

Explorando as funções de rotulagem do snorkel para textos da área de farmacovigilância

Leandro Zirondi de Sousa
(leandro.sousa@aluno.cefet-rj.br)

Orientadora: Prof^a. Dra. Kele Belloze

Rio de Janeiro
2020

Sumário

1. Introdução

2. Metodologia

3. Resultados

4. Discussão

5. Conclusão

6. Referências

Introdução

O problema de *datasets* rotulados

- ▶ Houve um avanço na área de aprendizado de máquina nos últimos anos graças ao uso de grandes *datasets* rotulados;
- ▶ Entretanto, tais *datasets* são extremamente custosos de se produzir, tanto por volume de dados necessários quanto por sua rotulação.


cord_uid	sha	source_x	title	doi	pmcid	pubmed_id
287145 unique values	[null] 63% 0ed3c6a5559cd73... 0% Other (111982) 37%	WHO 33% Medline 24% Other (130709) 43%	249034 unique values	[null] 38% 10.1016/j.scitotenv.2... 0% Other (186824) 62%	[null] 61% PMC35282 0% Other (116336) 39%	
ug7v899j	d1aafb78c866a2868b82 786f8929fd9c988897fb	PMC	Clinical features of culture-proven Mycoplasma pneumoniae infections at King Abdulaziz University Ho...	10.1186/1471-2334-1-6	PMC35282	11472636
02tnwd4n	6b8567729c2143a66d73 7eb8a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in lung disease?	10.1186/rr14	PMC59543	11667967

Figura: COVID-19 Open Research Dataset Challenge (CORD-19)

Snorkel

- ▶ Ferramenta que busca a automatização de rotulação de *datasets* em linguagem natural;
- ▶ Rotulação por meio de “*Weak Supervision*”.



snorkel

Farmacovigilância

- ▶ É a ciência e atividades relativas à detecção, compreensão e prevenção de eventos adversos relacionados a medicamentos, como define a Organização Mundial de Saúde (OMS);
- ▶ Se beneficiaria já que grande parte do seu conteúdo são textos em linguagem natural.

C. Descrição da reação adversa. Se o paciente ainda não se recuperou, assinale o campo "Data do fim da reação" com um traço.

Reação*	Data de início da reação*	Data do fim da reação*
1.		
2.		
3.		
Relato clínico do caso e das reações, com dados laboratoriais relevantes.		

Figura: Extrato do formulário para notificação de suspeita de reação adversa a medicamento do Centro de Vigilância Sanitária do Estado de São Paulo

Metodologia

A ferramenta Snorkel se estrutura em quatro etapas principais:

- ▶ Leitura do *corpus* e geração de sentenças;
- ▶ Definição das relações e os candidatos;
- ▶ Criação das *Labeling Functions* e *Gold Labels*;
- ▶ Treinamento do modelo.

Corpus e Sentenças

- ▶ O *corpus* é um conjunto de documentos alimentado ao Snorkel;
- ▶ Cada documento do *corpus* será dividido em sentenças;
- ▶ Nestas sentenças que ocorrerão o processamento.

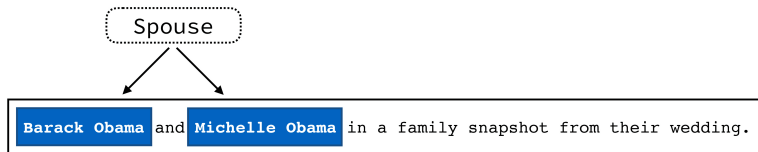


Figura: Exemplo de uma sentença

Relação e Candidatos

- ▶ A relação é o esquema do relacionamento que se quer extrair da sentença;
- ▶ É um esquema de relacionamento entre n-partes de uma sentença onde pode-se conter a informação procurada no texto.

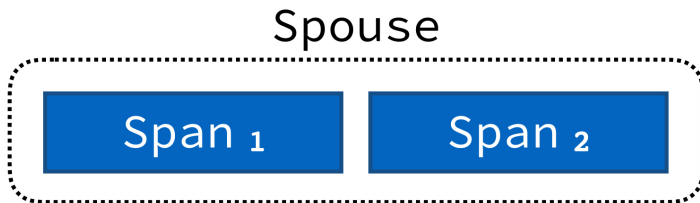


Figura: Exemplo de uma relação

Labeling Functions e Gold Labels

- ▶ *Labeling Functions* são funções utilizadas para validar candidatos a partir de algum conhecimento prévio, sendo divididas em duas categorias:
 - ▶ Padrão
 - ▶ *Distant Supervision*
- ▶ *Gold Labels* são sentenças rotuladas a mão pelo usuário, exemplificando a relação nas sentenças

Resultados

Problemas com Instalação e Versões

- ▶ A ferramenta recentemente teve uma mudança grande no seu último *release*, incorrendo em mudanças diversas;
- ▶ Documentação e suporte foram desafios enfrentados;
- ▶ O *script* de instalação gerou problemas, já que certas dependências não eram instaladas propriamente e geravam problemas apenas no meio da execução;
- ▶ A ferramenta foi instalada, configurada e testada executando-se os tutoriais disponibilizados cujo domínio do *corpus* era de notícias.

Datasets utilizados

- ▶ Com a ferramenta em funcionamento, foram definidos então os *datasets* necessários para sua execução, considerando o domínio da farmacovigilância:
 - ▶ Um *corpus*;
 - ▶ Para execução em português, é possível definir funções *matchers* com dicionários em português.
- ▶ Foram utilizados os *datasets* do trabalho "Detecção de sinais de eventos adversos de medicamentos em textos informais" da onde foram extraídos o *corpus* e os dicionários em português:
 - ▶ O *corpus* é um conjunto de *tweets*;
 - ▶ Os dicionários são um dicionário de medicamentos e um dicionário de eventos adversos.

Labeling Functions

- ▶ Com os candidatos em mãos, a próxima etapa é a definição das *labeling functions*;
- ▶ Porém, antes desse desenvolvimento foi visto a necessidade de se entender melhor os textos dos candidatos.

Discussão

LIWC e Ontologias

- ▶ O LIWC (Linguistic Inquiry and Word Count) em português, assim como ontologias biomédicas podem ser utilizados para verificar quais palavras e ideias estão sendo descritas entre os candidatos de uma sentença;
- ▶ O LIWC, por sua vez, poderia também ser utilizado numa *labeling function*, cruzando referências acerca de verbos e palavras que trazem a ideia de causalidade.

Estrutura da Língua

- ▶ A estrutura da língua portuguesa também precisa ser analisada para verificar as possibilidades de funções mais baseadas em estrutura da língua;
- ▶ "Alergia à alguma coisa";
- ▶ Um *corpus* formado por *tweets* em linguagem informal pode ter problemas com a estrutura.

Datasets rotulados em português

- ▶ Falta de *datasets* em português podem trazer problemas;
- ▶ Comparação final do resultado da execução com um conhecimento prévio de eventos adversos de cada medicamento;
- ▶ *Web scraping* poderia ajudar a resolver essa falta de *datasets*.

Conclusão

Conclusões

- ▶ Uso da língua portuguesa é viável, porém com ressalvas;
- ▶ Há uma falta de suporte para esta versão da ferramenta, com versão mais atuais cumprindo outras funções;
- ▶ Ainda falta terminar a execução da ferramenta com os *datasets* obtidos além de um estudo mais profundo das *Labeling Functions*.

6. Referências I



BOAS práticas de farmacovigilância para as Américas. [S.I.], 2011.



CHEN, Xiaoyi et al. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methyphenidate. [S.I.].



CUNHA, Alexandre Martins da. DETECÇÃO DE SINAIS DE EVENTOS ADVERSOS DE MEDICAMENTOS EM TEXTOS INFORMAIS. Rio de Janeiro, 2019. p. 134.



RATNER, Alexander et al. ASnorkel: Rapid Training Data Creation. Stanford, 2018. p. 14.



_____. Snorkel: Rapid Training Data Creation. [S.I.], 2017.



SEN, Chen et al. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. [S.I.], 2017.



ZHANG, Ce et al. DeepDive: Declarative Knowledge Base Construction. [S.I.].

Obrigado!