



CORPORA

Luísa Coheur

OVERVIEW

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- After this class, students should be able to explain:
 - What is a corpus and how it should be analysed, cleaned and used in experiments
 - Several concepts, such as annotators' agreement or wizard of oz
 - Different types of toxic data
 - Different data augmentation processes

TOPICS

OVERVIEW

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data augmentation
 - Data cleaning
 - Toxic data
 - Using Data: data splits
- Key takeaways
- Suggested readings

MOTIVATION

- The scientific method
 - Observation/Question: Identify a problem and formulate a [research question](#)
 - Hypothesis: Develop a [hypothesis](#) (testable and falsifiable prediction or tentative explanation that addresses the research question)
 - [Data Collection](#): Planning and gathering data
 - Experimentation: Designing and conducting [experiments](#) to test the hypothesis
 - Data Analysis: [check results](#), determine whether they support the hypothesis or not
 - Draw [conclusions](#)

MOTIVATION

- FACT: to **train and test** our models...



WE NEED

DATA

and sometimes

ANNOTATED

DATA

CORPORA

- A **corpus** is a collection of texts
 - Corpora (plural); corpus (singular)
- There are many, many, many corpora available with or without annotations
- There are many different types of annotations
 - Example:
 - Reviews are annotated as positive, neutral or negative
 - Words within a text are annotated with the correspondent morpho-syntactic category (verb, noun, etc.)
 - ...



**Sentiment
Analysis is an
NLP TASK!**



**PoS Tagging is an
NLP TASK (and we will
study how to do it)!**

CORPORA

- Some [companies/startups/whatever](#) gather/produce/sell corpora
 - Examples:
 - ELDA: Evaluations and Language Resources Distribution Agency
 - LDC: Linguistic Data Consortium
 - Kaggle
 - ...
- Some [conferences](#) and [journals](#) are exclusively dedicated to NLP resources, including corpora
 - Examples:
 - Language Resources and Evaluation conference (LREC)
 - Language Resources and Evaluation journal

CORPORA: EXAMPLES

Annotations

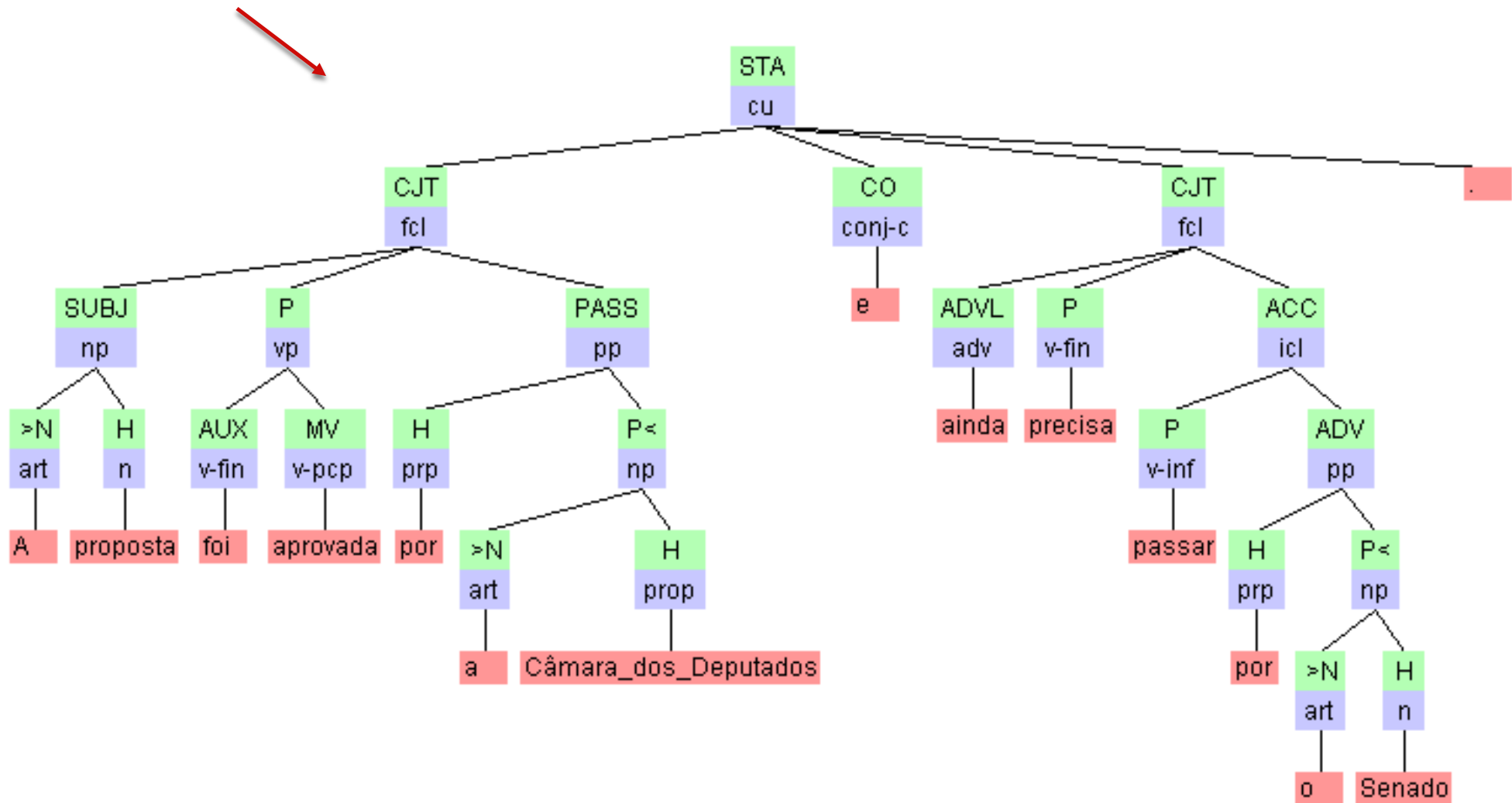


What is fibromyalgia ?	DESC:def	
What is done with worn or outdated flags ?	DESC:desc	
What does cc in engines mean ?	DESC:def	
When did Elvis Presley die ?	NUM:date	
What is the capital of Yugoslavia ?	LOC:city	
Where is Milan ?	LOC:city	
What is the speed hummingbirds fly ?	NUM:speed	
What is the oldest city in the United States ?	LOC:city	
What was W.C. Fields ' real name ?	HUM:ind	
What river flows between Fargo , North Dakota and Moorhead , Minnesota ?	LOC:other	
What do bats eat ?	ENTY:food	
What state did the Battle of Bighorn take place in ?	LOC:state	
Who was Abraham Lincoln ?	HUM:desc	
What do you call a newborn kangaroo ?	ENTY:termeq	
What are spider veins ?	DESC:def	
What day and month did John Lennon die ?	NUM:date	
What strait separates North America from Asia ?	LOC:other	

Li and Roth, Question Classification

CORPORA: EXAMPLES

Annotations



Example from Floresta Sintática (Linguatca)

CORPORA: EXAMPLES



- ABOUT
- MEMBERS
- COMMUNICATIONS
- LANGUAGE RESOURCES
- Data
- Obtaining Data
- Catalog
- By Year
- Top Ten Corpora**
- Projects
- Search
- Memberships
- Data Scholarships
- Tools
- Papers
- LR Wiki
- DATA MANAGEMENT
- COLLABORATIONS

Home > Language Resources > Data

Top Ten LDC Corpora

LDC2013T19	OntoNotes Release 5.0
LDC93S1	TIMIT Acoustic-Phonetic Continuous Speech Corpus
LDC2006T13	Web 1T 5-gram Version 1
LDC96L14	CELEX2
LDC99T42	Treebank-3
LDC2008T19	The New York Times Annotated Corpus
LDC93S10	TIDIGITS
LDC97S02	Switchboard-1 Release 2
LDC2006T06	ACE 2005 Multilingual Training Corpus
LDC2011T07	English Gigaword Fifth Edition

Example

CORPORA: EXAMPLES

Updates

As of April, 2015, TIDIGITS is also available in flac compressed wav. This package is available to licensees as an additional download. Not included in this version are the folders relating to handling the shortened sphere files of the original corpus.

Copyright

Portions © 1993 Trustees of the University of Pennsylvania

Available Media


- ☒ Web Download

Fees

\$0.00 1993 Member
\$500.00 Non-Member
\$250.00 Reduced-License
\$0.00 Extra Copy


Login for the applicable fee

CORPORA: EXAMPLES

 Featured Code Competition

Quora Insincere Questions Classification

Detect toxic content to improve online conversations

 Quora · 4,037 teams · 2 years ago

OverviewDataNotebooksDiscussionLeaderboardRules

New Topic

\$25,000
Prize Money



Dieter

22nd place

Augmentation for text

Posted in [quora-insincere-questions-classification](#) 2 years ago



117

Of course there are a lot of augmentation techniques for images, but what about text? Let's discuss some techniques:

1. Exchanging words with synonyms (see e.g. <https://arxiv.org/pdf/1502.01710.pdf>)
2. noising in RNN (<https://arxiv.org/pdf/1703.02573.pdf>)
3. Translation to other language and back (<https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/48038>)

CORPORA: EXAMPLES

The screenshot shows a web interface for a dataset. At the top, there is a search bar and buttons for 'Sign In' and 'Register'. The main header features the dataset title 'Friends Series Dataset' and subtitle 'Data about all 236 episodes of Friends Series', which are circled in red. Below this, the creator's name 'Mohammad Reza Ghari' and the update time 'updated 5 months ago' are displayed. A navigation bar includes links for 'Data', 'Tasks', 'Notebooks (6)', 'Discussion (1)', 'Activity', and 'Metadata', along with buttons for 'Download (56 KB)' and 'New Notebook'. A section below the navigation bar provides 'Usability' (10.0), 'License' (Data files © Original Authors), and 'Tags' (arts and entertainment). The 'Description' section contains two sub-sections: 'Context' and 'Content'.

Friends Series Dataset
Data about all 236 episodes of Friends Series

Mohammad Reza Ghari • updated 5 months ago

[Data](#) [Tasks](#) [Notebooks \(6\)](#) [Discussion \(1\)](#) [Activity](#) [Metadata](#) [Download \(56 KB\)](#) [New Notebook](#)

Usability 10.0 **License** Data files © Original Authors **Tags** arts and entertainment

Description

Context

Most of the times there are a lot of interesting insights behind the popular subjects online, specially the ones that are involved social interactions like votes, reviews and user-generated contents. Popular TV shows are in this category. Friends Sitcom TV show, is one of the most loved series and I found it useful to have a dataset online.

Content

This dataset consisting of 235 row each representing an episode of the show and 8 columns that are features of each episodes indexed on IMDB. I'll try to add more features in the future.

Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

BUILDING CORPORA

- Time-consuming and expensive
- Sometimes **experts** are needed to label the data
- Sometimes **the crowd** is used
 - Check, for instance, Amazon Mechanical Turk
 - Check the work of the Portuguese DefinedCrowd (currently Defined.ai)



DEFINED.AI

ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



THINK-PAIR-SHARE



Find examples of NLP tasks that require annotated data where the annotators do not need to be experts

BUILDING CORPORA: EXAMPLES DATA ALIGNED WITH CLUE-ALIGNER (Anabela Barreiro's project)

Text alignment is an NLP TASK!

The screenshot displays the CLUE-ALIGNER Alignment Tool interface, which is used for text alignment. The main window is titled "Alignment Tool (New Version)". It features a grid for aligning words from two different languages. The top row of the grid contains the English words: "we", "all", "agree", "that", "the", "trans", "european", "network", "must", "be", "a", "multimodal", "network". The left column contains the Portuguese words: "todos", "concordam", "em", "que", "a", "rede", "transeuropeia", "de", "transportes", "deve", "ser", "uma", "rede", "multimodal". The grid cells are colored black or white, indicating alignment. A dashed green box highlights the alignment between "concordam" and "agree", and "em que" and "that".

Navigation controls include a "Go to" field with the value "258" and "of 400", and buttons for "<-- Back" and "Next -->". A "Save" button is also present.

The "Multi-Word Units" panel on the right shows a list of aligned multi-word units, with a "Remove Multi-Word Unit" button above it. The units listed are:

- em que | that
- rede transeuropeia de transportes | trans - european network
- deve ser | must be
- rede multimodal | multimodal network
- todos | we all
- concordam em que | agree that

BUILDING CORPORA: EXAMPLES

TEXT SEGMENTATION (Pedro Mota's PhD)

=====

2: Kinematics: Describing motion.

Our first goal is understanding the motion of objects.

The first step is simple: merely DESCRIBING the motion of things.

1) We'll only talk about "particles": point like objects, whose structure is irrelevant.

2) We'll work in one dimension, e.g. a train moving back and forth on a straight track.

To describe motion, we need a few basic concepts, quantities, and definitions.

We'll use English language words but define them mathematically when possible.

You'll see that words like "velocity, acceleration, force, energy, momentum (which are often sloppy), are, in physics, totally distinct and well defined.

===== (A1, A2, A3)

1) POSITION: Where is the object?

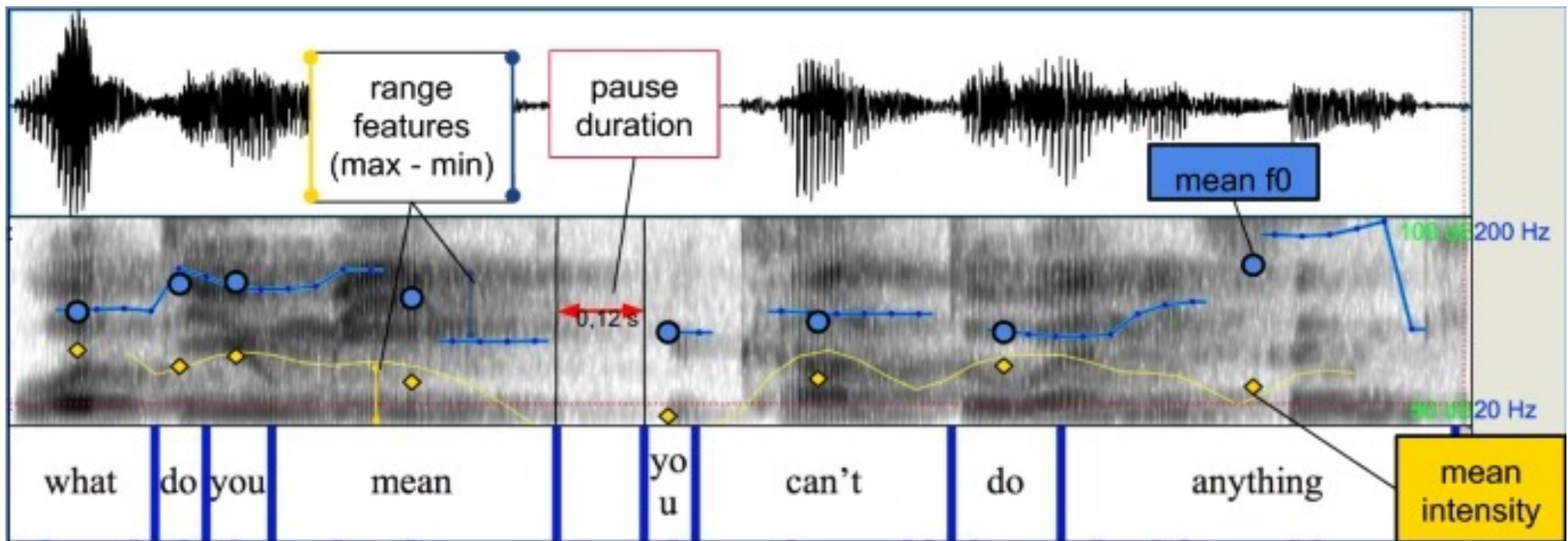
You need a reference frame to describe position.

A reference frame means a choice of axis and coordinate system: where is the origin, what units will we use to measure length, which direction will we call positive?

**Text
segmentation is
also an
NLP TASK!**

BUILDING CORPORA: EXAMPLES SPEECH RECOGNITION

**Speech
Recognition: a
course in P4!**



<https://link.springer.com/article/10.1007/s10579-021-09556-2>

BUILDING CORPORA: EXAMPLES

SIGN LANGUAGES

Arquivo Editar Anotação Trilha Tipo Buscar Visualizar Opções Janela Ajudar

Grade Texto Legenda Lexicon Comments Reconhecedores Metadados Controles

Volume: 100

Video_140_294.mp4

Mute Solo

Velocidade: 70

00:00:01.260 Seleção: 00:00:21.331 - 00:00:22.804 1473

Modo de Seleção Modo de Repetição (Loop)

LP_P1 transcrição livre [397]

Come_P1Literal [29]

LGP_P1Trans_Literal [101]

GLOSAS_P1 [456]

GLOSA_P1-M1 [451]

GLOSA_P1-M2 [153]

M2_ClassGram [144]

M1_ClassGram

Cultura, Arte, Teatro.

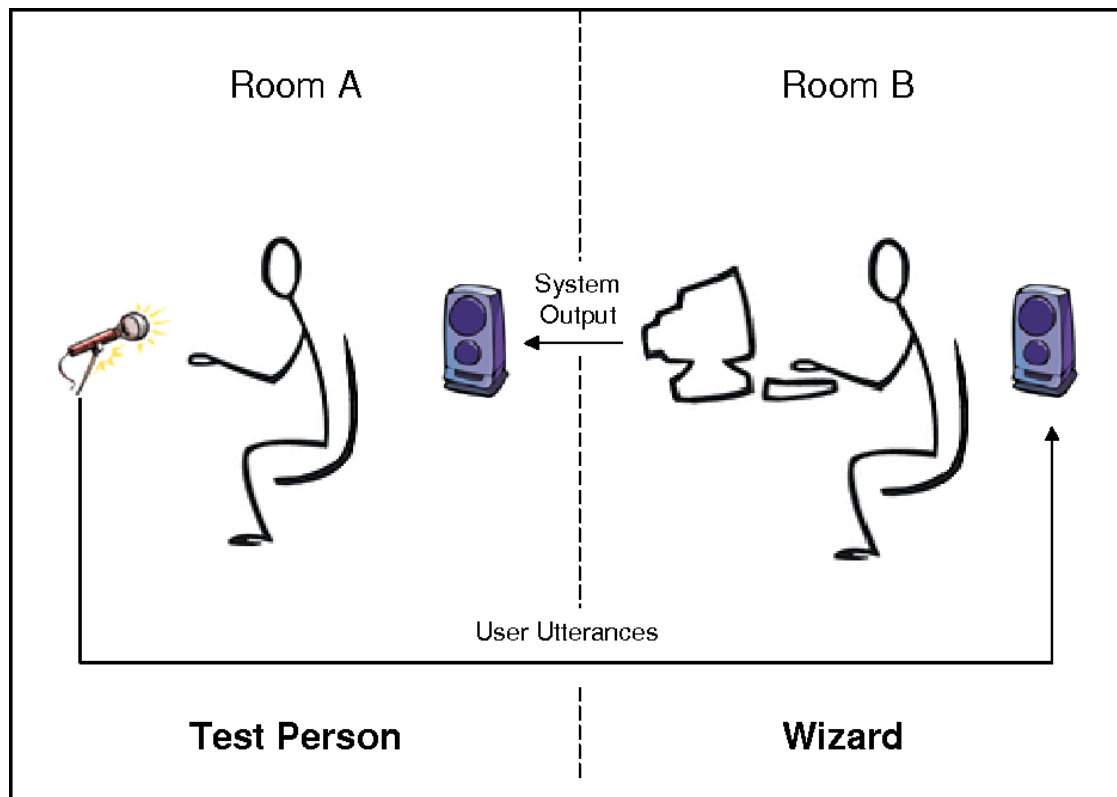
CULTURA ARTE TEATRO

CULTURA

CULTURA

BUILDING CORPORA: EXAMPLES AND IF WE HAVE NO CORPUS?

- If you don't have data that allows you to understand how your system will be used, try a **Wizard of Oz**.



Wizard of Oz: experiment in which subjects interact with a computer that they believe to be autonomous, although it is being operated by an unseen human

Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

INTER-ANNOTATOR AGREEMENT

- How good are the annotations?
 - A way to check their quality is to see the **agreement among annotators**. If the agreement is very low:
 - We have bad guidelines, or
 - The task is very difficult, or
 - The annotators did not pay much attention to their job
- Notice that if two humans don't agree... well... the machine will not work a miracle




INTER-ANNOTATOR AGREEMENT

- Different tasks need different metrics for annotator's agreement
- For each task, several metrics might exist
- Examples:
 - Cohen's kappa coefficient (two annotators), ← Sebenta and now
 - Fleiss Kappa (several annotators),
 - Window Difference (ex: text segmentation)
 - ...

ACTIVE LEARNING MOMENT



EXERCISE: ANNOTATORS AGREEMENT

- LABELS:
 - GEOGRAPHY, MUSIC, LITERATURE, HISTORY, SCIENCE
- Guidelines:
 - GOAL: Tag questions with the given labels according with their main focus
 - Examples:
 -  Stevie Nicks: "Edge Of ____" [Seventeen](#)
 -  Germany has Worms & this country that borders Germany has a district called Wormerveer
[The Netherlands \(Holland\)](#)
 -  In this 17th C. novel, Sancho Panza is ...
[Don Quixote](#)

EXERCISE: ANNOTATORS AGREEMENT

1. Annotate the corpus in the next slide (use your knowledge!)
2. Choose a colleague near you
3. Go to:



4. Calculate the agreement between you and your colleague
5. Discuss the results

TAGS: GEOGRAPHY, MUSIC, LITERATURE, HISTORY, SCIENCE



1	This book by Virginia Woolf inspired Michael Cunningham's novel "The Hours"	Mrs. Dalloway
2	The "amiable" former name of the Tongan archipelago	the Friendly Islands
3	The Rhine Valley occupies one-third of this 62-square-mile country; the Alps cover the rest	Liechtenstein
4	PBS fans know that "Evening at Pops" refers to this city's Pops	Boston
5	In 1996 he simultaneously published "The Regulators" as Richard Bachman & "Desperation" under this name	Stephen King
6	In 1843 Congress allocated \$30,000 to string one between Baltimore & Washington; it was completed in 1844	a telegraph wire
7	According to Chuck Jones, whenever possible, this force of nature was to be Wile E. Coyote's greatest enemy	gravity
8	This 1940 Disney film featured the music of Bach, Beethoven, Stravinsky, Schubert & Mussorgsky	Fantasia
9	The Babylonians kept abreast of the times using a form of this instrument seen here:	Sundial
10	Dying in 2009 at age 113, British WWI vet Henry Allingham was the last original surviving member of this group, formed 1918	the Royal Air Force
11	-273 Celsius	absolute zero
12	Perhaps the greatest violinist ever, this Italian could play a whole piece on just one string	Niccolo Paganini
13	Blink-182: "That's about the time she walked away from me, nobody likes you when you're ____"	23
14	Rolf Gruber & Mother Abbess of Nonnberg Abbey	The Sound of Music
15	In this song, David Bowie instructs, "Put on your red shoes and dance the blues"	Let's Dance
16	Longfellow wrote, "Tell me not" that "life is but an empty" this	dream
17	His 1543 book "Concerning the Revolutions of the Celestial Spheres" started an astronomical revolution	Nicholas Copernicus
18	Baby, boudoir & concert are 3 sizes of this type of piano	grand piano
19	Welcome MCR, this alt-rock group, to "The Black Parade", its 2006 concept album	My Chemical Romance
20	Iron filings are often used to demonstrate the presence of this field	magnetic field

Annotate and find the agreement [between you and your colleague](#)

EXERCISE: ANNOTATORS AGREEMENT

REFERENCE	ME						
		GEO	MUS	LIT	HIST	SCI	TOTAL
	GEO	2					2
	MUS	1	7				8
	LIT			2	1		3
	HIST				2		2
	SCI			1		4	5
TOTAL	3	7	3	3	4	*20*	

1	LITERATURE
2	GEOGRAPHY
3	GEOGRAPHY
4	MUSIC
5	LITERATURE
6	HISTORY
7	SCIENCE
8	MUSIC
9	SCIENCE
10	HISTORY
11	SCIENCE
12	MUSIC
13	MUSIC
14	MUSIC
15	MUSIC
16	LITERATURE
17	SCIENCE
18	MUSIC
19	MUSIC
20	SCIENCE

Quantify agreement with kappa results

	A	B	C	D	E	Total
A	2	0	0	0	0	2
B	1	7	0	0	0	8
C	0	0	2	1	0	3
D	0	0	0	2	0	2
E	0	0	1	0	4	5
Total	3	7	3	3	4	20

Number of observed agreements: 17 (85.00% of the observations)

Number of agreements expected by chance: 4.9 (24.25% of the observations)

Kappa= 0.802

SE of kappa = 0.103

95% confidence interval: From 0.601 to 1.000

The calculations above only consider exact matches between observers. If the categories (A, B, C...) are ordered, you may also wish to consider close matches. In other words, if one observer classifies a subject into group B and the other into group C, this is closer than if one classifies into A and the other into D. The calculation of weighted kappa, below, assumes the categories are ordered and accounts for how far apart the two raters are. This calculation uses linear weights.

EXERCISE: ANNOTATORS AGREEMENT

- Discussion:
 - His 1543 book "Concerning the Revolutions of the Celestial Spheres" started an astronomical revolution
 - LITERATURE? SCIENCE? HISTORY?
 - Unbalanced corpus (8 MUSIC in 20)
 - Consequences?



Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

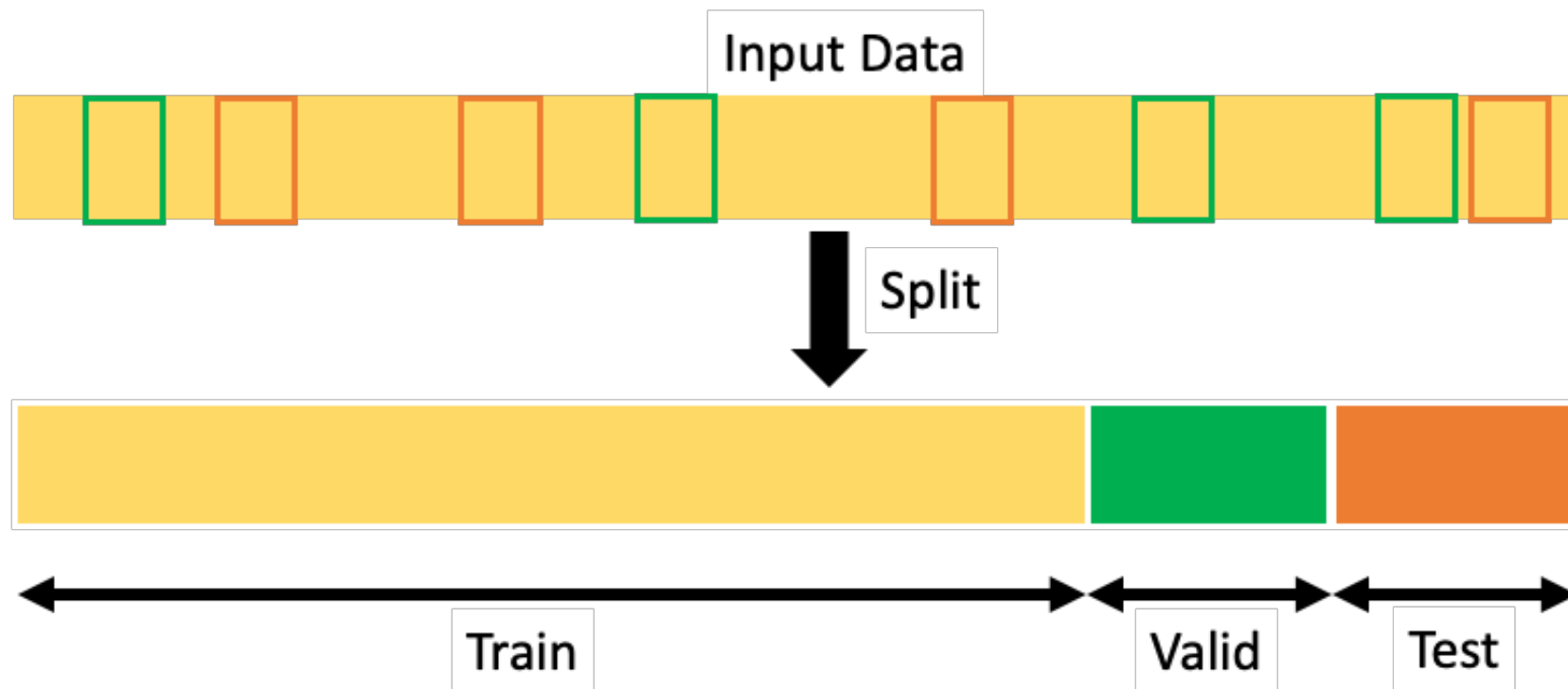
DATA SPLITS

- **Train set**: used to train the model
- **Test set**: used to evaluate the performance of the model after training
 - Tests the model's ability to generalize to new, unseen data
 - Results on the test set gives an indication of how the model will perform with real-world data
 - The test set is **NOT** used during training
 - **Data hygiene** (keeping the training data separate from the test) in the case of large language models (LLM) can be difficult. How to guarantee the separation train/test?

DATA SPLITS

- **Validation set**: used during the tuning process
 - It helps adjust hyperparameters (e.g., learning rate, depth, regularization) **without touching the test set**
 - It must remain independent from the test set to ensure an unbiased final evaluation
 - It is **not** used for training the model's weights
- **Development set**: used like a validation set, but sometimes also for preliminary assessment of the models
 - Sometimes development set and validation set are used interchangeably

DATA SPLITS



From <https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c>

DATA SPLITS

- **Reference**: refers to a set of data used as a benchmark or standard when comparing different models or when validating the outputs of a model
- **Gold standard or gold collection**: dataset that has been meticulously curated and is of highest quality. It often serves as a benchmark

DATA SPLITS

- Data Split
 - Usually: 90% train – 10% test or 80% train – 20% test
 - BigData: 99% train – 1% test
- Validation/development (dev) set:
 - Usually 10% of the training set

DATA SPLITS

- K-fold cross validation
 - Divide the dataset into K equal parts (folds/splits) and run K experiments
 - Each fold is used once as a test set while the remaining K-1 folds form the training set
 - For instance, for $K = 10$
 - divide the corpus in 10 parts (randomly. Why?)
 - train your system with 9 parts
 - evaluate in the remaining one
 - average over the rounds at the end
- Notice that in the deep learning days it might be complicated to train K models

Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

DATA AUGMENTATION

- Some techniques:
 - Synonym replacement: substitutes words in sentences with their **synonyms**
 - **Paraphrasing**: rewrite sentences or paragraphs differently
 - Rule-based augmentation: apply linguistic rules to sentences, such as changing the voice from active to passive
 - Example: The student wrote the report (active voice) vs. The report was written by the student (passive voice)
 - Back translation: translates text to another language and then back to the original language
 - Random swap: randomly swaps the position of words within sentences to create slight variations.

DATA AUGMENTATION

- Some techniques (cont.):
 - Text expansion: enrich the content with additional relevant text, such as explanatory clauses or descriptive phrases
 - Noise injection: introduce typos, spelling mistakes, or grammatical errors to mimic real-world imperfections in text
 - Entity substitution: replace named entities (like names, locations, and organizations) with other entities of the same type

ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



THINK-PAIR-SHARE

Try to match the following sentences with the possible techniques – *Synonym replacement, Paraphrasing (rule-based, back translation, random swap), text expansion, noise injection, entity substitution* – used to generate them from *Princess Mary entered the palace*:

1. Princess Marry entered the palace, the royal residence
2. Into the palace, princess Mary entered
3. Princess Mary walked into the palace
4. Princess Mary entered the place
5. The royal lady Mary made her way into palace
6. Princess Mary has entered the palace
7. Princess Joana entered the palace

Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

DATA CLEANING

- Data cleaning (= denoising) involves the detection and rectification of errors and inconsistencies in datasets

Overview

- Learning objectives
- Topics
 - Corpora: motivation, concept and examples
 - Building corpora
 - Agreement between annotators
 - Data splits
 - Data augmentation
 - Data cleaning
 - Toxic data
- Key takeaways
- Suggested readings

THINK-PAIR-SHARE

How hard can be the work of annotators of toxic datasets?

TOXIC DATA

- Toxicity in NLP: various forms of harmful, offensive, or inappropriate content that can manifest in text
- Examples:
 - Profanity: use of swear/curse words
 - Threats, Insults:
 - Cyberbullying: repeated online behaviours that intimidate or upset individuals

TOXIC DATA

- Examples (cont.):
 - **Misinformation and disinformation**: the spread of false or misleading information, either unintentionally (misinformation) or deliberately (disinformation)
 - **Stereotyping and generalizations**: statements that apply a generalized belief or opinion to all members of a group
 - **Hate speech**: communication that demeans a person or group based on characteristics such as race, religion, ethnic origin, sexual orientation, disability, or gender

ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



THINK-PAIR-SHARE

If I have a dataset that might contain hate speech,
whom should I hire?

Who should decide what is hate speech?

KEY TAKEAWAYS

KEY TAKEAWAYS

- Understand the importance of having corpora for NLP tasks
- Understand the importance of having good annotations
- Concepts associated with corpora, including the ones related with toxic language, and techniques of data augmentation

SUGGESTED READINGS

READINGS

- This slides
- Sebenta:
 - Methodology, corpora and evaluation
 - notice that it does not covers all these slides