



Classificação de publicações de autores ambíguos em repositórios bibliográficos digitais utilizando o modelo de PLN SciBERT

Lucas Damasceno
Natan Rodrigues

Brasília, 06 de Abril de 2024

Agenda

- Introdução
- Fundamentação Teórica
 - Desambiguação de nomes de autores ou *Author Name Disambiguation* - AND
 - Processamento de Linguagem Natural (PLN) - BERT
- Experimentos Iniciais
- Considerações Finais

Repositórios Bibliográficos Digitais

Disponibilizam de forma centralizada informações de citações, trabalhos científicos, autores e redes sociais acadêmicas.

- Exemplos: *DBLP*, *ArnetMiner*, *CiteSeerX*, *PubMED*

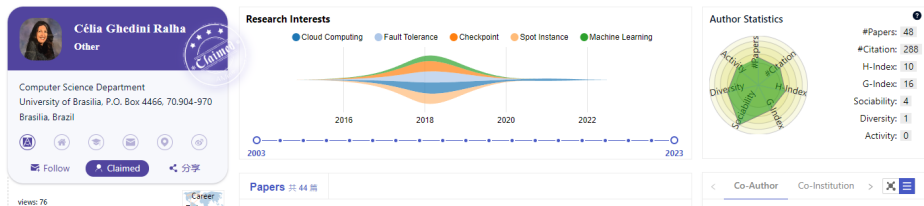


Figura: Página da pesquisadora Célia Ghedini Ralha no *ArnetMiner* (consulta realizada em 03-04-2024).

Repositórios Bibliográficos Digitais

- Diferentes autores podem compartilhar a mesma referência bibliográfica.
- Erros tipográficos ou abreviações.
- Afeta a integridade dos dados de um repositório digital.
- Um pesquisador pode ter várias entradas em um determinado repositório.

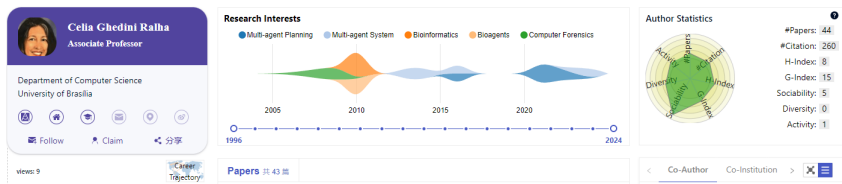


Figura: Outra página da pesquisadora Celia Ghedini Ralha no *ArnetMiner* (consulta realizada em 03-04-2024).

Ambiguidade de Nomes de Autores

Aspectos no tratamento de ambiguidade de nomes de autores:

- Vários autores com mesmo nome e/ou um autor com vários formatos de nome.
- Problema da ambiguidade de nomes de autores.
- Desambiguação de Nomes de Autores.
- Existem diversos trabalhos na literatura que visam a resolução do problema de AND.

Ambiguidade de Nomes de Autores

Segundo Ferreira et al. (2020)¹ são questões que devem ser consideradas para o desenvolvimento de soluções para o problema de AND:

- Poucos dados nas citações bibliográficas
- Eficiência
- Praticidade e custo
- Escalabilidade

¹Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2020). Automatic Disambiguation of Author Names in Bibliographic Repositories. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 12(1), 1–146. <https://doi.org/10.2200/S01011ED1V01Y202005ICR070>

Ambiguidade de Nomes de Autores - Projetos para resolução

- Algoritmo de aprendizado não-supervisionado multinível com clusterização hierárquica de grafos

Zhang, S., Xinhua, E., & Pan, T. (2019). A multi-level author name disambiguation algorithm. IEEE Access, 7, 104250–104257. <https://doi.org/10.1109/ACCESS.2019.2931592>

- Método de Resolução Progressiva de Entidades

Backes, T., & Dietze, S. (2022). Lattice-Based Progressive Author Disambiguation. Information Systems, 109, 102056. <https://doi.org/10.1016/j.is.2022.102056>

- Utilização de redes neurais para AND

Boukhers, Z., & Asundi, N. B. (2022). Whois? Deep Author Name Disambiguation Using Bibliographic Data. 201–215

- Abordagem multi-estratégica (comparação de *strings*, tratamento de ligações de nomes e análise de similaridade de redes sociais)

Rodrigues, N. S., Costa, A. R., Lemos, L. C., & Ralha, C. G. (2021). Multi-strategic Approach for Author Name Disambiguation in Bibliography Repositories. Em Information Management and Big Data. SIMBig 2020. Communications in Computer and Springer. https://doi.org/10.1007/978-3-030-76228-5_5

Problema

Considerando que os métodos de AND propostos na literatura podem enfrentar os desafios mencionados, esta pesquisa aborda a utilização de PLN para abordar as questões elencadas.

Justificativa:

- PLN pode lidar com a complexidade semântica dos textos bibliográficos.

Objetivos

O objetivo principal deste trabalho é o desenvolvimento de uma solução que aplique uma classificação utilizando PLN para o problema de AND em repositórios bibliográficos digitais.

Processamento de Linguagem Natural (PLN)

PLN é uma área da Inteligência Artificial que visa o desenvolvimento de métodos e algoritmos para permitir a compreensão, interpretação e geração de linguagem humana por computadores (Daniel, James H et al., 2007).



Técnicas de PLN

- Word2Vec - utiliza vetores para representar palavras em alta dimensionalidade capturando relações semânticas e sintáticas.
- Redes Neurais Convolucionais (RNC) - particularmente eficazes em tarefas como a classificação de textos e a análise de sentimentos.
- *Bidirectional Encoder Representations from Transformers* (BERT) - modelo de PLN baseado na arquitetura *transformer*, capaz de compreender o contexto das palavras em uma frase de forma bidirecional, considerando o contexto anterior e posterior.

BERT

- Permite uma boa compreensão das relações e significados das palavras em determinado texto.
- Considera tanto os termos que vêm antes quanto os que vêm depois de cada vocábulo (bidirecional).
- Utilização de *transfer learning* para adaptação a diferentes domínios.
- Pré-treinamento em grandes conjuntos de textos não rotulados.
- Ajustes para tarefas específicas com camadas de classificação.

BERT

Como o BERT funciona?

- Segmentação de palavras em partes menores e a inclusão dos *tokens* especiais [CLS] e [SEP] para formatar os dados de entrada.
- [CLS] é usado no início de cada entrada, representando o texto agregado para tarefas de classificação em nível de frase.
- [SEP] separa frases distintas dentro da mesma entrada, permitindo que o BERT lide com várias tarefas.
- Utilização de três tipos de *embedding*:
 - *Token*,
 - *Segment*, e
 - *Position*.

BERT

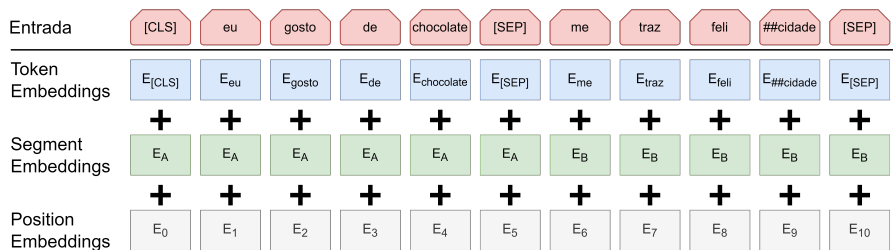


Figura: Pré-processamento textual do modelo BERT e criação de *embeddings* na camada inicial (adaptado de Devlin et al., 2018).

BERT

O pré-treinamento do BERT é composto por duas tarefas:

- Modelagem de Linguagem Mascarada.
- Previsão da Próxima Fase.

Input = [CLS] o homem foi à [MASK] loja [SEP]
ele comprou um litro de [MASK] leite [SEP]

Label = IsNext

Input = [CLS] o homem [MASK] foi à loja [SEP]
os gatos têm [MASK] sete vidas [SEP]

Label = NotNext

Figura: Tarefas de pré-treinamento do BERT (adaptado de Devlin et al., 2018; Eler, 2022).

Modelos BERT

<i>Modelo</i>	<i>Características</i>
BERT	Modelo original do Google
RoBERTa	BERT com treinamento mais longo e modelos maiores
DistilBERT	Versão compacta do BERT para inferência mais rápida
ALBERT	Versão leve do BERT com tamanho de modelo e etapas de treinamento reduzidas
SciBERT	Modelo BERT pré-treinado para tarefas científicas

Segundo Beltagy et al. (2019)², o SciBERT tem:

- Treinamento com 1.14M papers e 3.1B tokens do semanticscholar.org.
- Bom desempenho em tarefas de PLN no domínio científico.

²Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained Language Model for Scientific Text. [EMNLP](#)

Setup Experimental

Classificação vs Clusterização

Em tempo real verificamos que esse problema de AND deve ser tratado como um problema de Clusterização, pois dados reais e incrementais são não-rotulados. No entanto, utilizamos bases rotuladas e bem verificadas para teste da classificação utilizando o SciBERT.

- *ArnetMiner* - 2.
- Utilizar Título e *Abstract* (Resumo).
- 100 nomes de autores ambíguos.
- 208827 documentos.
- 39655 autores únicos.
- Autores, coautores, palavras-chave, resumos, títulos e instituições.
- Disponível em <https://github.com/neozhangthe1/disambiguation/?tab=readme-ov-file>.

Dataset - Cabeçalho

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

Quantidade de classe de autores: 16

Quantidade total de documentos: 198

	publication_title	author_id	abstract	ENCODE_AUTHOR
0	Induced life cycle transition from holocycly t...	5b5433f3e1cd8e4e1516badf	The Russian wheat aphid (RWA), <i>Diuraphis noxia</i> ...	0
1	Impact of alfalfa/cotton intercropping and man...	5b5433f3e1cd8e4e1516badf	A short-term study was carried out to evaluate...	0
2	A check-list of the Chinese Megalopodinae (Col...	5b5433e5e1cd8e4e15f7474c	Two genera and 33 taxa of Megalopodinae are re...	1
3	Key to the species of the genus <i>Aristochroa</i> Ts...	5b5433e5e1cd8e4e15f7474c	A key to all 14 species of <i>Aristochroa</i> Tschits...	1
4	A review of genus <i>Onycholabis</i> bates (Coleopter...	5b5433e5e1cd8e4e15f7474c	Species in the genus <i>Onycholabis</i> Bates are bri...	1

Figura: Cabeçalho do Conjunto de Dados

Dataset - WholsWho

- WholsWho é dos maiores datasets rotulados manualmente do mundo com mais de 1.000.000 de artigos construídos usando um processo de anotação interativo.

The screenshot shows the WholsWho website. The header has the WholsWho logo and a 'Home' link. The main banner features the text 'WholsWho' and 'Web-Scale Academic Name Disambiguation: the WholsWho Benchmark, Leaderboard, and Toolkit' over a 3D cube graphic. Below the banner, there are two buttons: 'From-scratch Name Disambiguation' and 'Real-time Name Disambiguation'. On the left, a section titled 'What is WholsWho?' provides a description of the dataset and toolkit. On the right, a 'Leaderboard for From-scratch Name Disambiguation' is displayed as a table.

What is WholsWho?

WholsWho owning, a world's largest manually-labeled benchmark with over 1,000,000 papers built using an interactive annotation process,

A regular leaderboard with comprehensive tasks, i.e., From-scratch Name Disambiguation, Real-time Name Disambiguation, and Incorrect Assignment Detection. The historical contests of WholsWho have already attracted more than 3,000 researchers.

An easy-to-use toolkit encapsulating the entire pipeline as well as the most powerful features and baseline models for tackling the tasks.

Please refer to the WholsWho paper for more details.

Chen et al. 2023

Leaderboard for From-scratch Name Disambiguation

Rank	Method	Organization	References	Metric (P-F1)
1	BOND	-	-	0.89719
2	SND-all	-	-	0.89216
3	ECNU_AIDA	-	-	0.89140
4	Complex808	-	-	0.88594

Execução

- Tratamento dos Dados de Entrada
 - Iteração sobre arquivos .json para extração de informações relevantes.
- BERT
 - *SciBERT*
 - Hiperparâmetros utilizados:
 - Tamanho do vetor de *embedding* do SciBERT: 256
 - Tamanho do lote: 64
 - Taxa de aprendizado: 2×10^{-5}
 - Número de épocas de treinamento: 20

Execução - Escolha dos Hiperparâmetros

- Andrade et al. (2020)³ aborda o sucesso das representações contextuais, como as baseadas em BERT, na melhoria da eficácia de algumas tarefas de PLN.







Claudio M.V. de Andrade et al.

Information Processing and Management 60 (2023) 103336

Table 11
Tuning Parameters of the deep learning methods.

Methods	Parameters
BERT	initial learning rate: 2×10^{-5} ; batch_size: 64; max_epochs: 5; max_len: 256
XLNet	initial learning rate: 5×10^{-5} ; batch_size: 32; max_epochs: 10; patience: 5; max_len: 150
RoBERTa	initial learning rate: 5×10^{-5} ; batch_size: 32; max_epochs: 10; patience: 5; max_len: 150
DistilBERT	initial learning rate: 5×10^{-5} ; batch_size: 32; max_epochs: 10; patience: 5; max_len: 256
BART	initial learning rate: 5×10^{-5} ; batch_size: 16; max_epochs: 10; patience: 5; max_len: 256
CNN1	batch size: 32; max epochs: 10; patience: 5; max length: 256; vector length: 300
CNN2	batch size: 64; max epochs: 100; patience: 30; max length: 256; vector length: 300

Figura: Hiperparâmetros utilizados neste estudo.

³de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., & Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations—or “The classifier does not matter when the (text) representation is so good!”. *Information Processing & Management*, 60(4), 103336.      

Execução - Código e Hardware para execução

- Disponível em https://github.com/natansr/int_mineracao_dados_seminario_final.git
- GPU - A100 - Google Colab

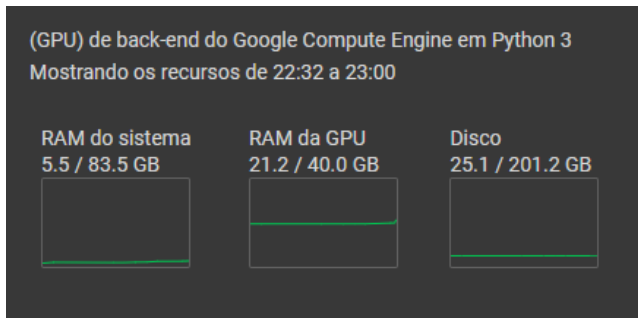


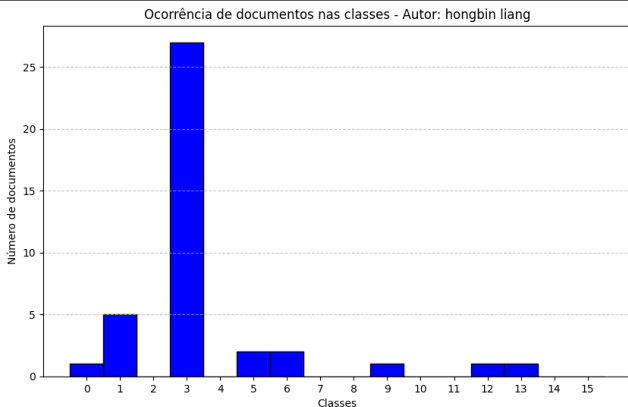
Figura: Hardware para experimento.

Métricas de Avaliação

- **Precisão (Precision):** Mede a proporção de verdadeiros positivos em relação ao número total de previsões positivas feitas pelo modelo.
- **Revocação (Recall):** Mede a proporção de verdadeiros positivos em relação ao número total de positivos reais na amostra.
- **F1-Score:** Média harmônica entre precisão e revocação, balanceando a importância de ambas as métricas.
- **Acurácia (Accuracy):** Mede a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo.

Teste com autor - Hongbin Liang

- Classes: 16
- Documentos totais: 198
- Treinamento (80%): 158
- Teste (20%): 40



Teste com autor - Hongbin Liang

Relatórios de classificação:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.83	1.00	0.91	5
3	0.96	1.00	0.98	27
5	0.50	1.00	0.67	2
6	1.00	1.00	1.00	2
9	0.00	0.00	0.00	1
12	0.00	0.00	0.00	1
13	0.00	0.00	0.00	1
accuracy			0.90	40
macro avg	0.41	0.50	0.44	40
weighted avg	0.83	0.90	0.86	40

Apresentação dos Resultados

Autor	Documentos	Classes	Acurácia (%)
Hongbin Liang	198	16	90.00
Wen Chang Chen	312	13	95.00
Yongsheng Zhao	299	51	63.33
Jing Luo	682	174	58.09
Toda base	208827	39655	X?

Tabela: Resumo dos Autores e Resultados

Considerações finais

- Resultados preliminares demonstram a complexidade do problema de AND e a necessidade de uma maior investigação.
- O *F-measure* médio é competitivo para classificação de um nome em específico de autor. Também, observamos uma relação entre a quantidade de classes e a acurácia.
- Exploração de métricas específicas ou empíricas para avaliar a qualidade das representações semânticas e classificação obtidas pelo modelo SciBERT.
- Utilizar outros modelos de PLN para comparação de complexidade de processamento.
- Buscar outros modelos para resolução do problema em um cenário não-supervisionado.

Obrigada pela participação e atenção!



Lucas Damasceno
Natan Rodrigues

lucasadmfv@gmail.com, natan5souza@gmail.com