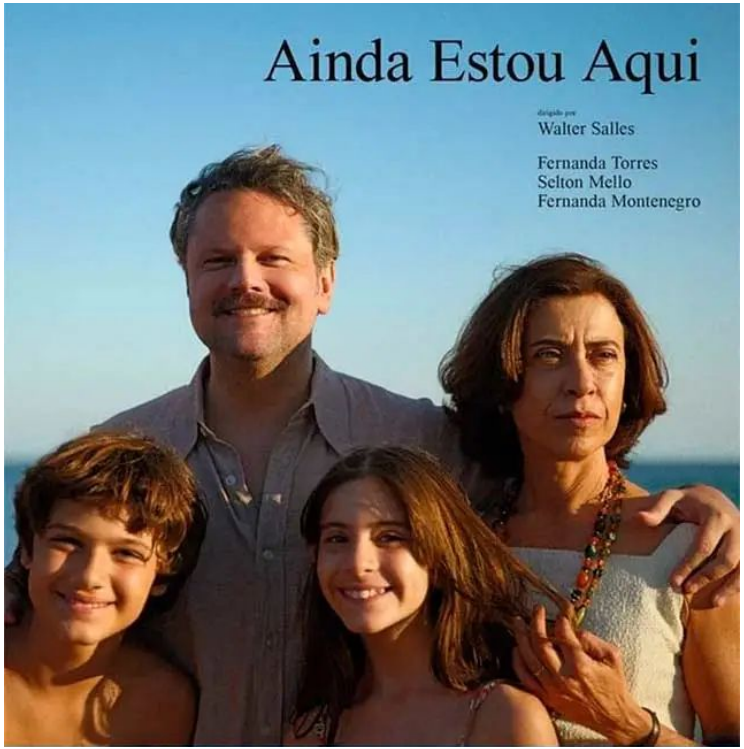
O que trata o filme "Ainda Estou Aqui"?

Em dezembro de 1970, Eunice Paiva (Fernanda Torres) e seu marido, o ex-deputado e engenheiro, Rubens Paiva (Selton Mello), moram com a família no Rio de Janeiro, enquanto tentam seguir com uma vida normal durante a Ditadura Militar.

Qual a moral do filme "Ainda Estou Aqui"?

O filme **fala sobre a importância de aceitar a dor para conseguir superá-la**. Possui uma história que muitos diriam que é clichê, mas se assistida com abertura e curiosidade, apresenta experiências comuns às pessoas que experimentaram traumas emocionais.

O que é sumarizar?



1) Tempo/Espaço

Quantas horas tem o filme "Ainda Estou Aqui"?

Duração: **135 min.**

2) Informatividade

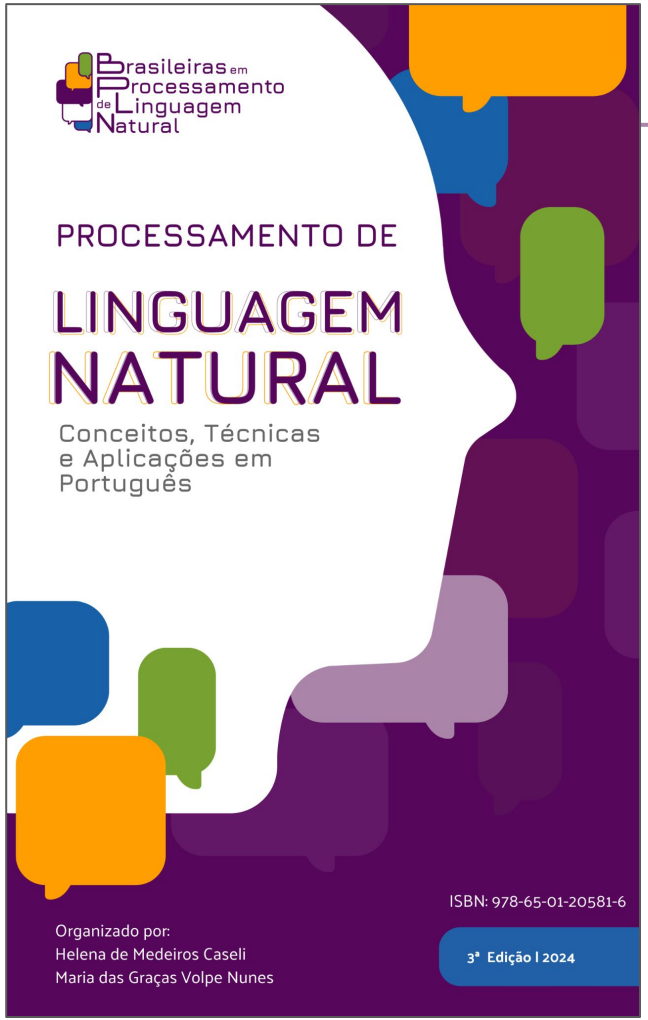
O que trata o filme "Ainda Estou Aqui"?

Em dezembro de 1970, Eunice Paiva (Fernanda Torres) e seu marido, o ex-deputado e engenheiro, Rubens Paiva (Selton Mello), moram com a família no Rio de Janeiro, enquanto tentam seguir com uma vida normal durante a Ditadura Militar.

3) Visão de mundo

Qual a moral do filme "Ainda Estou Aqui"?

O filme **fala sobre a importância de aceitar a dor para conseguir superá-la**. Possui uma história que muitos diriam que é clichê, mas se assistida com abertura e curiosidade, apresenta experiências comuns às pessoas que experimentaram traumas emocionais.



O que é sumarizar?

"não conseguiremos reproduzir exatamente o que aconteceu, mas faremos o máximo de esforço para transmitir uma mensagem o mais próximo do original, valorizando o ***conteúdo*** *mesmo em detrimento da* ***forma****."*

Souza, Cardoso e Paixão (2024, p.1)

Algumas perspectivas [Fiad, 2013; Rino e Pardo, 2003]

LINGUÍSTICA e COMPUTACIONAL

- Existir um texto-fonte
- Apresentar uma ideia ou tópico principais
- Identificar informações relacionadas entre si e conteúdo relevante
- Identificar um propósito comunicacional
- Manter informatividade e coerência
- Aplicar operações linguísticas que possam manter e/ou modificar o texto
- Produzir o gênero textual *resumo/sumário*

Sumarização automática

Processo genérico

Entendendo o processo de SA

Exemplo [CSTNews - Cardoso et al., 2011]

Texto-fonte 1	Texto-fonte 2
O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque. Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador –Cahe descartou pancreatite ou úlcera. "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.	BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. "Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. "Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación. No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

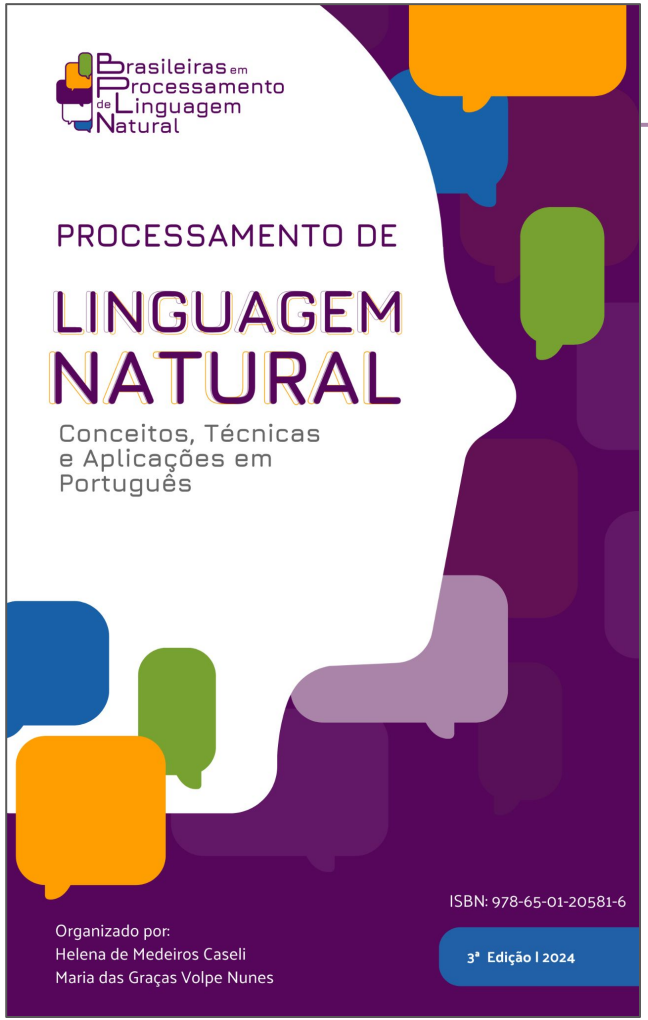
Exemplo [CSTNews - Cardoso et al., 2011]

Sumário humano multidocumento	Sumário automático multidocumento
BUENOS AIRES - Maradona voltou a ter problemas de saúde no fim de semana. <Texto-fonte 2> Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". <Texto-fonte 1> Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta. <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>	Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe. <Texto-fonte 2> "Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación". <Texto-fonte 1> Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas. <Texto-fonte 1>

Sumarização

Permeia o dia a dia das pessoas

- sinopse de novelas
- resumo de notícias
- resenhas de livros e filmes
- abstracts de artigos científicos
- passagens sobre páginas de internet



O que é importante?

*"A identificação de **conteúdo relevante** não é uma tarefa trivial. (...) Conforme se superavam alguns desafios, outros foram identificados e passaram a pautar muitos estudos, como a **sumarização multilíngue e abstrativa**"*

Souza, Cardoso e Paixão (2024, p.7)

Conceitos básicos

Diversas classificações:

- **Função**: informativo, indicativo ou crítico
- **Público-alvo**: genéricos ou específicos
- **Tipo**: extrativos ou abstrativos
- **Abordagem**: superficial, profunda ou híbrida
- **Fonte**: monodocumento ou multidocumento
- **Idioma**: monolíngue ou multilíngue

Taxa de compressão: quanto de informação será incluída no sumário

Avaliação

Qualidade dos sumários (DUC-2005)

Critérios difíceis de medir de forma automática?

- Gramaticalidade: ausência de erros de ortografia, pontuação e sintaxe
- Não redundância: ausência de informações repetidas
- Clareza referencial: clara identificação dos elementos do texto que se referem a outros componentes dentro do sumário
- Foco: uma sentença deve se relacionar com o restante do sumário
- Estrutura e coerência: organização do sumário

Avaliação

Informatividade

- Quanto da informação relevante dos **sumário de referência** é preservada no **sumário automático** - sobreposição de conteúdo

  ROUGE - Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004):

  Compara automaticamente a quantidade de n-gramas em comum entre um sumário automático e um ou mais sumários de referência.

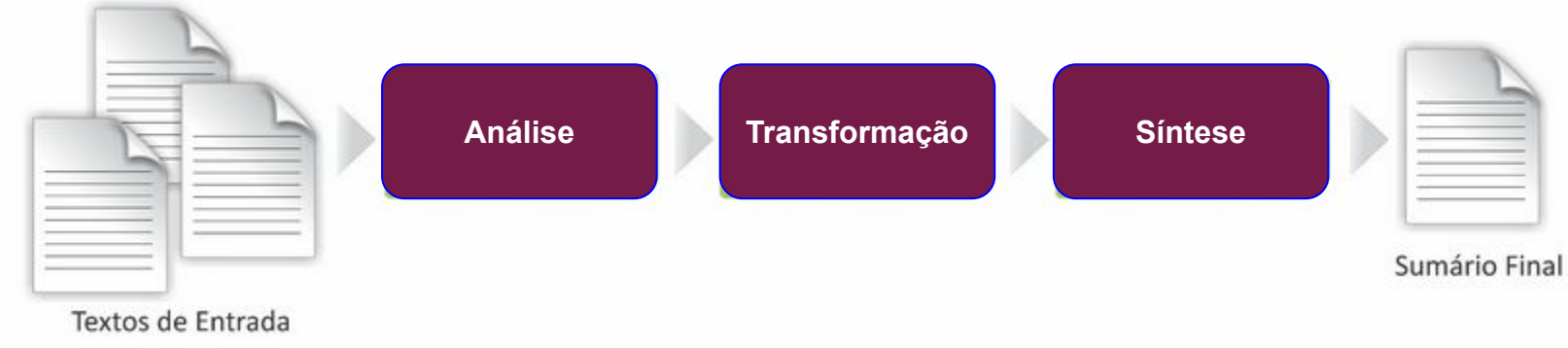
  Variações: ***Rouge-Ngram, Rouge-L, Rouge-W, Rouge-S***

Avaliação [Fabbri et al., 2021; Zhu; Bhat, 2020]

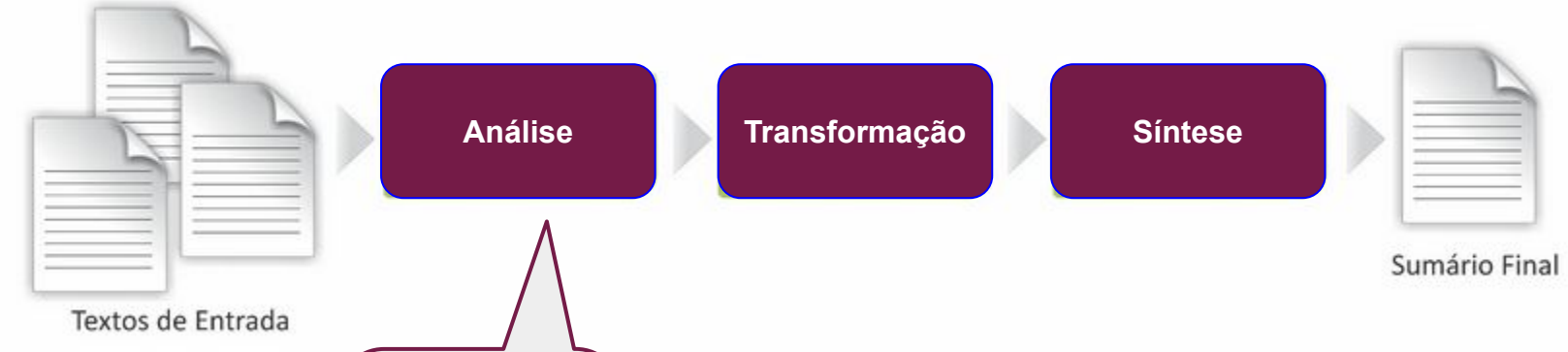
Questionamentos

- E quando não temos o sumário de referência?
- Produção de sumários é uma tarefa intelectual e que sofre influência da familiaridade com o assunto, atitude e disposição do produtor.
- Métricas baseadas apenas na sobreposição de vocabulário não são eficazes para o cenário atual (sumários abstrativos)
- Diversos conjuntos de dados de SA com pouca discussão sobre sua qualidade, indicando ser necessário buscar outras formas de avaliação

Etapas da SA


Análise
Transformação
Síntese
Textos de Entrada
Sumário Final

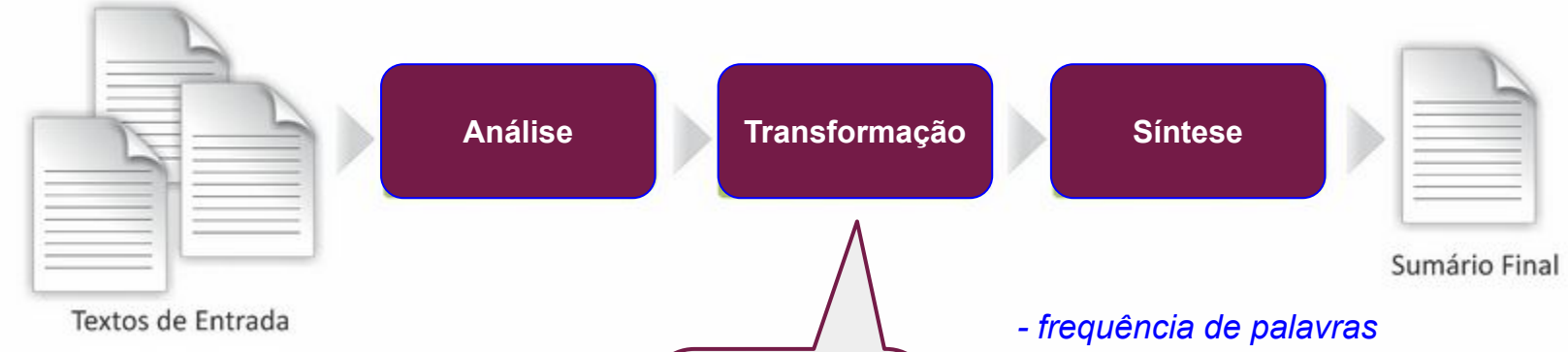
Etapas da SA



textos-fonte são analisados e transformados em uma forma mais estruturada

- *pré-processamento*
- *análise discursiva*
- *análise semântica*
- *etc*

Etapas da SA



Com algum critério de *ranking*, as partes mais relevantes são selecionadas

- frequência de palavras
- posição das palavras/sentenças no texto
- frases indicativas
- relacionamentos discursivos
- relacionamentos semânticos
- algoritmos de AM, modelos de língua, uma variedade de possibilidades

Etapas da SA


Textos de Entrada
Análise
Transformação
Síntese
Sumário Final
gera-se o sumário em LN. Ajustes: tratamento de correferência, ordenação de sentenças, etc

Uma infinidade de possibilidades [Zhang et al., 2024]

Modelos estatísticos de língua

transformação do texto em características estatísticas

Uma infinidade de possibilidades [Zhang et al., 2024]

Modelos estatísticos de língua

transformação do texto em características estatísticas

simples e eficientes

extração de relações gramaticais e contextuais

Uma infinidade de possibilidades [Zhang et al., 2024]

Modelos de Word Embedding

conversão do texto em vetores contínuos, permitindo a captura dos significados

Modelos estatísticos de língua

transformação do texto em características estatísticas

Uma infinidade de possibilidades [Zhang et al., 2024]

representação de palavras e aprendizado de padrões +

ambiguidades contextuais, captura semântica profunda −

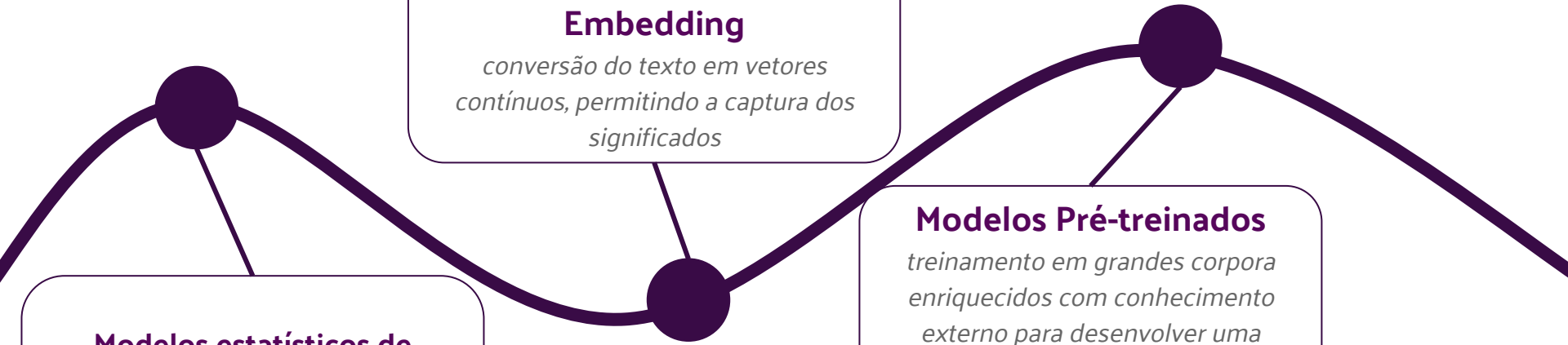
Modelos de Word Embedding

conversão do texto em vetores contínuos, permitindo a captura dos significados

Modelos estatísticos de língua

transformação do texto em características estatísticas

Uma infinidade de possibilidades [Zhang et al., 2024]



Modelos de Word Embedding

conversão do texto em vetores contínuos, permitindo a captura dos significados

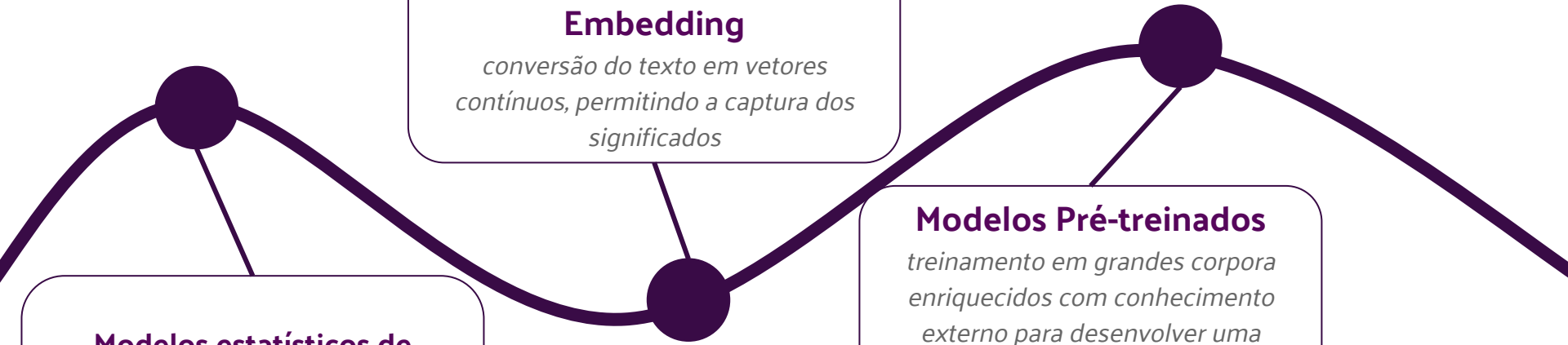
Modelos estatísticos de língua

transformação do texto em características estatísticas

Modelos Pré-treinados

treinamento em grandes corpora enriquecidos com conhecimento externo para desenvolver uma representação "textual universal"

Uma infinidade de possibilidades [Zhang et al., 2024]



Modelos de Word Embedding

conversão do texto em vetores contínuos, permitindo a captura dos significados

Modelos estatísticos de língua

transformação do texto em características estatísticas

Modelos Pré-treinados

treinamento em grandes corpora enriquecidos com conhecimento externo para desenvolver uma representação "textual universal"

captura de significados profundos

requerem adaptações à SA

cursos elevados para trainamento *

O que temos para língua portuguesa

Diversos *corpora* e ferramentas

Corpus

- Temário (Pardo e Rino, 2003)
- CSTNews - multidocumento (Cardoso et al, 2011)
    - diversas camadas de anotação linguística
- OpiSums-PT: sumários opinativos - multidocumento (López et al., 2015)

O que temos para língua portuguesa

Diversos corpora e ferramentas

Ferramentas

- GistSumm (Pardo, 2002)
    - 1o. sumarizador automático - abordagem estatística e monodocumento
- CSTTool (Aleixo; Pardo, 2008)
    - para análise semiautomática para SA multidocumento com base na CST (Cross-document Theory)

Além de várias pesquisas que utilizam modelos discursivos (Cap. 11 do livro), abordagens híbridas.

O que temos para língua portuguesa

Mudança de paradigmas

- Utilização de LLMs [Barros, 2022; Paiola, 2022]
- Diferentes domínios, como o jurídico [Feltrin et al., 2023] e código de programação [Pontes et al., 2022]

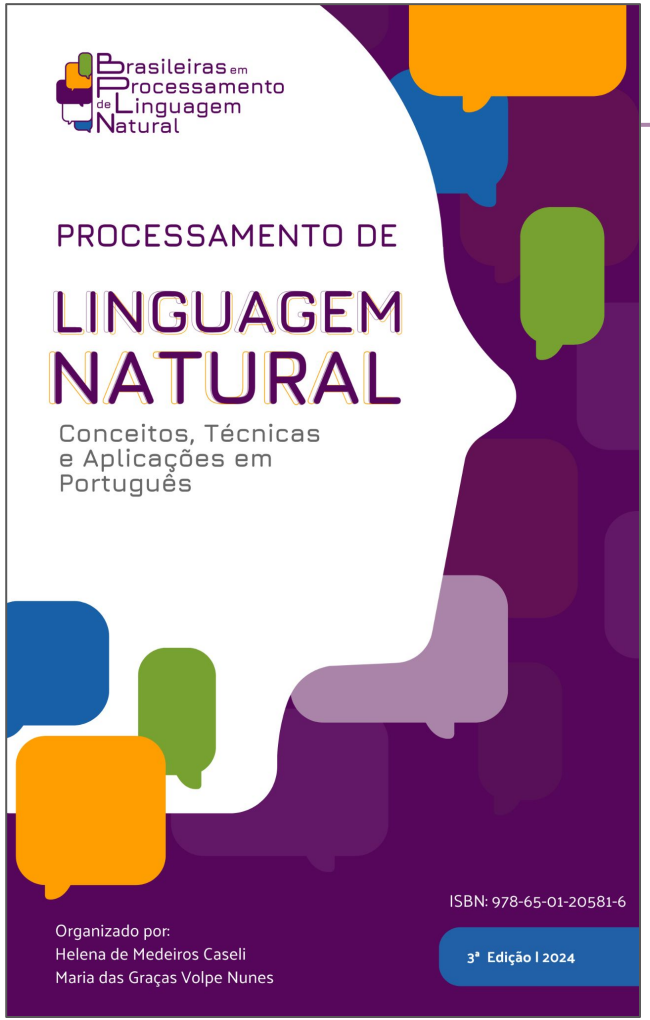
Considerações finais

Considerações finais

- A abordagem **Transformers** deve ser o limiar entre o antes e o depois em SA (*e como quase tudo em PLN!*)

- Algumas questões ainda continuam em aberto
  - Quais as **motivações** de continuar pesquisando SA?
  - Quais as considerações sobre **Gênero e tipo textuais** devem ser levadas em conta?
  - Quão importante é o **usuário** do sistema durante o processo de SA?

- Alguns **desafios** foram superados, mas há outros ainda



Ainda há o que ser feito!

"alguns [desafios] ainda persistem, como a seleção de conteúdo relevante e personalizado, avaliação do sumário automático sem a necessidade de compará-lo com o sumário produzido por humanos, geração de abstract semelhante a um sumário produzido por humano) e quais outros poderão ser superados ***na*** *e* ***com a sumarização****.*

Souza, Cardoso e Paixão (2024, p.19)

Agradecimentos

- Parceria na elaboração deste capítulo
- Brasileiras em PLN

Referências

ABBRI, A. R. et al. SummEval: Re-evaluating Summarization Evaluation. **Transactions of the Association for Computational Linguistics**, v. 9, p. 391–409, 2021.

FIAD, R. S. Reescrita, dialogismo e etnografia. **Linguagem em (Dis) curso**, v. 13, p. 463–480, 2013.

RINO, L. H. M.; PARDO, T. A. S. A Sumarização Automática de textos: principais características e metodologias. Anais do XXIII Congresso da Sociedade Brasileira de Computação. **Anais**... 2003.

CARDOSO, P. C. F. et al. CSTNews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. Proceedings of the 3rd RST Brazilian Meeting. **Anais**...2011.

LIN, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. **Anais**...Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Disponível em: <https://aclanthology.org/W04-1013>

FABBRI, Alexander R. et al. Summeval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistic**s, v. 9, p. 391-409, 2021.

ZHU, Wanzheng; BHAT, Suma. GRUEN for evaluating linguistic quality of generated text. **arXiv preprint arXiv:2010.02498**, 2020.

JIN, Hanlei et al. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. **arXiv preprint arXiv:2403.02901**, 2024.