Processamento de Linguagem Natural

Experimentos em PLN e Morfologia Computacional

Marlo Souza[1]
[1]Universidade Federal da Bahia - Brasil

23 de julho de 2024

Sumário

Datasets e *Corpora*

Morfologia Computacional e Morfossintaxe

Datasets e *Corpora*

Dataset

Um dataset é um conjunto de dados, i.e. elementos que, organizados e tratados, produzem informação. Em PLN, *datasets* são compostos de dados linguísticos, que podem ser:

Dataset

Um dataset é um conjunto de dados, i.e. elementos que, organizados e tratados, produzem informação. Em PLN, *datasets* são compostos de dados linguísticos, que podem ser:

- Palavras com suas classificações gramaticais

Dataset

Um dataset é um conjunto de dados, i.e. elementos que, organizados e tratados, produzem informação. Em PLN, *datasets* são compostos de dados linguísticos, que podem ser:

- Palavras com suas classificações gramaticais
- Palavras ou unidades maiores, que podem ser classificadas como pessoa

Dataset

Um dataset é um conjunto de dados, i.e. elementos que, organizados e tratados, produzem informação. Em PLN, *datasets* são compostos de dados linguísticos, que podem ser:

- Palavras com suas classificações gramaticais
- Palavras ou unidades maiores, que podem ser classificadas como pessoa
- Pares de frases classificadas quanto ao seu grau de similaridade

Corpus

Um corpus é um *conjunto de dados linguísticos organizado*. O material que compõe um corpus (os textos) é coletado *com algum propósito* e *contém exemplos reais do uso da língua*.

Corpus

Um corpus é um *conjunto de dados linguísticos organizado*. O material que compõe um corpus (os textos) é coletado *com algum propósito* e *contém exemplos reais do uso da língua*.

Um corpus, para ser considerado um dataset linguístico no PLN, precisa ter tamanho o suficiente para permitir avaliar de maneira confiável o resultado de uma análise automática.

Corpus

Um corpus é um *conjunto de dados linguísticos organizado*. O material que compõe um corpus (os textos) é coletado *com algum propósito* e *contém exemplos reais do uso da língua*.

Um corpus, para ser considerado um dataset linguístico no PLN, precisa ter tamanho o suficiente para permitir avaliar de maneira confiável o resultado de uma análise automática.

A diferença pode ser também a nível de utilidade quando um corpus anotado passa a ser usado como material para treinar sistemas, ele também pode ser visto como um dataset.

Anotação Linguística e dados rotulados

A anotação é uma atividade de classificação: temos um conjunto de classes (as etiquetas) – também chamado de *tagset* – previamente definidas e critérios que guiam a classificação.

Anotação Linguística e dados rotulados

A anotação é uma atividade de classificação: temos um conjunto de classes (as etiquetas) – também chamado de *tagset* – previamente definidas e critérios que guiam a classificação.

1. TRISTE$_{SUBST}$[0] é$_{V}$[0] uma$_{ART}$[0] palavra$_{SUBST}$[0] de$_{PREP}$[0] 6$_{NUM}$[0] letras$_{SUBST}$[0] [frase neutra]

2. Um$_{NUM}$[0] dos$_{PREP+ART}$[0] livros$_{SUBST}$[0] mais$_{ADV}$[0] tristes$_{ADJ}$[–] que$_{PRON-Rel}$[0] já$_{ADV}$[0] li$_{V}$[0] [frase neutra?]

3. Sofri$_{V}$[–] com$_{PREP}$[0] a$_{ART}$[0] protagonista$_{SUBST}$[0] a$_{PREP}$[0] cada$_{PRON}$[0] nova$_{ADJ}$[0] página$_{SUBST}$[0]; sofri$_{V}$[–] quando$_{ADV}$[0] o$_{ART}$[0] livro$_{SUBST}$[0] acabou$_{V}$[0] [frase positiva]

4. Nunca$_{ADV}$[0] sofri$_{V}$[–] tanto$_{ADV}$[0] para$_{PREP}$[0] ler$_{V}$[0] um$_{ART}$[0] livro$_{SUBST}$[0] [frase negativa]

Anotação Linguística e dados rotulados

A anotação é uma atividade de classificação: temos um conjunto de classes (as etiquetas) – também chamado de *tagset* – previamente definidas e critérios que guiam a classificação.

1. TRISTE$_{SUBST}$[0] é$_{V}$[0] uma$_{ART}$[0] palavra$_{SUBST}$[0] de$_{PREP}$[0] 6$_{NUM}$[0] letras$_{SUBST}$[0] [frase neutra]

2. Um$_{NUM}$[0] dos$_{PREP+ART}$[0] livros$_{SUBST}$[0] mais$_{ADV}$[0] tristes$_{ADJ}$[–] que$_{PRON-Rel}$[0] já$_{ADV}$[0] li$_{V}$[0] [frase neutra?]

3. Sofri$_{V}$[–] com$_{PREP}$[0] a$_{ART}$[0] protagonista$_{SUBST}$[0] a$_{PREP}$[0] cada$_{PRON}$[0] nova$_{ADJ}$[0] página$_{SUBST}$[0]; sofri$_{V}$[–] quando$_{ADV}$[0] o$_{ART}$[0] livro$_{SUBST}$[0] acabou$_{V}$[0] [frase positiva]

4. Nunca$_{ADV}$[0] sofri$_{V}$[–] tanto$_{ADV}$[0] para$_{PREP}$[0] ler$_{V}$[0] um$_{ART}$[0] livro$_{SUBST}$[0] [frase negativa]

É fundamental ter em mente que a anotação é uma atividade interpretativa por parte do (grupo de) anotador(es).

Datasets: pra quê?

A existência de datasets linguísticos, ou corpora padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN por três motivos principais:

Datasets: pra quê?

A existência de datasets linguísticos, ou corpora padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN por três motivos principais:

- São necessários para treinar modelos

Datasets: pra quê?

A existência de datasets linguísticos, ou corpora padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN por três motivos principais:

- São necessários para treinar modelos
- São necessários para avaliar e comparar modelos

Datasets: pra quê?

A existência de datasets linguísticos, ou corpora padrão ouro, é fundamental para uma série de tarefas e aplicações de PLN por três motivos principais:

- São necessários para treinar modelos
- São necessários para avaliar e comparar modelos
- São necessários para estabelecer uma interpretação comum do fenômeno e determinar caminhos futuros de desenvolvimento

Características de um bom dataset linguístico

Uma avaliação adequada, justa, abrangente, detalhada, sistemática e transparente é um passo essencial ao se desenvolver, construir, analisar, comparar e usar tecnologias de linguagem.

Características de um bom dataset linguístico

Uma avaliação adequada, justa, abrangente, detalhada, sistemática e transparente é um passo essencial ao se desenvolver, construir, analisar, comparar e usar tecnologias de linguagem.
Toda avaliação está situada em um contexto e deve ser feita de acordo com ele. Assim, apesar de determinadas práticas avaliativas serem difundidas dentro da área, não existe uma forma correta e definitiva de se avaliar um sistema, apenas questões que devem ser consideradas nas práticas de avaliação.

Avaliação de sistemas de PLN

9

As avaliações de sistemas podem ser classificadas em relação às seguintes dimensões:

- manual ou automática
- intrínseca e extrínseca, ou direta e indireta
- formativa ou sumativa
- por usuários ou por especialistas
- componente ou ponta-a-ponta
- transparente ou às escuras
- investigativa ou experimental
- qualitativa ou quantitativa
- objetiva ou subjetiva
- supervisionada ou não-supervisionada

Leaderboards

Uma prática popular se resume a tentar melhorar uma métrica em um leaderboard.

Model	Clemscore	% Played	Quality Score
claude-2.1	36.38	83.08	43.79
claude-2	33.71	82.12	41.05
gpt-3.5-turbo-0613	32.53	91.96	35.37
gpt-3.5-turbo-1106	30.45	77.12	39.49
openchat_3.5	19.72	57.57	34.26
sheep-duck-llama-2-70b-v1.1	17.12	40.82	41.93
Yi-34B-Chat	16.77	63.76	26.3
WizardLM-70b-v1.0	16.7	51.65	32.34
Mixtral-8x7B-Instruct-v0.1	16.53	57.68	28.66
tulu-2-dpo-70b	15.9	54.49	29.18
claude-instant-1.2	15.44	59.61	25.91
CodeLlama-34b-Instruct-hf	10.34	23.96	43.15

Figura: Leaderboard extraído do benchmark clembench (Madureira,2024)

Leaderboards

Embora esse paradigma ajude a fomentar um progresso mensurável, ele vem acompanhado de muitas críticas: foco na métrica não no uso real da tecnologia; corrompimento da métrica, etc. Ainda assim, avaliações conjuntas de tarefas podem trazer consideráveis evoluções para sua solução, especialmente para tarefas novas.

Elementos de uma avaliação experimental em PLN

- Hipóteses e Experimentos: avaliações são projetaddas com base numa metodologia experimental na qual tentamos validar uma hipótese de forma controlada
- Material de referência: dados rotulados ou históricos são comumente necessários para permitir a avaliação da performance do sistema em relação aos objetivos e hipóteses pré-estabelecidas
- Metodologia de avaliação: é preciso garantir que o esforço de avaliação seja correto (i.e. confiável) e replicável.
- Métricas experimentais: conjunto de medidas que são utilizadas na avaliação quantitativas de sistema para representar sua performance

Partição dos dados

Dados que seja utilizados no processo de desenvolvimento da solução não devem ser usados para sua avaliação, uma vez que o sistema foi desenvolvido especificamente para tratar desses caos. Assim, é necessário realizar uma partição dos dados, de forma que os dados de avaliação não tenham sido observados pelo sistema ou pelos projetistas durante o processo de desenvolvimento.

Partição dos dados

Dados que seja utilizados no processo de desenvolvimento da solução não devem ser usados para sua avaliação, uma vez que o sistema foi desenvolvido especificamente para tratar desses caos. Assim, é necessário realizar uma partição dos dados, de forma que os dados de avaliação não tenham sido observados pelo sistema ou pelos projetistas durante o processo de desenvolvimento. Técnicas de partição de dados

- Partição dos dados: treino, teste, desenvolvimento
- Validação cruzada

Baselines e Validação estatística

Comumente devemos comparar a performance do sistema proposto com uma solução inicial básica ou pré-existente para garantir que a performance do sistema justifica sua aplicação. A tais sistemas chamams de *baseline*.

Baselines e Validação estatística

Comumente devemos comparar a performance do sistema proposto com uma solução inicial básica ou pré-existente para garantir que a performance do sistema justifica sua aplicação. A tais sistemas chamams de *baseline*.
A elaboração de hipóteses experimentais usualmente envolvem a comparação da performance do sistema em relação aos *baselines* escolhidos e a validação das mesmas se dá através de testes estatísticos.

Análise de Erros

A simples validação estatística da hipótese não nos fornece base para compreender o funcionamento do sistema proposto - do contrário recaímos num reducionismo (como discutido anteriormente sobre os leaderboards).

Análise de Erros

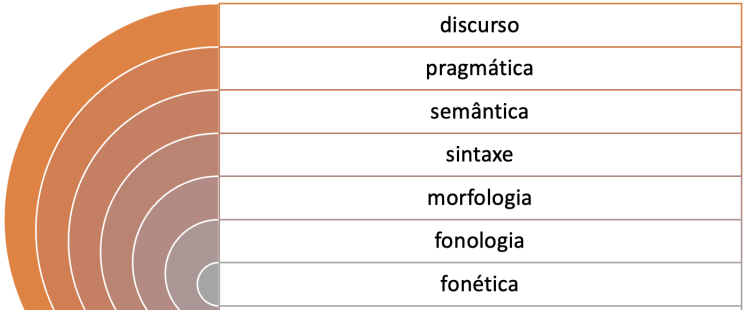
A simples validação estatística da hipótese não nos fornece base para compreender o funcionamento do sistema proposto - do contrário recaímos num reducionismo (como discutido anteriormente sobre os leaderboards).
É preciso então investigar de formas diferentes, de preferência de forma qualitativa, a performance do sistema, levando em consideração aspectos não quantificados.

Análise de Erros

A simples validação estatística da hipótese não nos fornece base para compreender o funcionamento do sistema proposto - do contrário recaímos num reducionismo (como discutido anteriormente sobre os leaderboards).
É preciso então investigar de formas diferentes, de preferência de forma qualitativa, a performance do sistema, levando em consideração aspectos não quantificados.
Uma tal forma é a análise qualitativa de erros, que nos permite entender de forma minuciosa o comportamento do sistema e quais suas limitações de aplicação. Ela consiste em analisar manualmente as saídas do sistema e os erros cometidos pelo mesmo a fim de levantar hipóteses para melhorias futuras.

Morfologia Computacional e Morfossintaxe

Níveis de Análise Linguística



Morfologia

A morfologia é a área de linguística que estuda a estrutura e o processo de formação formação das palavras de uma língua.

Morfologia

A morfologia é a área de linguística que estuda a estrutura e o processo de formação formação das palavras de uma língua.

$$\text{casa} \xrightarrow[\textit{diminutivo}]{} \text{casinha}$$

Morfologia

A morfologia é a área de linguística que estuda a estrutura e o processo de formação formação das palavras de uma língua.

$$\text{casa} \xrightarrow[\textit{diminutivo}]{} \text{casinha}$$

A menor unidade dotada de significado é chamada de **morfema** (e.g., inha). Em uma explicação simples, podemos dizer que os morfemas são os pedacinhos que se juntam para formar as palavras

Palavras e tokens

O conceito de palavra em linguística é um conceito complexo.

Palavras e tokens

O conceito de palavra em linguística é um conceito complexo.Segundo Cabré (1999, p. 20), as palavras são as unidades de referência da realidade empregadas pelos falantes.

Palavras e tokens

O conceito de palavra em linguística é um conceito complexo.Segundo Cabré (1999, p. 20), as palavras são as unidades de referência da realidade empregadas pelos falantes.

O léxico constitui o léxico consiste no conjunto das palavras de uma língua e dos padrões que possibilitam a criatividade do falante.

Palavras e tokens

O conceito de palavra em linguística é um conceito complexo.Segundo Cabré (1999, p. 20), as palavras são as unidades de referência da realidade empregadas pelos falantes.

O léxico constitui o léxico consiste no conjunto das palavras de uma língua e dos padrões que possibilitam a criatividade do falante.

Em PLN, empregamos o conceito de *token*, ou palavra computacional, que significa qualquer sequência de caracteres à qual se atribui um valor. Exemplo: a sentença

"Eu sempre viajo para Campinas, para Salvador e para Belém"

possui os tokens (''eu", "sempre", "viajo", "para", "Campinas", ",", "Salvador", "e", "Belém" e ".")

A Morfologia Computacional

20

A morfologia computacional é a área do PLN que estuda técnicas para a análise de propriedades morfológicas do texto ou da língua.

A Morfologia Computacional

A morfologia computacional é a área do PLN que estuda técnicas para a análise de propriedades morfológicas do texto ou da língua.
São tarefas comumente estudadas no processamento morfológico:

- a sentenciação
- a tokenização
- a tokenização em subpalavras (descoberta de morfemas ou de morfotokens)
- a normalização
- a análise morfossintática
- a análise de atributos morfológicos

A Sentenciação

A sentenciação (ou sentenciamento) é o processo de segmentação do texto em sentenças, ou seja, é o processo de identificação de unidades textuais de processamento onde se definem os limites de cada sentença.

A Sentenciação

A sentenciação (ou sentenciamento) é o processo de segmentação do texto em sentenças, ou seja, é o processo de identificação de unidades textuais de processamento onde se definem os limites de cada sentença.
Estratégias comuns empregadas para a sentenciação:

- Abordagens baseadas em regras
- Abordagens baseadas em aprendizado de máquina supervisionado
- Abordagens baseadas em aprendizado de máquina não supervisionado,

A Sentenciação

A sentenciação (ou sentenciamento) é o processo de segmentação do texto em sentenças, ou seja, é o processo de identificação de unidades textuais de processamento onde se definem os limites de cada sentença.
Estratégias comuns empregadas para a sentenciação:

- Abordagens baseadas em regras
- Abordagens baseadas em aprendizado de máquina supervisionado
- Abordagens baseadas em aprendizado de máquina não supervisionado,

Fui à clínica do Dr. Nilo.

A Tokenização

A separação em unidades linguísticas mínimas é denominada tokenização. No caso do português é feita partindo da separação das palavras através de delimitadores.

A Tokenização

A separação em unidades linguísticas mínimas é denominada tokenização. No caso do português é feita partindo da separação das palavras através de delimitadores.Outra tarefa frequente da tokenização é a separação de palavras contraídas, por exemplo, a palavra “da” é separada em dois tokens: “de”+“a”

A Tokenização de subpalavras

Uma tarefa relacionada à tokenização, denominada tokenização de subpalavras, busca identificar fragmentos de palavras que possuem conteúdo relevante no processamento morfológico de uma palavra. Apesar de ter se popularizado recentemente com sua utilização em modelos de linguagem, é estudado no contexto de morfologia computacional não supervisionada para identificação de morfemas.

A Normalização

A normalização é a tarefa que converte as palavras para alguma forma padrão. São exemplos de normalização: conversão de versões abreviadas de palavras, conversão para caracteres minúsculos , lematização (e.g., estabelecendo que "somos" é uma conjugação do verbo "ser") e radicalização (e.g., estabelecendo que "retrabalho" tem o radical "trabalho").

A Análise Morfossintática

A análise morfossintática, ou PoS tagging, é uma técnica fundamental na área de PLN que envolve a atribuição de etiquetas gramaticais a cada palavra em um texto, com base na sua classe gramatical e em suas características morfológicas.

		DET	NOUN	ADJ	VERB	ADV	PUNCT		
	<s>	O	gato	preto	correu	rapidamente	.	</s>	

Figura: Exemplo de anotação de PoS (Finatto et al, 2024)

Anotação de atributos morfológicos

A anotação de atributos morfológicos, também conhecida como atribuição de features morfológicas (ou somente feats), é uma importante tarefa que envolve a marcação ou identificação de informações específicas sobre as características gramaticais e morfológicas de palavras em um texto. Esses atributos morfológicos incluem características como número, gênero, modo, tempo, pessoa e outras informações semelhantes.

					Mood=Ind Number=Sing Person=3 Tense=Past			
		Gender=Masc Number=Sing	Gender=Masc Number=Sing	Gender=Masc Number=Sing				
		O	GATO	PRETO	CORRER	RAPIDAMENTE	.	
		DET	NOUN	ADJ	VERB	ADV	PUNCT	
	<s>	O	gato	preto	correu	rapidamente	.	</s>

Figura: Exemplo de anotação de features morfológicas (Finatto et al, 2024)