

From Documents to Entities: A journey through Natural Language Processing tasks and domains

Pedro Henrique Luz de Araujo

Departamento de Ciência da Computação, Universidade de Brasília, Brasília – DF, Brazil

pedro.luz@aluno.unb.br

Banca:

Teófilo E. de Campos (Orientador - UnB)

Alexandre Rademaker (IBM Research)

Thiago de Paulo Faleiros (UnB)

Luís Paulo F. Garcia (Suplente - UnB)

7 de agosto de 2020

Sumário

- 1 Introdução
- 2 Classificação de texto
- 3 Modelagem de Tópicos
- 4 Reconhecimento de Entidade Nomeada
- 5 Ligação de Entidade
 - Introdução
 - Trabalhos relacionados
 - Plano de trabalho

Introdução

- Dados textuais estão em constante produção.
 - ▶ Posts de redes sociais, livros, notícias, publicações oficiais, processos judiciais.
- Dados precisam ser estruturados para gerar conhecimento.
 - ▶ Aprendizado de Máquina e Processamento de Linguagem Natural permitem análise de texto em uma escala inatingível por humanos.

Aprendizado Profundo para Textos

- O uso de redes neurais profundas representou avanços em uma gama de tarefas de processamento de texto.
 - ▶ Análise de sentimento;
 - ▶ Tradução por máquina;
 - ▶ Inferência em textos naturais.
- Problema: Generalização para diferentes domínios.
 - ▶ Transferência de aprendizado ajuda; mas dados rotulados de domínios específicos ainda são necessários—ajuste fino e avaliação.

Exemplos

DODF

O GOVERNADOR DO DISTRITO FEDERAL, no uso das atribuições que lhe confere o artigo 100, incisos XXVI e XXVII, da Lei Orgânica do Distrito Federal, resolve [...]

Acórdão STF

HABEAS CORPUS 110.260 SÃO PAULO RELATOR : MIN. LUIZ FUX
PACTE.(S) :LAERCIO BRAZ PEREIRA SALES IMPTE.(S)
:DEFENSORIA PÚBLICA DA UNIÃO PROC.(A/S)(ES) :DEFENSOR
PÚBLICO-GERAL FEDERAL COATOR(A/S)(ES) :SUPERIOR
TRIBUNAL DE JUSTIÇA

Objetivos I

- Explorar tarefas de NLP em diferentes domínios.
- Abordagem documentos → entidades.
- Objetivos específicos:
 - 1 propor conjuntos de dados para classificação de textos dos domínios jurídico e de textos oficiais e comparar modelos;
 - 2 treinar modelos de tópicos em textos jurídicos, analisar a semântica dos tópicos obtidos;
 - 3 propor conjunto de dado para reconhecimento de entidades nomeadas (NER) com entidades do domínio jurídico e treinar um modelo com os dados;
 - 4 propor um sistema de ligação de entidade (EL) para domínios com poucos recursos.

Objetivos II

- Items 1 a 3 já desenvolvidos. Publicações:
 - ▶ Luz de Araujo, P. H. et al. VICTOR: a dataset for Brazilian legal documents classification. [1]
 - ▶ Luz de Araujo, P. H. et al. Inferring the source of official texts: can SVM beat ULMFiT? [2]
 - ▶ Luz de Araujo, P. H. et al. LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. [3]
- Planos para o item 4.

Classificação de texto

VICTOR: a dataset for Brazilian legal documents classification.

- Conjunto de dados com mais de 40.000 Recursos Extraordinários.
- Rótulos a nível de página (tipo de documento) e a nível de processo (tema de repercussão geral).
- Comparação de modelos BoWs, CNNs, LSTMs. Uso de CRF para explorar natureza sequencial das páginas de um processo.

Inferring the source of official texts: can SVM beat ULMFiT?

- Conjunto de dados com textos do DODF rotulados (órgão de origem) e não rotulados.
- Treino e comparação de um método de transferência de aprendizado estado-da-arte (ULMFiT [4]) com modelos BoW.
- BoW + SVM competitivo quando comparado com o ULMFiT.

Modelagem de Tópicos

Topic modelling Brazilian Supreme Court lawsuits

- Usa LDA para modelar Recursos Extraordinários.
- Treino e análise de modelos com 10 e 30 tópicos.
- Vetorização de textos usando tópicos (30, 100, 300 e 1000 dimensões) para classificação de tema.

Reconhecimento de Entidade Nomeada

LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text

- Conjunto de 70 documentos legais (acórdãos, leis e portarias) com anotação de entidades, totalizando 318.073 tokens; destas, 44.513 pertencentes a entidades.
- Entidades de 4 tipos genéricos (pessoa, tempo, organização e local) e 2 específicos ao domínio (legislação e jurisprudência).
- Treino de modelo char-biLSTM-CRF usando os dados.

Ligação de Entidade

- Um passo além de NER: ligar as menções extraídas a entidades específicas de uma base de conhecimento.

Lula?

A **lula** tem oito braços, para a captura de alimento, e dois tentáculos, com função na reprodução.

Motivação II

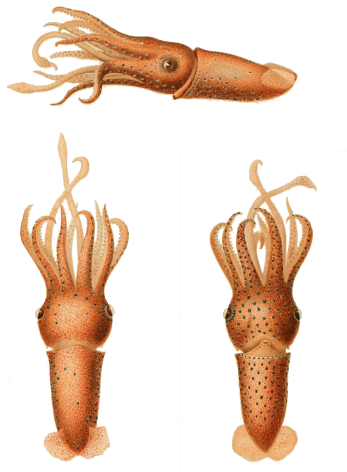


Figura: Essa lula...



Figura: ...ou esse Lula?

Três passos:

- 1 Detecção de menção: extração de potenciais menções a entidades—idêntico a NER quando restringe-se menções a entidades nomeadas.
 - 2 Geração de candidatos: o sistema seleciona um conjunto de possíveis candidatos para cada menção.
 - 3 Desambiguação de entidades: o sistema seleciona a entidade mais provável para cada menção.
- Ligação ponta-a-ponta versus apenas desambiguação.

O desafio

- Potencialmente milhões de entidades.
- Diversidade de menções.
 - ▶ Big blue vs IBM.
- Ambiguidade de menções.
 - ▶ Paris?

A solução (?)

- Utilizar recursos adicionais:
 - ▶ Estatísticas de frequência ligação;
 - ▶ Informação estruturada;
 - ▶ Tabelas de aliases.
- Conjuntos de dados massivos:
 - ▶ Wikipedia.

O problema

- Cenários com poucos recursos:
 - ▶ Sem quantidade massiva de dados anotados no domínio alvo;
 - ▶ Sem estatísticas de frequência;
 - ▶ Sem descrições textuais canônicas de entidade;
 - ▶ Sem dados estruturados sobre entidades.
- Domínios específicos: médico e legal.

Trabalhos relacionados I

- 12 publicações de 2016 a 2020, analisando:
- Capacidades:
 - ▶ Ponta-a-ponta?
 - ▶ Global?
- Recursos:
 - ▶ Estatísticas?
 - ▶ Dados estruturados?
 - ▶ Dicionário de entidade?

Trabalhos relacionados II

Tabela: Comparação dos trabalhos lidos.

Autores	Ano	Capacidades		Recursos		
		Ponta-a-ponta	Global	Estatísticas	Dados estr.	Dicionário
Tsai et al. [5]	2016		✓	✓		
Ganea et al. [6]	2017		✓	✓		✓
Pappu et al. [7]	2017	✓	✓	✓		✓
Upadhyay et al. [8]	2018		✓	✓	✓	
Kolitsas et al. [9]	2018	✓	✓	✓		✓*
Gillick et al. [10]	2019				✓	✓
Le et al. [11]	2019				✓	
Logeswaran et al. [12]	2019					✓
Le et al. [13]	2019		✓	✓	✓	✓*
Broscheit [14]	2019	✓				
Wu et al. [15]	2019					✓
Onoe et al. [16]	2020			✓	✓	

* Indiretamente: usa embeddings de entidade treinados com dicionário de entidade.

- Ponta-a-ponta: realiza detecção de menção—caso contrário, assume-se fornecimento das fronteiras de menção.
- Global: informação global.
- Estatísticas: estatísticas de frequências entidade-menção.
- Dados estr.: dados estruturados.
- Dicionário: dicionário de entidade.

- Sistema de ligação de entidades para cenários com poucos recursos:
 - ▶ independência de dicionário de entidades;
 - ▶ independência de estatísticas de frequência;
 - ▶ independência de dados estruturados.
- Um passo além do trabalho em zero-shot.
- Possibilita trabalhar com bases de conhecimento que consistem somente em IDs de entidade sem descrição textual.

- Usar transferência de aprendizado (i.e. modelos pre-treinados com modelagem de linguagem).
- Tratar a tarefa como problema de aprendizado de distância:

$$L = \sum_{i=1}^n \max(\|f(x_i^a) - f(x_i^p)\|^2 - \|f(x_i^a) - f(x_i^n)\|^2 + \alpha, 0), \quad (1)$$

- Desafio: como codificar entidades sem dicionário de entidades?
 - ▶ Pular codificação de entidade?
 - ▶ Agregar codificações de menções?
 - ▶ Automatizar descritor de entidade?

Conjuntos de datos

- Wikipedia.
- Wikia zero-shot corpus [12].
- TACKBP-2010 [17].

$$\text{Recall@k} = \frac{n}{m} . \quad (2)$$

$$\text{Unnormalised accuracy} = \frac{c}{m} . \quad (3)$$

$$\text{Normalised accuracy} = \frac{d}{n} . \quad (4)$$

Cronograma

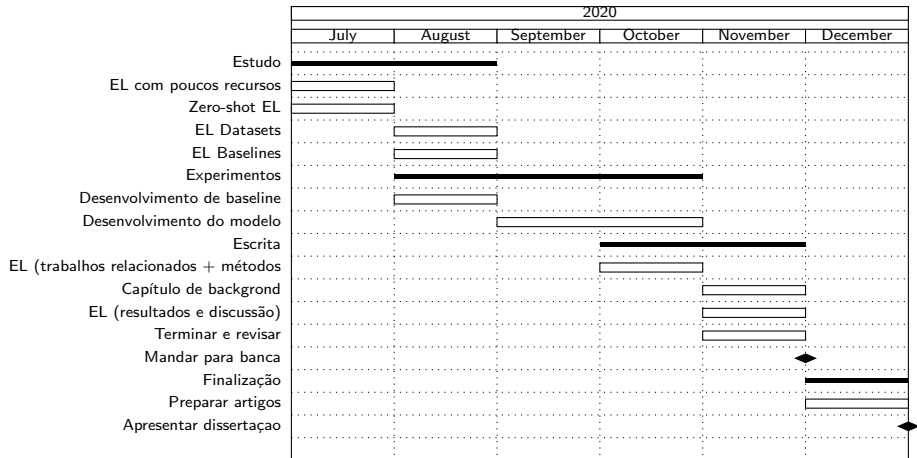


Figura: Plano de ataque mensal.

Referências I



Pedro H. Luz de Araujo, T. E. de Campos, F. Ataidez Braz, and N. Correia da Silva, “VICTOR: a dataset for Brazilian legal documents classification,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 1449–1458, European Language Resources Association, May 2020.






Pedro H. Luz de Araujo, T. E. de Campos, and M. Magalhaes Silva de Sousa, “Inferring the source official texts: can SVM beat ULMFiT?,” in *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), (Evora, Portugal), Springer, March 2-4 2020.




Code and data available from

<https://cic.unb.br/~teodecampos/KnEDLe/>.




Referências II

-  [Pedro H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, “Lener-br: a dataset for named entity recognition in brazilian legal text,” in *International Conference on the Computational Processing of Portuguese \(PROPOR\)*, \(Canela, RS, Brazil\), September 24-26 2018.](#)
-  [J. Howard and S. Ruder, “Fine-tuned language models for text classification,” *CoRR*, vol. abs/1801.06146, 2018.](#)
-  [C.-T. Tsai and D. Roth, “Cross-lingual wikification using multilingual embeddings,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, \(San Diego, California\), pp. 589–598, Association for Computational Linguistics, June 2016.](#)




Referências III

-  O.-E. Ganea and T. Hofmann, “Deep joint entity disambiguation with local neural attention,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 2619–2629, Association for Computational Linguistics, Sept. 2017.
-  A. Pappu, R. Blanco, Y. Mehdad, A. Stent, and K. Thadani, “Lightweight multilingual entity extraction and linking,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM*, (New York, NY, USA), p. 365–374, Association for Computing Machinery, 2017.
-  S. Upadhyay, N. Gupta, and D. Roth, “Joint multilingual supervision for cross-lingual entity linking,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 2486–2495, Association for Computational Linguistics, Oct.-Nov. 2018.




Referências IV

-  N. Kolitsas, O.-E. Ganea, and T. Hofmann, “End-to-end neural entity linking,” in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, (Brussels, Belgium), pp. 519–529, Association for Computational Linguistics, Oct. 2018.
-  D. Gillick, S. Kulkarni, L. Lansing, A. Presta, J. Baldrige, E. Ie, and D. Garcia-Olano, “Learning dense representations for entity retrieval,” 2019.
-  P. Le and I. Titov, “Distant learning for entity linking with automatic noise detection,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4081–4090, Association for Computational Linguistics, July 2019.

Referências V

-  L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, and H. Lee, “Zero-shot entity linking by reading entity descriptions,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3449–3460, Association for Computational Linguistics, July 2019.
-  P. Le and I. Titov, “Boosting entity linking performance by leveraging unlabeled documents,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 1935–1945, Association for Computational Linguistics, July 2019.
-  S. Broscheit, “Investigating entity knowledge in BERT with simple neural end-to-end entity linking,” in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, (Hong Kong, China), pp. 677–685, Association for Computational Linguistics, Nov. 2019.

Referências VI

-  L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer, “Zero-shot Entity Linking with Dense Entity Retrieval,” *arXiv e-prints*, p. arXiv:1911.03814, Nov. 2019.
-  Y. Onoe and G. Durrett, “Fine-Grained Entity Typing for Domain Independent Entity Linking,” *arXiv e-prints*, p. arXiv:1909.05780, Sept. 2019.
-  H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, “Overview of the TAC 2010 knowledge base population track,” in *Text Analysis Conference*, vol. 3, 2010.