

Annotation Basics and Challenges

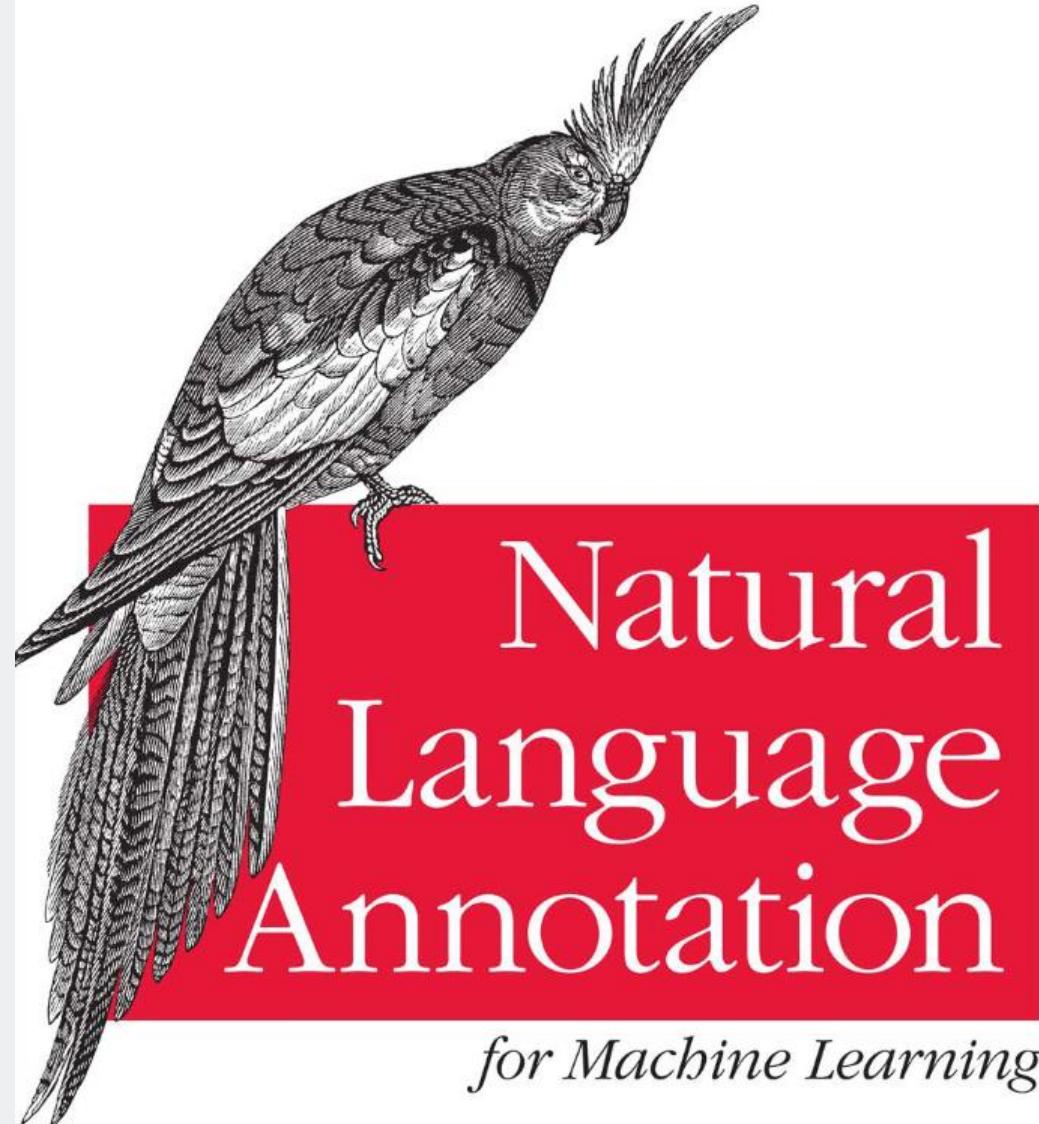
Allan Hanbury

Book

<https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/>

Available online via the TU Wien Library

Also used some slides from the book:
https://sites.google.com/site/brandeisnaml/readings/NLAML_CS216-2013.pdf



Contents

- Introduction
- Selecting an Annotation Task
- MATTER – Model
- MATTER – Annotate
- MATTER – Revise
- Annotation Process Example

Introduction

What is annotation?

- Getting a human to associate a label (metadata) with specific content in a document or file
- Examples
 - Image labeling
 - Spam detection
 - Date and event labeling
 - Document classification

What is annotation for?

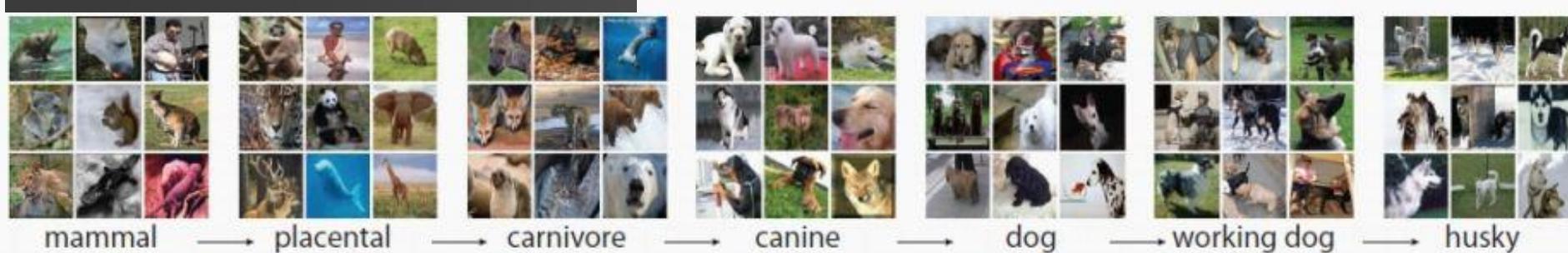
- Create an **annotated corpus**
 - a corpus (plural corpora) or text corpus is a language resource consisting of a large and structured set of texts (Wikipedia)
- Used for:
 - Training and validation data
 - Test data for evaluation

The data that transformed AI research—and possibly the world

In 2006, Fei-Fei Li started ruminating on an idea.

Li, a newly-minted computer science professor at University of Illinois Urbana-Champaign, saw her colleagues across academia and the AI industry hammering away at the same concept: a better algorithm would make better decisions, regardless of the data.

But she realized a limitation to this approach—the best algorithm wouldn't work well if the data it learned from didn't reflect the real world.



- S: (n) [Eskimo dog, husky](#) (breed of heavy-coated Arctic sled dog)
 - [direct hypernym / inherited hypernym / sister term](#)
 - S: (n) [working dog](#) (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
 - S: (n) [dog, domestic dog, Canis familiaris](#) (a member of the genus *Canis* (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
 - S: (n) [canine, canid](#) (any of various fissiped mammals with nonretractile claws and typically long muzzles)
 - S: (n) [carnivore](#) (a terrestrial or aquatic flesh-eating mammal) "terrestrial carnivores have four or five clawed digits on each limb"
 - S: (n) [placental, placental mammal, eutherian, eutherian mammal](#) (mammals having a placenta; all mammals except monotremes and marsupials)
 - S: (n) [mammal, mammalian](#) (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
 - S: (n) [vertebrate, craniate](#) (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
 - S: (n) [chordate](#) (any animal of the phylum Chordata having a notochord or spinal column)
 - S: (n) [animal, animate being, beast, brute, creature, fauna](#) (a living organism characterized by voluntary movement)
 - S: (n) [organism, being](#) (a living thing that has (or can develop) the ability to act or function independently)
 - S: (n) [living thing, animate thing](#) (a living (or once living) entity)
 - S: (n) [whole, unit](#) (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
 - S: (n) [object, physical object](#) (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
 - S: (n) [physical entity](#) (an entity that has physical existence)
 - S: (n) [entity](#) (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

More than 14 million images in ImageNet for around 22000 Synsets (sets of synonyms) (~600 images per Synset)

Jungle gym

A structure of vertical and horizontal rods where children can climb and play

1129 pictures

53.33%
Popularity
Percentile

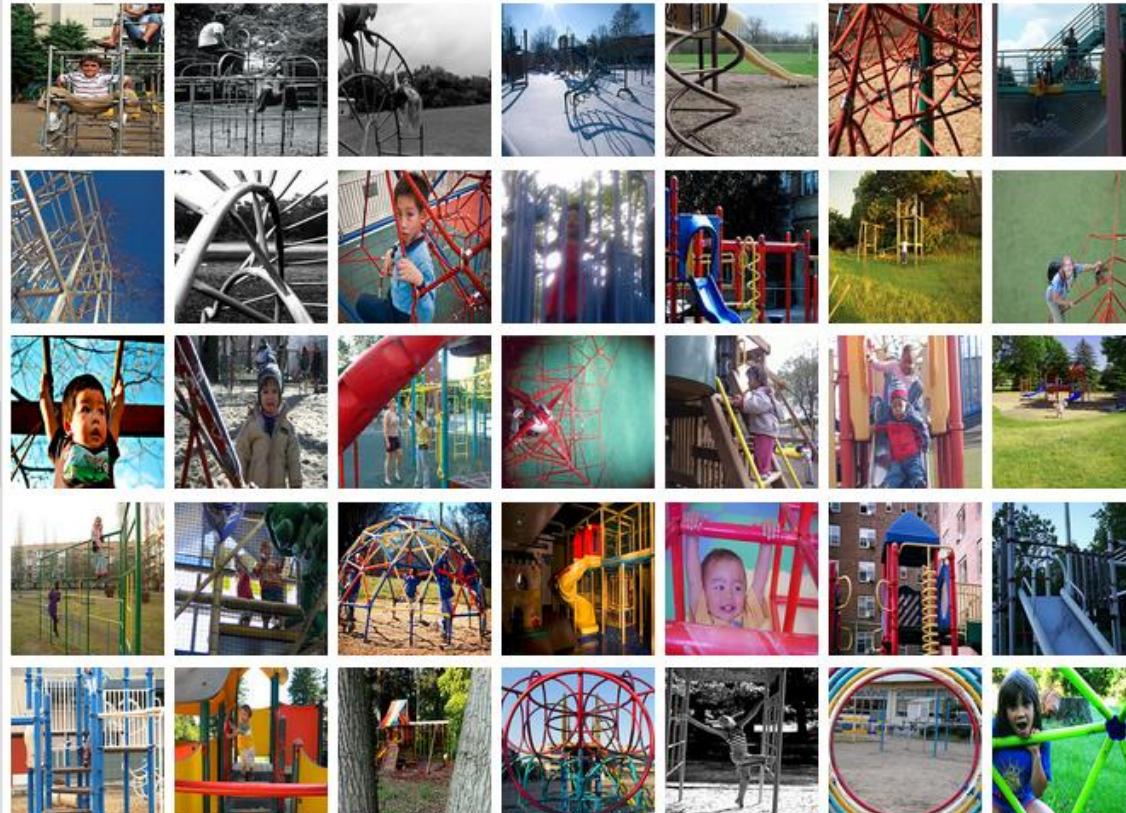


- swing (1)
- Frisbee (0)
- jungle gym (0)
- dollhouse, doll's house (0)
- stick horse (0)
- teddy, teddy bear (0)
- doll, dolly (8)
- Meccano, Meccano set (0)
- water pistol, water gun, squir
- slide, playground slide, slidir
- slingshot, sling, catapult (0)
- yo-yo (0)
- kite (2)
- pogo stick (0)
- balloon (0)
- kaleidoscope (0)
- jack-in-the-box (0)
- seesaw, teeter, teeter-totter, te
- pinata (0)
- top, whirligig, teetotum, spinn
- ball (2)
- cockhorse (0)
- popgun (0)
- train set (0)
- pea shooter (0)
- rattle (0)
- hobby, hobbyhorse, rocking h
- pinwheel, pinwheel wind coll
- hula-hoop (0)
- jumping jack (0)
- Lego, Lego set (0)
- playhouse, wendy house (1) ▾

Treemap Visualization

Images of the Synset

Downloads



*Images of children synsets are not included. All images shown are thumbnails. Images may be subject to copyright.

Prev 1 2 3 4 5 6 7 8 9 10 ... 32 33 Next

Bean sprout

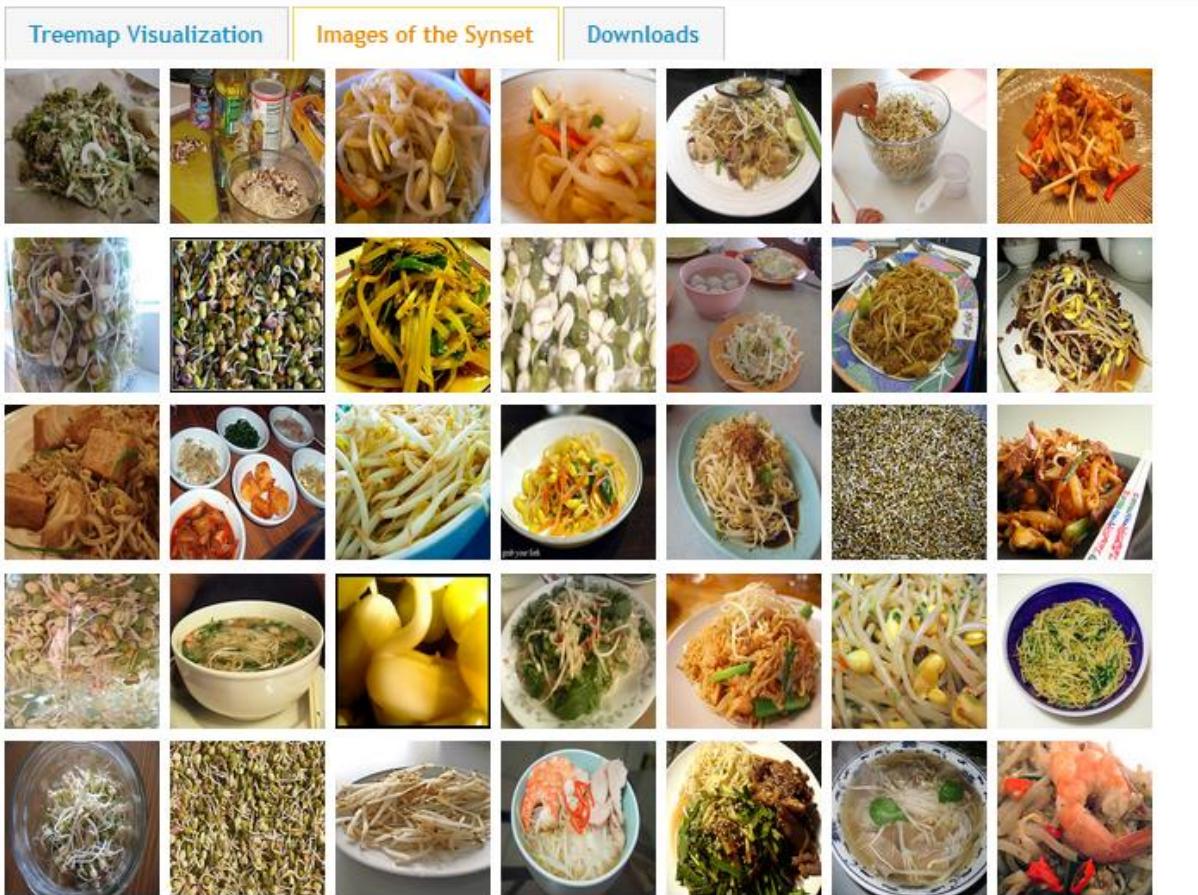
Any of various sprouted beans: especially mung beans or lentils or edible soybeans

1135 pictures

65.06% Popularity Percentile

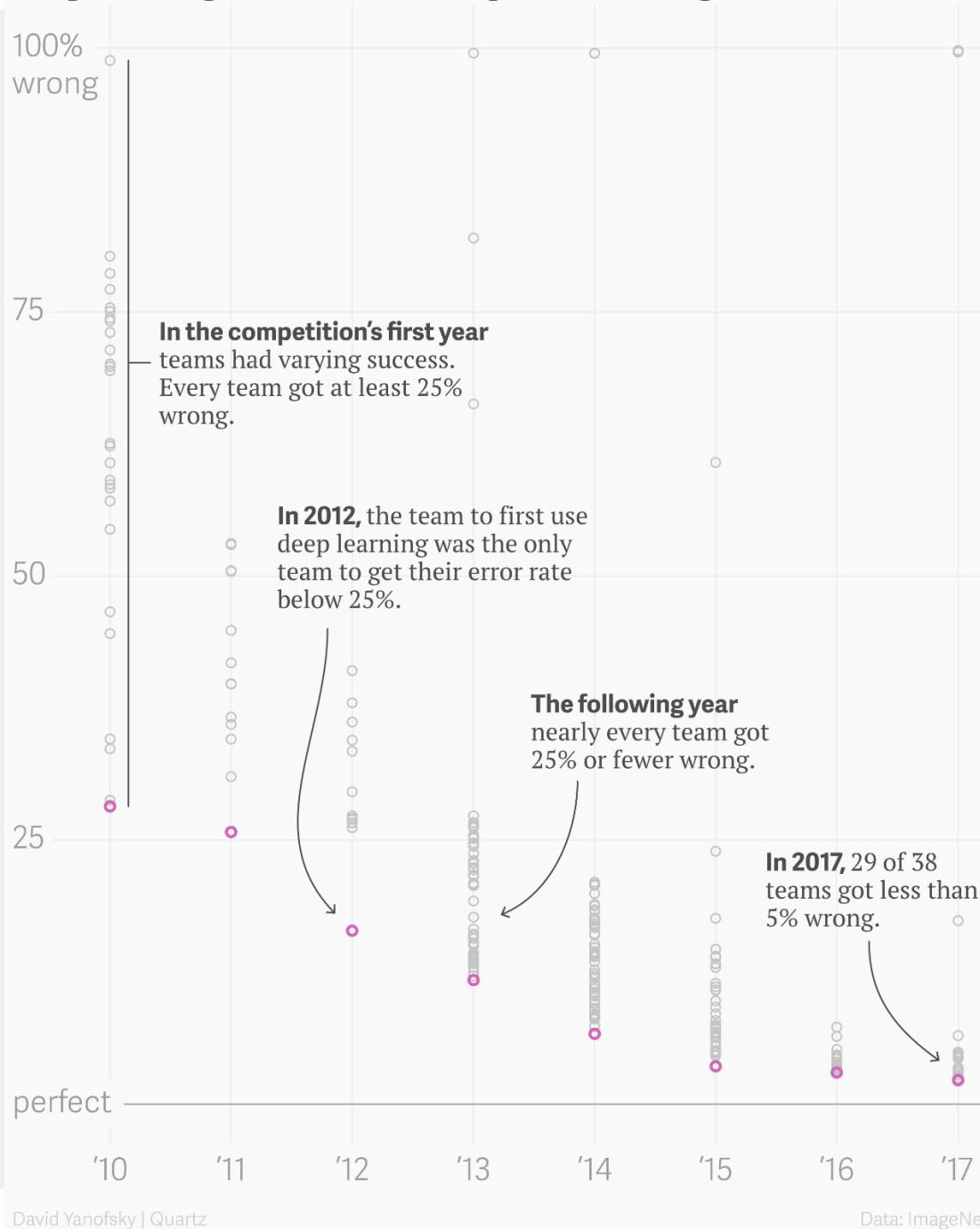
Wordnet IDs

- wax bean, yellow bean (0)
- Fordhooks (0)
- lima bean (1)
- sieva bean, butter bean, butter bean (0)
- fava bean, broad bean (0)
- green soybean (0)
- shell bean (5)
- fresh bean (13)
- flageolet, haricot (0)
- common bean (20)
- field soybean (0)
- soy, soybean, soya, soya bean (0)
- bean, edible bean (24)
- lentil (0)
- snow pea, sugar pea (0)
- sugar snap pea (0)
- split-pea (0)
- green pea, garden pea (3)
- marrowfat pea (0)
- cajan pea, pigeon pea, dahl (0)
- field pea (0)
- pea (7)
- chickpea, garbanzo (0)
- black-eyed pea, cowpea (0)
- legume (37)
- potherb (0)
- chop-suey greens (0)
- **bean sprout (0)**
- alfalfa sprout (0)
- sprout (2)
- beet green (0)
- chard, Swiss chard, spinach beet (0)
- buttercrunch (0)



Prev 1 2 3 4 5 6 7 8 9 10 ... 32 33 Next

ImageNet Large Scale Visual Recognition Challenge results



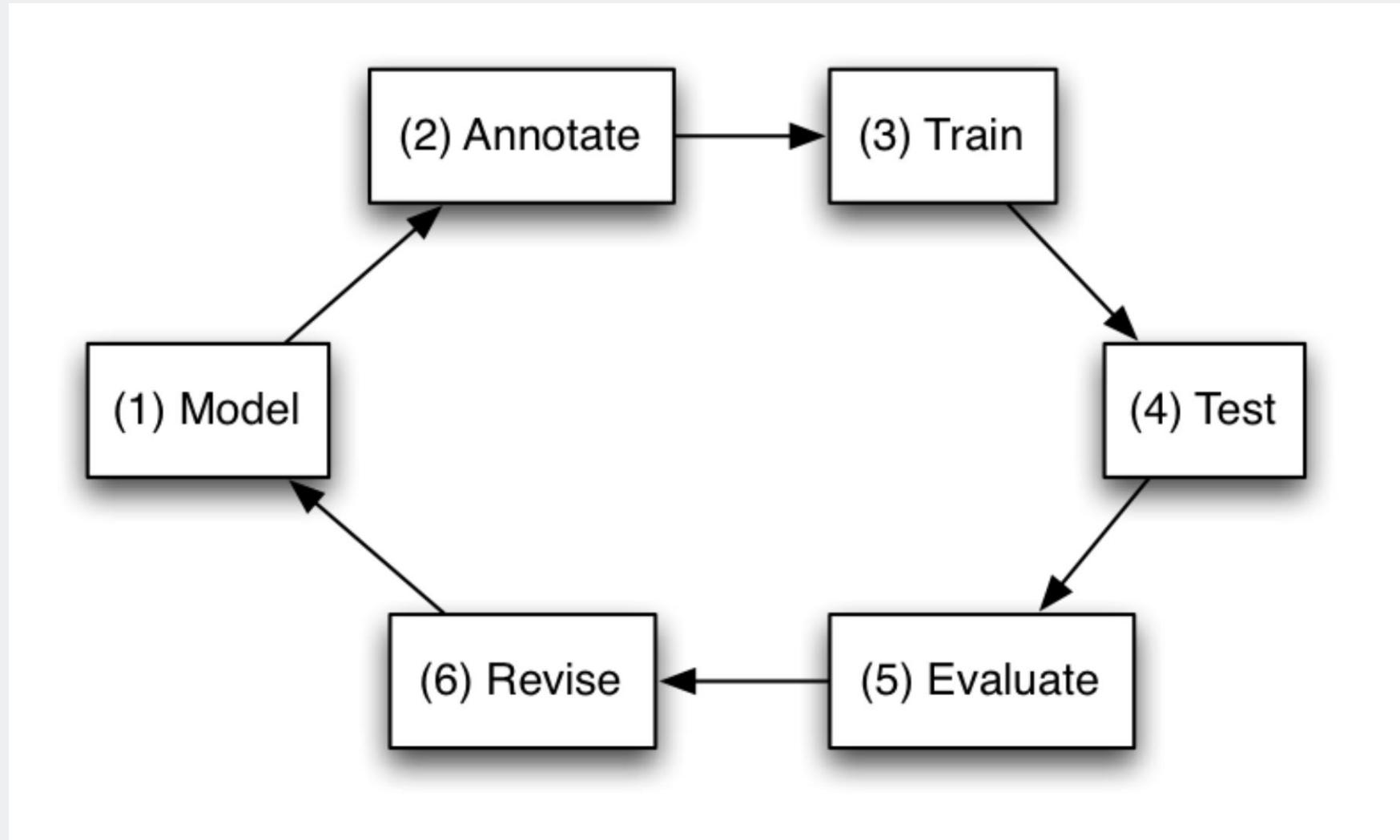
Aspects of Text Annotation

- Structure of language (word classes, syntax)
- Meaning of words (word sense disambiguation)
- Interpretation of meaning (semantics)
- Document structure (discourse)
- Speech sounds/parts of words (phonetics, morphology)

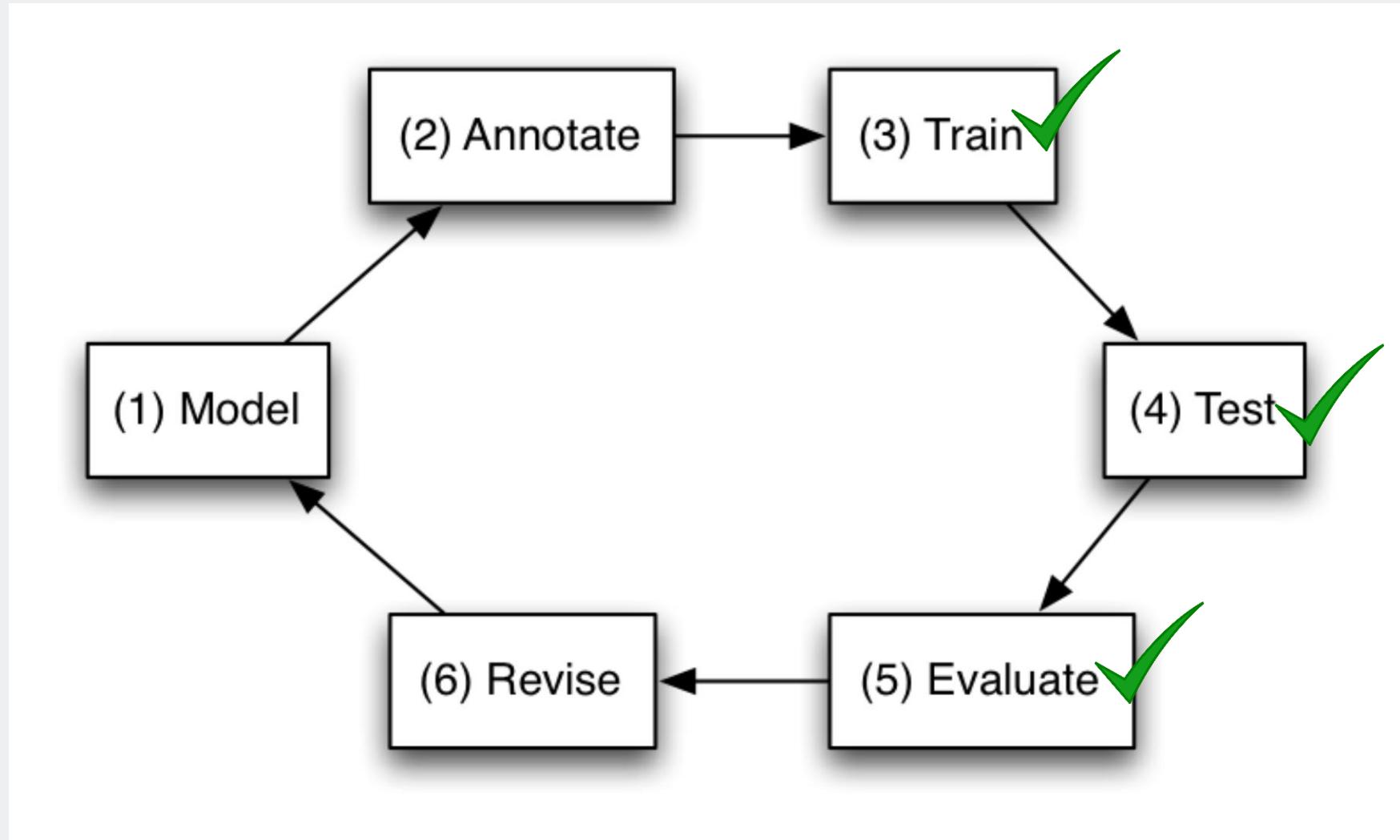
Creating Annotations

- Interpreting language isn't simple; people don't always agree on meaning
- An annotated corpus needs to be consistent to train/test an algorithm
- Today we'll focus on the **best practice** for creating an annotated corpus

The MATTER Cycle



The MATTER Cycle



Selecting an Annotation Task

Choosing a Goal

- Annotation tasks need to be focused on a particular goal
- Goals can take lots of different forms
- Questions to answer:
 - What will the annotation be used for?
 - What will the overall outcome of the annotation be?
 - Where will the corpus come from?
 - How will the outcome be achieved?

Outcome of the Annotation

- What will the annotation be used for?
 - database population,
 - linguistic analysis,
 - summarization,
 - timeline creation,
 - ...
- How in-depth will your classification be?

Where will the Corpus Come From?

- Needs to be representative and balanced as required by your goal
 - Representative means that the corpus contains all possible types of text for the area under consideration (related to sampling)
 - Balanced means that there are a sufficient number of examples of each of the types of text
- Size of the corpus – trade-off between utility and practicality
- Look at existing corpora

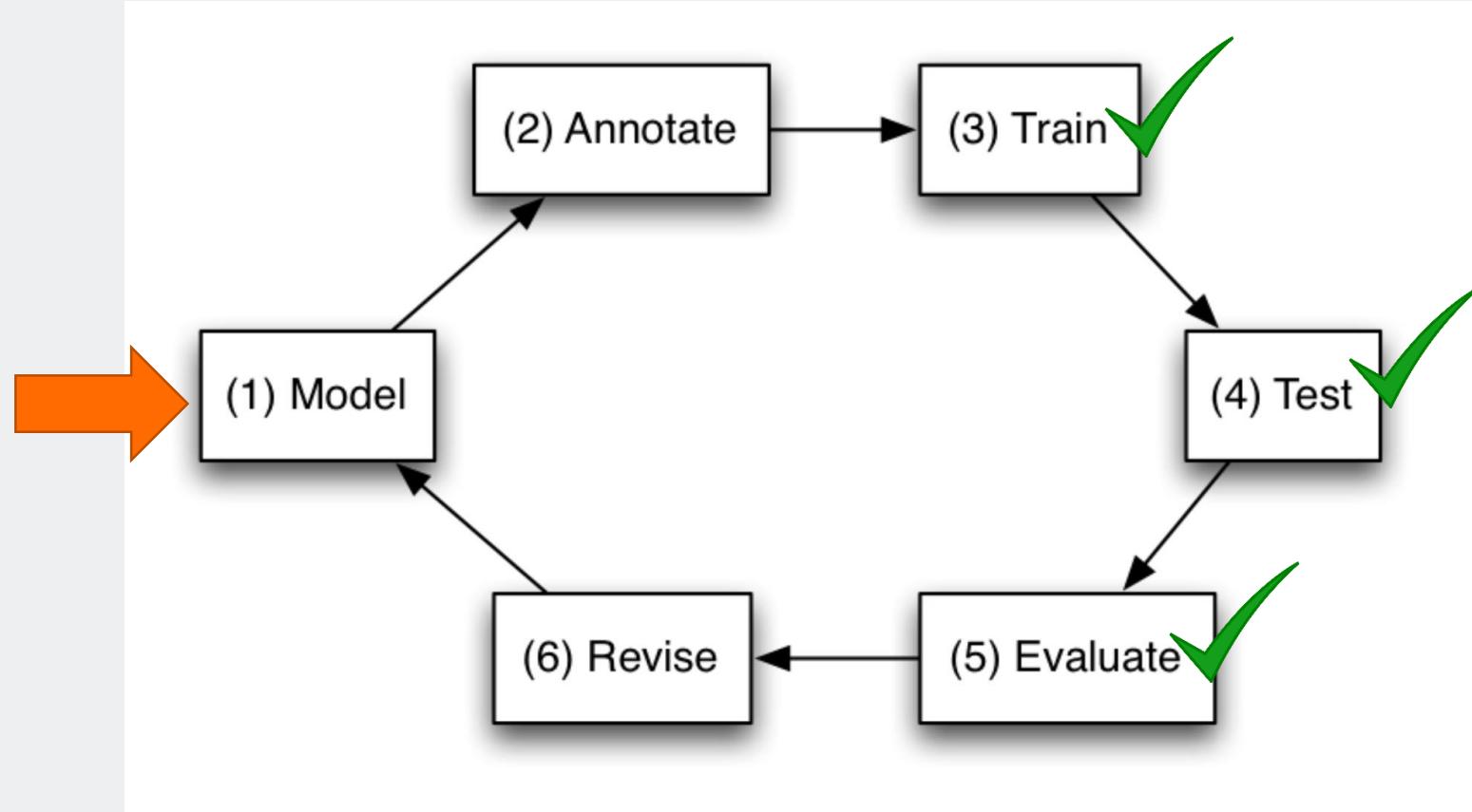
Achieving the Goal

- What aspects of the text will help you with your classification?
 - All the text in the document?
 - Individual words?
 - Relations between words? Between documents?

Things to Consider

- Scope of the task:
 - Styles/sources of text
 - Level of detail
- Purpose of the annotation
- Resources (money, people) available for the annotation

MATTER – Model



Model

- How will the annotation be represented?
- What exactly will be annotated?
 - Whole document
 - Text sections/entities
 - Relations between entities
- What will be used as the annotation vocabulary?

Example: Film Summaries – Genre Classification

- Goal: Use the film summary to determine the genre of the film being described
- Whole document annotation – one genre (or multiple genres) per document
- Vocabulary:
 - **IMDB:** Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film Noir, Game-Show, History, Horror, Music, Musical, Mystery, News, Reality-TV, Romance, Sci-Fi, Sport, Talk-Show, Thriller, War, Western
 - **Netflix (top level):** Action & adventure, Classic Movies, Comedies, Dramas, Horror movies, Romantic movies, Sci-Fi & Fantasy, Sports movies, Thrillers, Documentaries, TV Show, Teen TV shows, Children & family movies, Anime, Independent movies, Foreign movies, Music, Christmas, Others

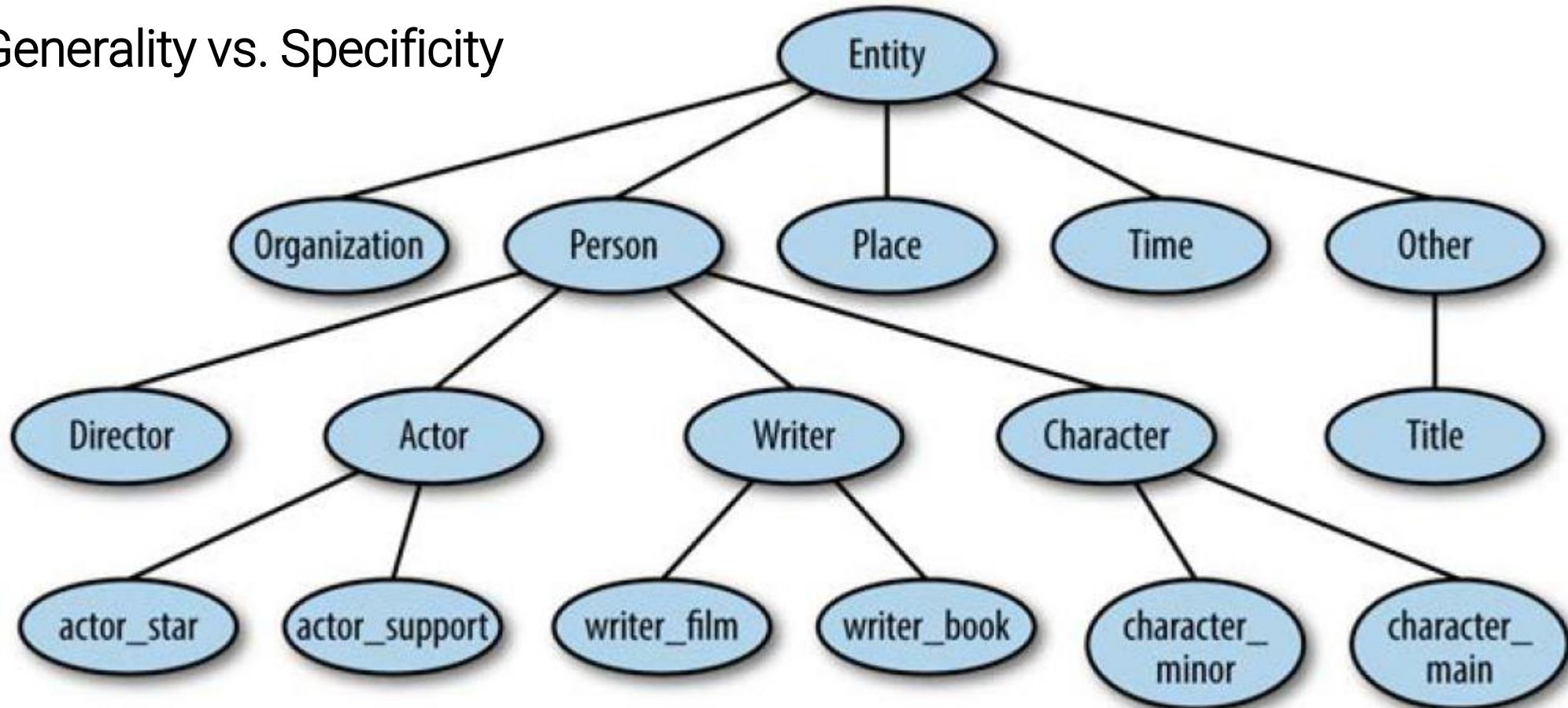
Example: Film Summaries – Named Entities

- Goal: Identify the Named Entities in the film summaries
- Associate text sections with entity types
- Vocabulary: film_title, director, writer, actor, character, ...
- Annotation example:

“... directed by <director>Steven Spielberg</director>, and starring <actor>John Cleese</actor>, <actor>Graeme Garden</actor> ...”

Example: Film Summaries – Named Entities

Generality vs. Specificity



Example: Film Summaries – Semantic Roles

- Goal: How are the Named Entities related to each other?
- Link entities with relations
- Relations: acts_in, acts_as, directs, writes, character_in, ...

Example: Film Summaries – Semantic Roles

“In <film_title>Love, Actually</film_title>, writer/director <writer><director>Richard Curtis</director></writer> weaves a convoluted tale about characters and their relationships. Of particular note is <actor>Liam Neeson</actor> (<film_title>Schindler’s List</film_title>, <film_title>Star Wars</film_title>) as <character>Daniel</character> ...”

The diagram illustrates the semantic roles and relationships between entities in the sentence. It uses colored arrows to map words to their corresponding entities and roles:

- The word "Love, Actually" is mapped to the entity "<film_title>Love, Actually</film_title>" (purple).
- The word "writer/director" is mapped to the entity "<writer><director>Richard Curtis</director></writer>" (red and orange).
- The word "weaves" is mapped to the entity "<actor>Liam Neeson</actor>" (green).
- The word "characters" is mapped to the entity "<character>Daniel</character>" (blue).

Relationships are indicated by arrows:

- A red arrow labeled "directs" points from "writer/director" to "<film_title>Love, Actually</film_title>".
- An orange arrow labeled "writes" points from "writer/director" to "<actor>Liam Neeson</actor>".
- A green arrow labeled "acts_in" points from "<actor>Liam Neeson</actor>" to "<film_title>Love, Actually</film_title>".
- A green arrow labeled "acts_in" points from "<actor>Liam Neeson</actor>" to "<film_title>Star Wars</film_title>".
- A blue arrow labeled "character_in" points from "<character>Daniel</character>" to "<film_title>Star Wars</film_title>".
- A blue arrow labeled "acts_as" points from "<character>Daniel</character>" to "<character>Daniel</character> ...".

Another Example: Time Stamping Events

April 25, 2010

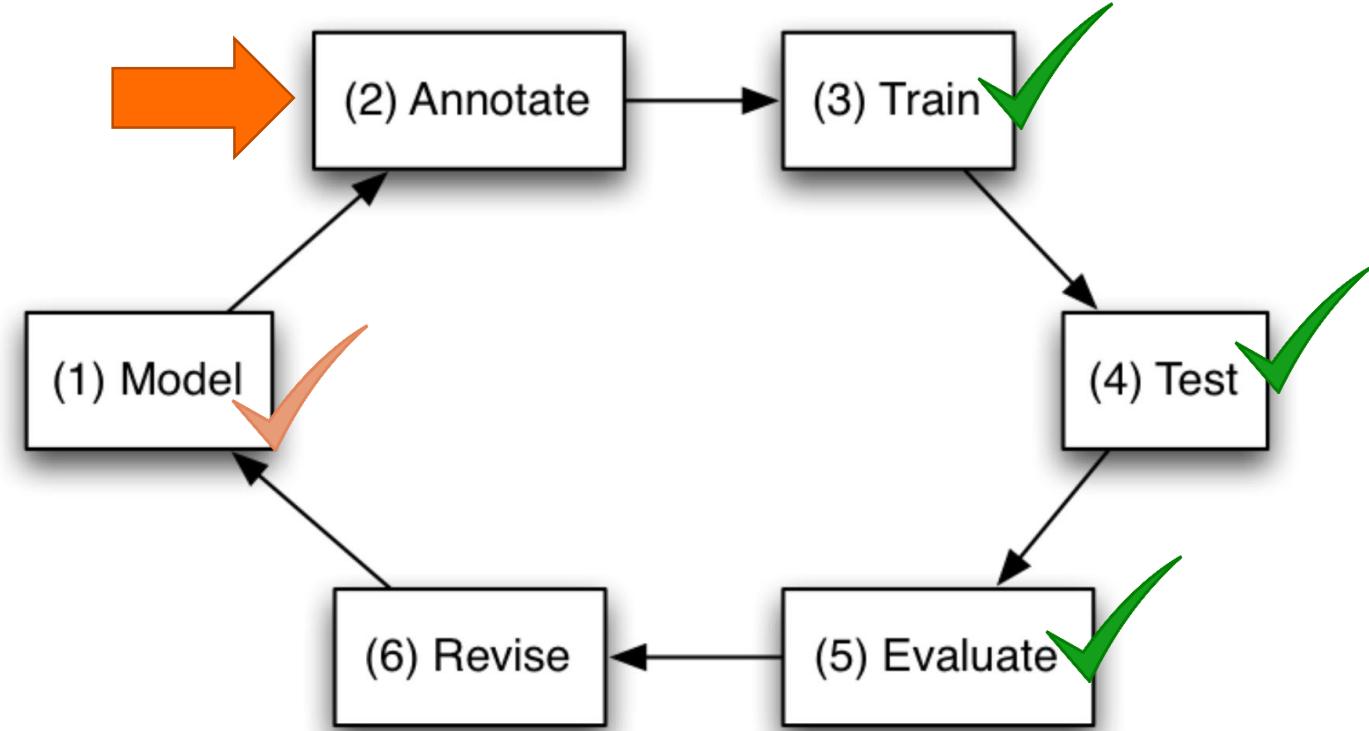
- President Obama paid tribute Sunday to 29 workers killed in an explosion at a West Virginia coal mine earlier this month, saying they died "in pursuit of the American dream." The blast at the Upper Big Branch Mine was the worst U.S. mine disaster in nearly 40 years. Obama ordered a review earlier this month and blamed mine officials for lax regulation.

Another Example: Temporal Ordering of Events

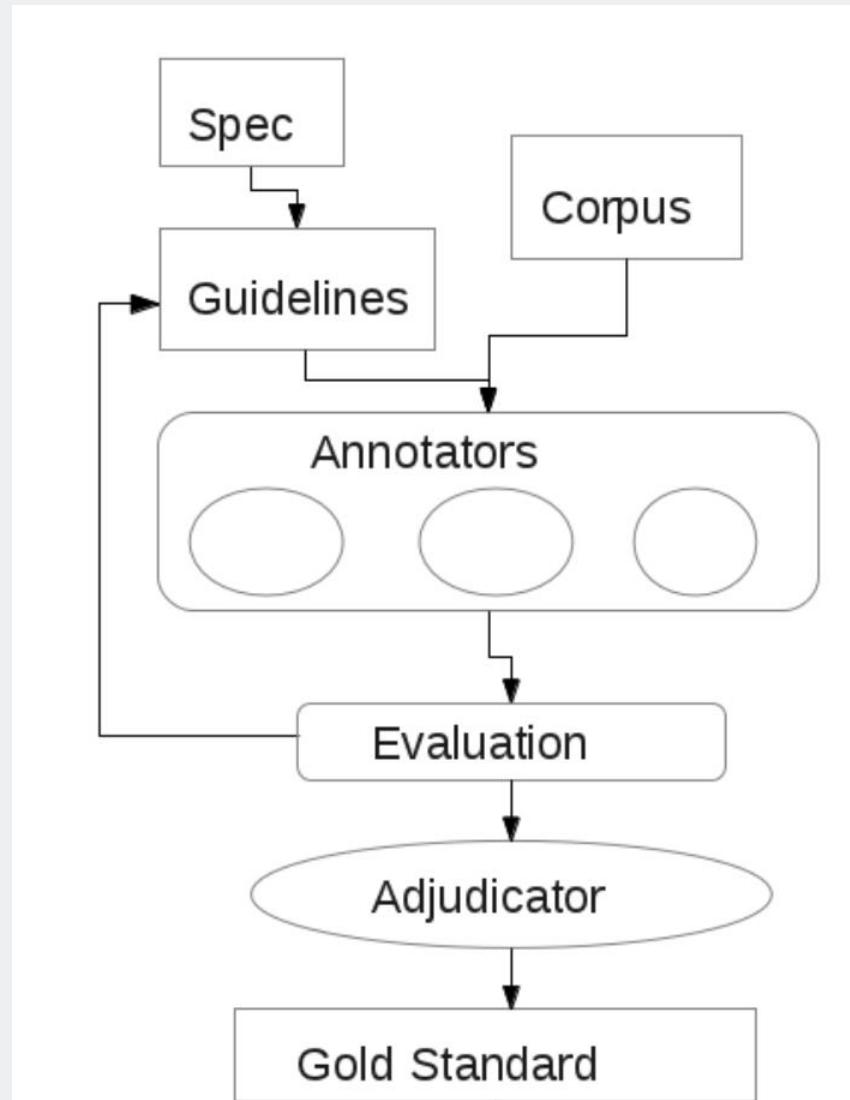
April 25, 2010

- President Obama **paid tribute** **Sunday** to 29 workers **killed** in an **explosion** at a West Virginia coal mine **earlier** this month, **saying** they **died** "in pursuit of the American dream." The **blast** at the Upper Big Branch Mine was the worst U.S. mine **disaster** in nearly 40 years. Obama **ordered** a **review** **earlier** this month and **blamed** mine officials for lax regulation.

MATTER – Annotate



Annotation Process



Specification

- Which model will be used to represent the annotations?
- How many annotations are necessary? What percentage of the annotations will be done by multiple annotators?
- Should the annotation task be broken down into sub-tasks?
- Are the resources available to create this number of annotations?
- Will quality control be done? If so, how?

Choosing Annotators

- What expertise do the annotators need?
 - Domain-expertise for a specific field
 - Linguistics
 - Both?
- How will you compensate the annotators?
 - Payment. How much?
 - Other incentive
- How are the annotators recruited?
 - Advertisement
 - Contacts
 - Annotation company
 - Crowd-sourcing
 - ...

Annotation Guidelines

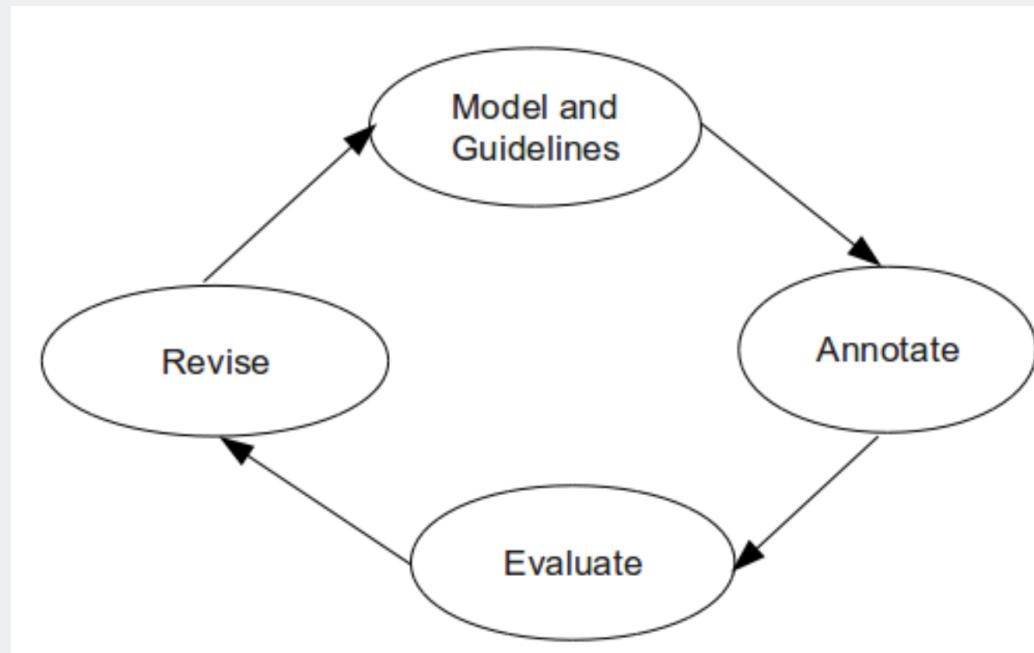
- Directions for the annotators
 - What, where, why, how
- Reusable
- Different from specification
 - Focus on how, why
- Balance between length and level of detail

Guideline Considerations

- What is the goal of the project?
- What is each tag called and how is it used?
- What parts of the text do you want annotated, and what should be left alone?
- How will the annotation be created?
 - order of the annotation of concepts
 - how the annotation tool works
 - ...

MAMA

- Model/Guidelines – Annotate – Model/Guidelines – Annotate – ...



- For the first cycles, use few documents and annotators

Writing the Annotation Guidelines

- Give guidelines to cover all eventualities
- Example: Annotation of sentiment of film reviews as positive or negative
- What to do about a review like this?
 - “This movie was all right. The special effects were good, but the plot didn’t make a lot of sense. The actors were funny, which helped, but the music was really distracting.”
- Cover this in the guidelines...

Guidelines for Movie Review Sentiment Labelling

What is the goal of the project?

To label movie reviews as being positive or negative.

What is each tag called and how is it used?

We have two labels, “positive” and “negative,” and each review will be labeled with one of them, based on the tone of the review. **Reviews that are not specifically positive or negative will be labeled as “negative.”**

What parts of the text do you want annotated, and what should be left alone?

Each review will be given a single label, which will be applied to the entire document.

How should the annotation be created?

[Describe the annotation software]

Guidelines for Movie Review Genre Labelling

What is the goal of the project?

To label film summaries with genre notations.

What is each tag called and how is it used?

We have 26 tags that can be applied to each summary as needed.

What parts of the text do you want annotated, and what should be left alone?

Each label will apply to the entire document.

How should the annotation be created?

[Describe the annotation software]



Kindle Customer

★★★★★ **This was a Disney *failure*?!!?**

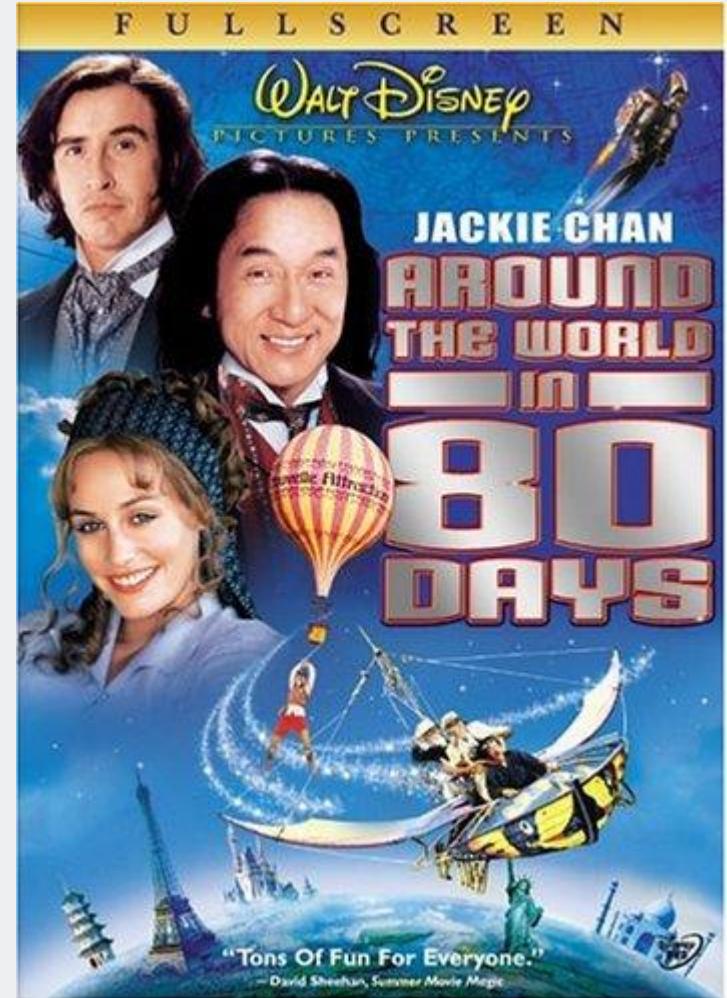
Rezension aus den Vereinigten Staaten vom 21. Juli 2018

Verifizierter Kauf

I loved this! I've been a great fan of Disney movies my entire life and I can't believe this lost money. It has all the requirements. It's not exactly like the book because it *isn't* the book. I don't recall the book having life sized adventure, excitement, spectacular scenery and learning opportunities. I think the comedy is a plus, the romance is kid friendly. In addition, who doesn't cheer for Jackie Chan's martial arts with good guy ninjas? Failure? Please.

From the Guidelines: "We have 26 tags that can be applied to each summary as needed."

Tags: Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film Noir, Game-Show, History, Horror, Music, Musical, Mystery, News, Reality-TV, Romance, Sci-Fi, Sport, Talk-Show, Thriller, War, Western



- Guideline improvements:
 - List how many labels may be applied to each review
 - How similarly do annotators interpret the genres? → Add definition of each genre
 - The genre labels describe different aspects of the film
 - Events in the film: Action, Adventure, Crime, Romance
 - Setting: Historical, Sci-Fi, Fantasy
 - Production circumstances: Animation, Talk-Show, Reality-TV→ How can this be handled?
 - If the annotator knows the movie, can he/she use this knowledge in selecting the labels, or must it be based solely on the text? → State this in the Guidelines

Example Guidelines: Sentence Annotation in Legal Text (1)

Es folgt eine Auflistung der verfügbaren Merkmale, aufgegliedert in die folgenden Kategorien:

- Volumen (Abschn. 2.1)
- Fläche (Abschn. 2.2)
- Höhe (Abschn. 2.3)
- Stellplätze, Garagen, Parkgebäude (Abschn. 2.4)
- Geschosse (Abschn. 2.5)
- Vorbauten (Abschn. 2.6)
- Einfriedungen (Abschn. 2.7)
- Nutzung und Widmung (Abschn. 2.8)
- Grossbauvorhaben, Hochhäuser, Einkaufszentren, Geschäftsstrassen (Abschn. 2.9)
- Laubengänge, Durchfahrten, Arkaden (Abschn. 2.10)
- Ausgestaltung und Sonstiges (Abschn. 2.11)
- Dach (Abschn. 2.12)
- Strasse und Gehsteige (Abschn. 2.13)
- Lage, Gelände und Planzeichen (Abschn. 2.14)
- Meta (Abschn. 2.15)

Example Guidelines: Sentence Annotation in Legal Text (2)

2.12.8 AnteilDachbegruenung

Erläuterung :

Spezifiziert den minimalen Anteil der begrünten Dachfläche relativ zur gesamten Dachfläche.

Wert :

Eine positive Zahl [%]

Beispiele :

„Auf den als Gemischtes Baugebiet/Betriebsbaugebiet und mit Dächer von Gebäuden [...] im Ausmaß von mindestens 40% ents zu begrünen.“
(7443_10)

2.2 Kategorie: Fläche

Enthält Merkmale, welche eine Fläche betreffen. Siehe auch Abschnitt 2.9 für Merkmale bezüglich der Fläche speziell bei Grossbauvorhaben, Einkaufszentren, etc.

2.2.1 BBBebaubareFlaecheGesamterBauplatz

Erläuterung :

Beschränkt die bebaubare Fläche auf dem Bauplatz relativ zur Fläche des gesamten Bauplatzes.

Wert :

Eine positive Zahl [%]

Rechtsmaterie :

WBO §82/5 :

(5) Die durch Nebengebäude in Anspruch genommene Grundfläche ist auf die nach den gesetzlichen Ausnutzbarkeitsbestimmungen bebaubare Fläche und die die nach § 5 Abs. 4 lit. d durch den Bebauungsplan beschränkte bebaubare Fläche des Bauplatzes anzurechnen. Im Gartensiedlungsgebiet ist die mit einem Nebengebäude bebaute Grundfläche auf die Ausnutzbarkeitsbestimmungen eines Bauloses dann anzurechnen, wenn die bebaubare Fläche im Bebauungsplan mit mindestens 100 m² festgesetzt ist.

Guidelines for Named Entity Annotation

- How should the annotators decide how long each tagged span should be?
- How should complex cases of people's names be handled?
E.g. "Stefan and Anna Warchalowski"
- Should every mention of a Named Entity be annotated or only the first mention?
- What about an entity playing two different roles?
E.g. "Boston City Hall"
- How are possessive constructions handled?
E.g. "John Hughes' *The Breakfast Club*"
- ...

...
Mechanism of insulin release in normal pancreatic beta cells - insulin production is more or less constant within the beta cells. Its release is triggered by food, chiefly food containing absorbable glucose.

Insulin is the principal hormone that regulates uptake of glucose from the blood into most cells (primarily muscle and fat cells, but not central nervous system cells). Therefore, deficiency of insulin or the insensitivity of its receptors plays a central role in all forms of diabetes mellitus.

Humans are capable of digesting some carbohydrates, in particular those most common in food; starch, and some disaccharides such as sucrose, are converted within a few hours to simpler forms, most notably the monosaccharide glucose, the principal carbohydrate energy source used by the body. The rest are passed on for processing by gut flora largely in the colon. Insulin is released into the blood by beta cells (β -cells), found in the islets of Langerhans in the pancreas, in response to rising levels of blood glucose, typically after eating. Insulin is used by about two-thirds of the body's cells to absorb glucose from the blood for use as fuel, for conversion to other needed molecules, or for storage.

Insulin is also the principal control signal for conversion of glucose to glycogen for internal storage in liver and muscle cells. Lowered glucose levels result both in the

Choosing an Annotation Environment

- Standalone or online?
- Existing tool or write your own?
- Questions specific to the annotation task:
 - Does the learning curve of the tool match the capabilities of the annotators?
 - What parts of the task must the annotation software support?
 - What are the units of the annotation task? (characters, entities, documents, ...)
 - Should the annotation task be divided into parallel/sequential sub-tasks?
 - What other software features are necessary? (annotator management, POS taggers, quality control, ...)

Example: INCEption

Active Learning

Session

Layer Named entity

Text Illinois

Label LOC

Score 1

Delta 1

Accept **Reject** **Skip**

Learning History

- Berkeley http://www.wikidata.org/entity/Q464678 skipped
- Berkeley http://www.wikidata.org/entity/Q168756 skipped
- Tesla PER accepted
- Science OTH rejected
- Tesla PER accepted

Annotation

1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017.

2 The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.

3 He served in the Illinois State Senate from 1997 until 2004.

Illinois Senate

upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States

Layer Surface form

Annotation

Layer Named entity

Text Illinois

identifier illi

value Illinois

- Illinois Senate
- Illinois River
- Governor of Illinois
- Alton
- Illinois Country
- Illinois Territory

Example: Annotation for few concepts

MAP

A more significant decline in performance is detected between runs 706 and 707 (switching from automatic filtering of the RequestText field to a manual edit), at the .025 level. In our training, run 605 achieved the best results for R-Prec and MAP, measures which were used to differing degrees for TREC-2006. This result was duplicated in the runs for TREC-2007.

Challege

Collection

Garbage

Measure

None

Run

Last 100 garbaged words

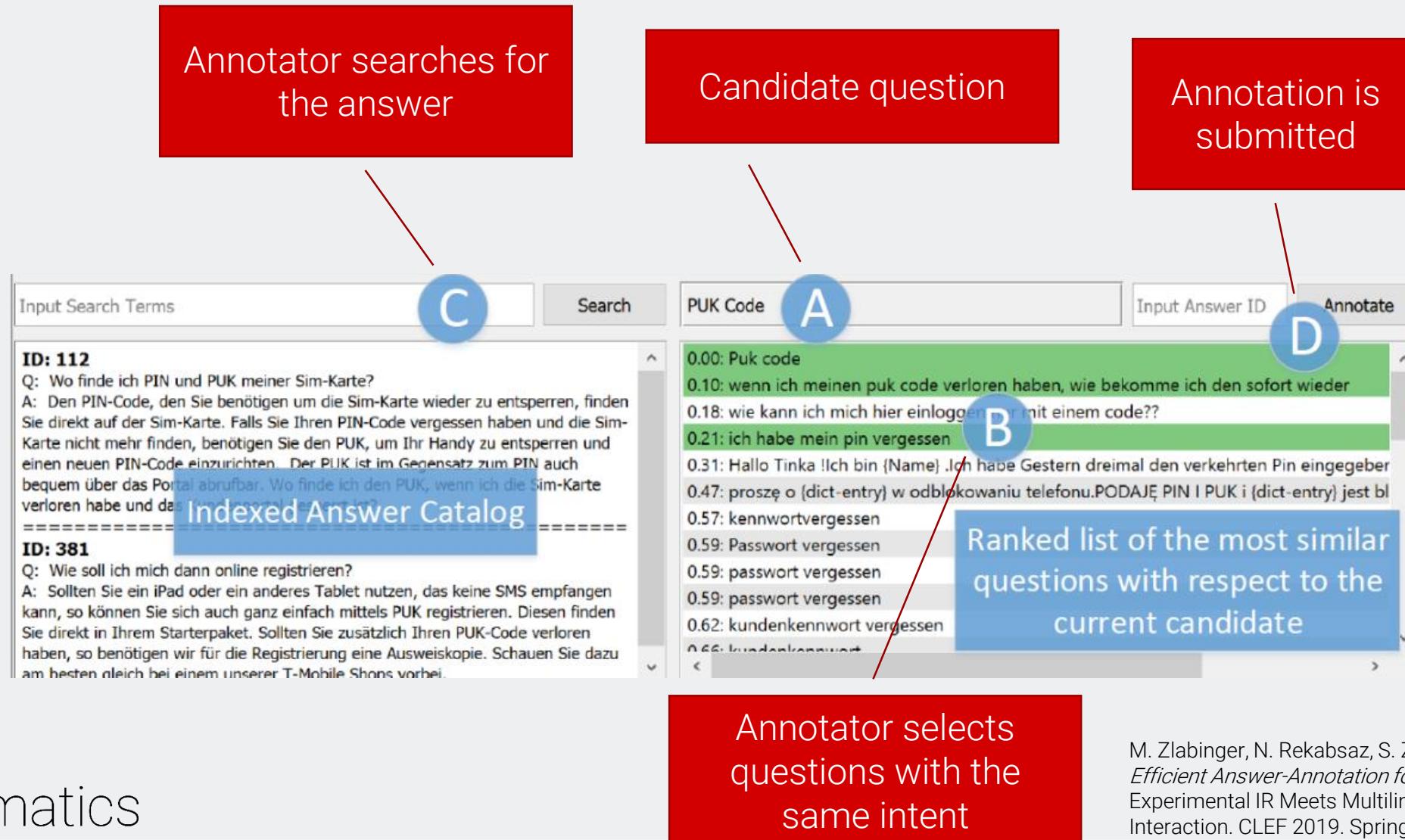
where current absolute prepared linking statistic missing

Skip

Task

Test Collection

Example: Question-Answering Interface



Example: Marking relevant text of relevant document extracts

what is durable medical equipment consist of :

If you can't use a cane or walker safely, but you have enough upper body strength or have someone who's available to help, you may qualify for a manual wheelchair. The most appropriate manual wheelchair for you may have to be rented first, even if you eventually plan to buy it. 2 Types of equipment (continued)What is the reimbursement criteria for Scooters? Power-operated vehicle/scooter If you can't use a cane or walker, or can't operate a manual wheelchair, you may qualify for a power-operated scooter, if you can safely get in and out of it and are strong enough to sit up and safely operate the controls. Note: If you don't need a scooter on a long-term basis, you can rent the equipment to lower your costs.

1 Wrong 2 Topic 3 Partial 4 Perfect

what is durable medical equipment consist of :

If you can't use a cane or walker safely, but you have enough upper body strength or have someone who's available to help, you may qualify for a manual wheelchair. The most appropriate manual wheelchair for you may have to be rented first, even if you eventually plan to buy it. 2 Types of equipment (continued)What is the reimbursement criteria for Scooters? Power-operated vehicle/scooter If you can't use a cane or walker, or can't operate a manual wheelchair, you may qualify for a power-operated scooter, if you can safely get in and out of it and are strong enough to sit up and safely operate the controls. Note: If you don't need a scooter on a long-term basis, you can rent the equipment to lower your costs.

1 Wrong 2 Topic 3 Partial 4 Perfect

Example: Sentence Annotation in Excel

7020.xlsx [Read-Only] - Excel

7020.xlsx [Read-Only] - Excel				
File	Home	Insert	Page Layout	Formulas
Paste	Cut	Copy	Format Painter	Clipboard
D5	A	B	C	D
1	Sentence	AttributeType	Attribute	AttributeType
2	Die roten Planzeichen gelten als neu festgesetzt.			
3	Für die rechtliche Bedeutung der roten Planzeichen ist die beiliegende „Zeichenerklärung für den Flächenwidmungsplan und den Bebauungsplan“ (§§ 4 und 5 der Bauordnung für Wien) vom September 1996 maßgebend, die einen Bestandteil dieses Beschlusses bildet.			
4	Für die Querschnitte der Verkehrsflächen gemäß § 5 Abs. 2 lit. c der BO für Wien wird bestimmt, daß bei einer Straßenbreite unter 10,0 m entlang der Fluchlinien Gehsteige mit mindestens 0,8 m Breite, bei einer Straßenbreite von 10,0 m bis unter 16,0 m entlang der Fluchlinien Gehsteige mit mindestens 1,5 m Breite und bei einer Straßenbreite ab 16,0 m entlang der Fluchlinien Gehsteige mit mindestens 2,0 m Breite herzustellen sind.	Strassen_und_Gehsteige	GehsteigbreiteMin	Volumen Flaeche Hoehe Stellplaetze_Garagen_Parkgebäude Geschosse Vorbauten Einfriedungen Nutzung_Widmung
5	In der Grinzingier Straße, Perntergasse, Heiligenstädter Straße, Gallmeyergasse und in der Hohe Warte zwischen Gallmeyergasse und Geweygasse sind Vorkehrungen zu treffen, daß das Pflanzen von mindestens einer Baumreihe möglich ist.			
6	Bestimmungen gemäß §4 Abs. 3 der BO für Wien: Seite - 2 -			
7	Für die mit BB 3 bezeichneten Flächen werden gesonderte Widmungen für zwei übereinander liegende Räume derart getroffen, daß der bis zur Deckenoberkante der unterirdischen Objekte reichende Raum dem Bauland/Wohngebiet und der Raum darüber dem Grünland/Erholungsgebiet Parkanlage, Grundfläche für öffentliche Zwecke zugeordnet wird.			
8	Die Grenze der Widmungen verläuft im Niveau von +13 m über Wiener Null.			
9				

Evaluating the Annotations

- Inter-Annotator Agreement (IAA) – measures:
 - How good are the annotators?
 - How good are the Guidelines?
 - How reproducible is the annotation process?
- IAA indicates what could be improved in the MAMA cycle
- Commonly used scores:
 - Cohen's Kappa (two annotators)
 - Fleiss's Kappa (more than two annotators)

Cohen's Kappa

- $\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$
- $\Pr(a)$ is the relative observed agreement between the annotators
- $\Pr(e)$ is the expected agreement between the annotators, if each annotator were to randomly pick a category for each annotation

Cohen's Kappa: Calculation Example (1)

- Two annotators, A and B, assigned the labels “positive,” “neutral,” and “negative” to a set of 250 movie reviews

$$\bullet \Pr(a) = \frac{(54+18+72)}{250} = 0.576$$

- A and B chose the positive label 85 times, $\frac{85}{250} = 0.34$. The probability that A and B chose “positive” together by chance is $0.34 \times 0.34 = 0.116$

- Repeat for the “neutral” and “negative” classes

$$\bullet \Pr(e) = 0.116 + 0.077 + 0.146 = 0.339$$

$$\bullet \kappa = \frac{(0.576 - 0.339)}{(1 - 0.339)} = 0.359$$

	B positive	B neutral	B negative
A positive	54	28	3
A neutral	31	18	23
A negative	0	21	72

Cohen's Kappa: Calculation Example (2)

- $\kappa = 0.359$
- Can be interpreted according to the table on the right, but often useful to look at results for similar annotation tasks
- Further interpretation:
 - Confusion over how to spot a neutral review → update that part of the Guidelines
 - B labelled some reviews negative that A labelled positive → Check that B fully understands the task and Guidelines
- The first round of annotation often has poor IAA

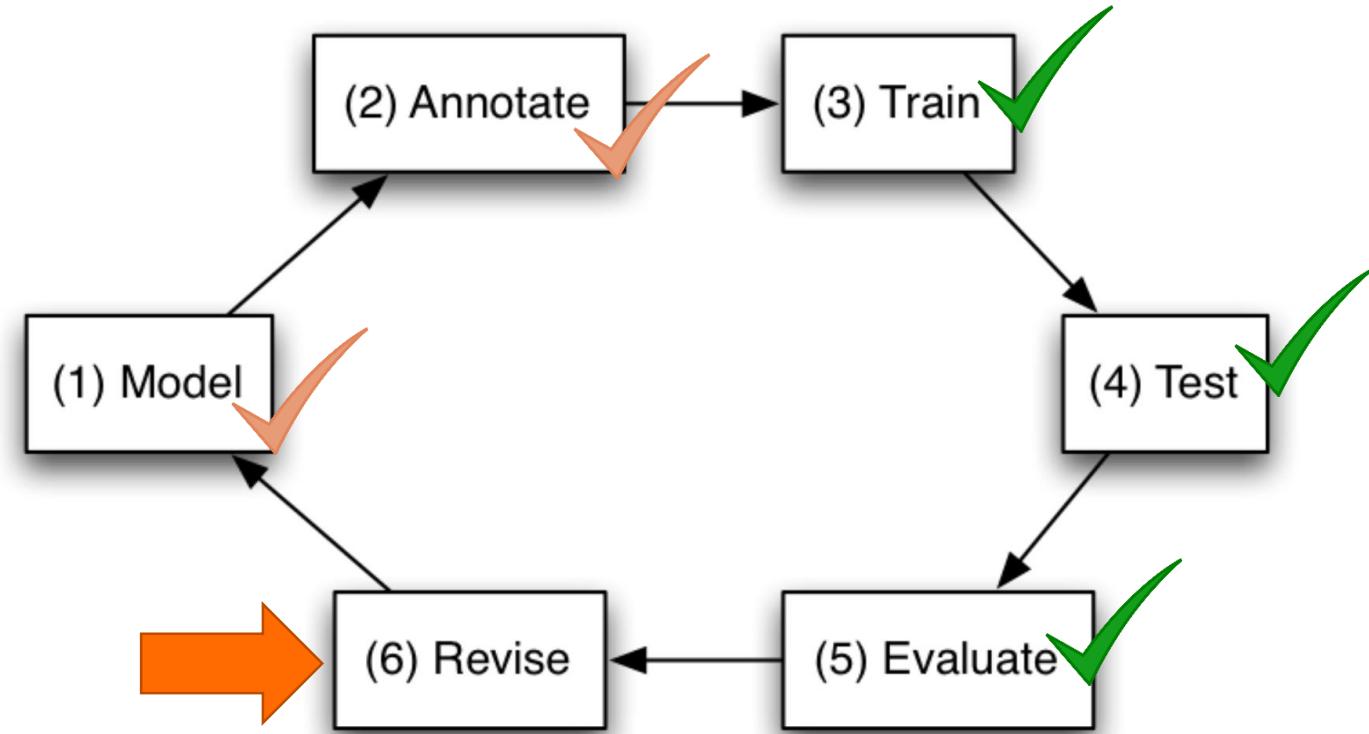
	B positive	B neutral	B negative
A positive	54	28	3
A neutral	31	18	23
A negative	0	21	72

K	Agreement level
< 0	poor
0.01–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	perfect

Creating the Gold Standard (Adjudication)

- One or more adjudicators
- Smooth over errors between annotators
 - Stay in line with the goal, specification, and guidelines
 - Just because two (or more) annotators agree on a label doesn't necessarily mean that the label is correct
- Look for things that were missed
- End result: Gold Standard

MATTER – Revise



Revision

- What can you improve in the process?
 - Corpus
 - Model and Specification
 - Annotation
 - Training and Testing
- Document everything about the annotation process (for reproducibility)

Annotation Process Example

Goal

Automatically extract the

- **Population**
- **Intervention**
- **Comparison**
- **Outcome**



P-I-C-O Elements

from the abstract and title of randomized controlled trials.

Efficacy and safety of three ciclesonide doses vs placebo in children with asthma: the RAINBOW study.

Pedersen S¹, Potter P, Dachev S, Bosheva M, Kaczmarek J, Springer C, et al.

 Author information

1. Population

Efficacy and safety of three doses of ciclesonide in children with asthma.

PATIENTS AND METHODS: This was a multicentre, double-blind, placebo-controlled, 12-week study of ciclesonide 40, 80 or 160 µg once daily, compared to placebo. Children (6-11 years) were randomised 1:1 to treatment via a metered-dose inhaler. The primary variable was change from baseline in time to first lack of efficacy (TLOE). Secondary variables included: asthma symptom score (AQLQ), forced expiratory volume in 1 s (FEV(1)), body height and body mass index. Safety assessments included: adverse events (AEs), urinary cortisol excretion and body height.

2. Intervention

RESULTS: In total, 1073 children improved with all doses of ciclesonide, but ciclesonide was significantly better than placebo in time to first TLOE, but ciclesonide was not significantly better than placebo in QoL. There were no differences between groups in AEs. The rates of AEs were comparable between all treatment groups (approximately 30%), and there were no significant differences between groups in body height or urinary cortisol excretion.

3. Comparison

“Is the I better than the C?”

4. Outcome

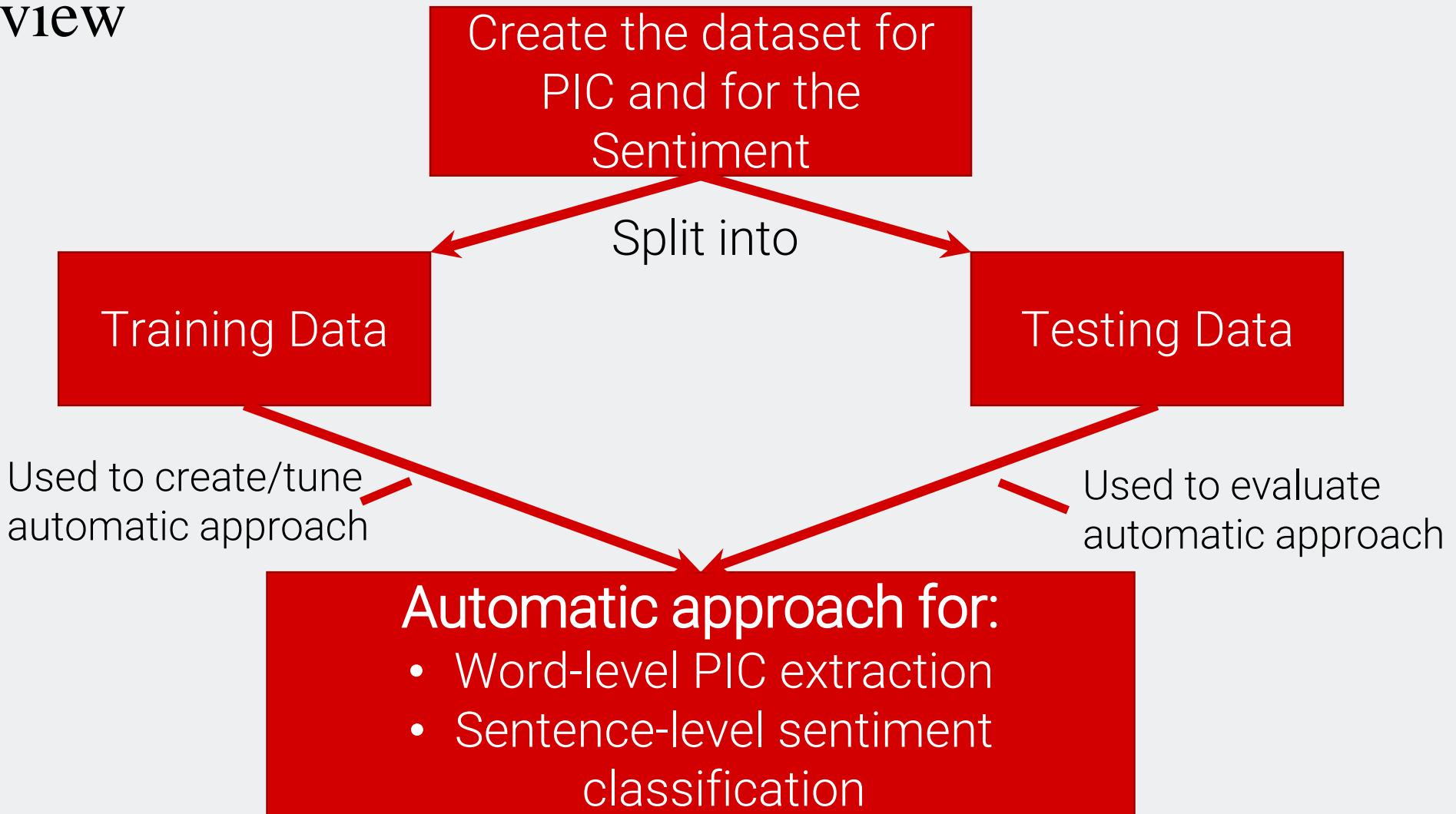
CONCLUSIONS: Ciclesonide 40-160 µg once daily is effective and well tolerated in children with persistent asthma; its efficacy and safety are unaffected by the use of a spacer. clinicaltrials.gov registration number: [NCT00384189](https://clinicaltrials.gov/ct2/show/NCT00384189).

Use case

Query:  children with asthma

Rank	Treatment	Evidence
#1	Ciclesonide	<p>RCT 1: <i>Ciclesonide versus other inhaled corticosteroids for chronic asthma in children.</i></p> <p>...</p> <p>RCT N: <i>Efficacy and safety of three ciclesonide doses vs placebo in children with asthma: the RAINBOW study.</i></p>
#2	Theophylline	<p>RCT 1: <i>A comparison of oral choline theophyllinate and beclomethasone in severe perennial asthma in children.</i></p> <p>...</p>
#3	Accolate	<p>RCT 1: <i>Effectiveness and tolerability of zafirlukast for the treatment of asthma in children.</i></p> <p>...</p>
...		

Overview



Machine Learning Prototype

PubMed ID:

20619624

Submit

Legend: Population **Population** Intervention **Intervention** Comparison **Comparison** Conclusion Sentence Positive **Conclusion Sentence Positive** Conclusion Sentence Neutral **Conclusion Sentence Neutral**

Title: Efficacy and safety of three ciclesonide doses vs placebo in children with asthma: the RAINBOW study.

Abstract:

OBJECTIVE

To evaluate the efficacy and safety of three doses of ciclesonide (with or without spacer) in children with persistent asthma.

PATIENTS AND METHODS

...

CONCLUSIONS

Ciclesonide 40-160 µg once daily is effective and well tolerated in children with persistent asthma; its efficacy and safety are unaffected by the use of a spacer. [clinicaltrials.gov](#) registration number: NCT00384189.

Data creation

- 6 human annotators (2 linguists, 4 persons from the medical domain) annotated title and abstract of Randomized Controlled Trials (RCTs)
 - Annotated PIC at **word-level**
 - Annotated the Sentiment at **sentence-level** (using the conclusion sentences)

First Annotation Prototype

Problem:
Inter-annotator
Agreement of
~20%

Why?
Too much
freedom through
free text input

<p>Title</p> <p>Ceramic water filters impregnated with silver nanoparticles as a point-of-use water-treatment intervention for HIV-positive individuals in Limpopo Province, South Africa: a pilot study of technological performance and human health benefits.</p>	<p>ID</p> <p>5579735</p>
<p>Abstract</p> <p>Waterborne pathogens present a significant threat to people living with the human immunodeficiency virus (PLWH). This study presents a randomized, controlled trial that evaluates whether a household-level ceramic water filter (CWF) intervention can improve drinking water quality and decrease days of diarrhea in PLWH in rural South Africa. Seventy-four participants were randomized in an intervention group with CWFs and a control group without filters. Participants in the CWF arm received CWFs impregnated with silver nanoparticles and associated safe-storage containers. Water and stool samples were collected at baseline and 12 months. Diarrhea incidence was self-reported weekly for 12 months. The average diarrhea rate in the control group was 0.064 days/week compared to 0.015 days/week in the intervention group ($p < 0.001$, Mann-Whitney). Median reduction of total coliform bacteria was 100% at enrollment and final collection. CWFs are an acceptable technology that can significantly improve the quality of household water and decrease days of diarrhea for PLWH in rural South Africa.</p>	
<p>Text Assignment</p> <p>Mark text in the title/abstract or write it down manually</p> <p>Population + Intervention + Comparator + Outcome + Sample Size +</p>	<p>Need Advice?</p> <p>HIV-positive individuals</p> <p>Add as Population?</p> <p>Add Discard</p>

Second Annotation Prototype

1. Select a sentence.
The sentence is shown below with an annotation from a medical pipeline (disease, drug, anatomy, person, ...)

Abstract

AIMS

The aim of this study was to compare the effects of brimonidine and timolol on retinal nerve fiber layer (RNFL) thickness in ocular hypertensive patients.

METHODS

This was a prospective, comparative, and unmasked study. For 12 months, 38 eyes of 19 patients with primary open-angle glaucoma received brimonidine tartrate 0.2%, and 40 eyes of 20 patients received timolol maleate 0.5%. Intraocular pressure (IOP) was measured every 2 months, and RNFL thickness was assessed using Scanning laser polarimetry (GDx) at baseline and at 12 months.

RESULTS

Mean IOP reduction was similar in both groups. Within each treatment group, the RNFL thickness for ellipse average ($P = 0.004$), superior ($P = 0.035$), temporal ($P = 0.003$), inferior ($P < 0.0001$), and nasal averages ($P = 0.044$) were significantly decreased from baseline in timolol at 12 months. However, the RNFL thickness for ellipse average and four quadrants showed no significant change from baseline in brimonidine. The between-group difference in RNFL change showed a significant reduction for ellipse average ($P = 0.02$), temporal ($P = 0.005$), and inferior averages ($P = 0.016$) following timolol therapy, as compared to brimonidine.

CONCLUSIONS

There appear to be less progression for RNFL damage following brimonidine 0.2% therapy compared to timolol 0.5% in ocular hypertensive patients over 1 year. This finding does not correlate with IOP reduction.

Previous Sentence 1 Next Sentence

Next Document

Current Sentence (To annotate, click an arbitrary start and end token)

Drug
Drug
Drug
Anatomy
The aim of this study was to compare the effects of brimonidine and timolol on retinal nerve fiber layer (RNFL)
Anatomy
Person
thickness in ocular hypertensive patients .

3. Select the correct class and "Save"

Title
Effects of oral erythromycin on esophageal pH and pressure profiles in patients with gastroesophageal reflux disease. B

Abstract
Erythromycin, a possible ...
Effects of oral erythromycin on esophageal pH and pressure profiles in **patients with gastroesophageal reflux disease**. D

Population Intervention Compare
A

Save

Previous Sentence 0 Next Sentence C Return to Overview

Drug Disease Person Disease
Effects of oral erythromycin on esophageal pH and pressure profiles in patients with gastroesophageal reflux disease. C

Title

Short term correction of anaemia with recombinant human erythropoietin and reduction of cardiac output in end stage renal failure.

Abstract

Previous Sentence

3 ▾

Next Sentence

Return to Overview

Current Sentence (To annotate, click an arbitrary start and end token)

Eleven children with end stage renal failure and anaemia (haemoglobin concentration < 90 g/l) were enrolled into a single blind , placebo controlled , crossover study to assess the cardiovascular effects of reversing anaemia using subcutaneous human recombinant erythropoietin (r-HuEpo) .

```
graph TD; children[children] -- Person --> endStageRenalFailure[end stage renal failure]; children[children] -- Person --> anaemia[anaemia]; endStageRenalFailure -- Disease --> haemoglobin[haemoglobin concentration < 90 g/l]; anaemia -- Disease --> haemoglobin; haemoglobin -- Anatomy --> cardiovascular[cardiovascular effects]; rHuEpo[recombinant erythropoietin] -- Drug --> subcutaneousHuman[r-HuEpo]
```

Annotations

Compare	placebo	
Population	children with end stage renal failure and anaemia (haemoglobin concentration < 90 g/l)	
Intervention	subcutaneous human recombinant erythropoietin (r-HuEpo)	

Creation of training data (Sentiment)

Abstract

PURPOSE

This study was designed to investigate the efficacy of arthrocentesis with and without injection of sodium hyaluronate (SH) into the upper joint space in the treatment of temporomandibular joint (TMJ) internal derangements.

PATIENTS AND METHODS

Forty-one TMJs in 5 males and 26 females aged 14 to 53 years comprised the study material. The patients' complaints were limited mouth opening, TMJ pain and tenderness, and joint noises during function. The study was performed in 2 groups: arthrocentesis plus intra-articular injection of SH versus arthrocentesis alone. The study was performed in 1 group and arthrocentesis plus intra-articular injection of SH was performed in the other group. The study contained patients with disc displacement with reduction and without reduction. The study was performed immediately after the procedure, on postoperative day 1, and at 1, 3, 6, and 12 months. The study assessed jaw function, and clicking sounds in the TMJ were assessed using a visual analog scale (VAS). The study was recorded at each follow-up visit.

RESULTS

Both techniques increased maximal mouth opening, lateral movements, and function, while reducing TMJ pain and noise.

CONCLUSIONS

Although patients benefitted from both techniques, arthrocentesis with injection of SH seemed to be superior to arthrocentesis alone.



Although patients benefitted from both techniques, arthrocentesis with injection of SH seemed to be superior to arthrocentesis alone.

The conclusion of the abstract is used to determine the sentiment

Annotation Guidelines (1)

- 12 Page Document with:
 - Extensive documentation on how PIC is defined
 - General Guidelines
 - Extensive Examples
 - Common Patterns
- Annotation of P, I and C are done concurrently by each annotator

INTERVENTION & COMPARISON

Sentence: This double-blind, double-dummy, parallel-group study was designed to show that a pharmacokinetically enhanced formulation of oral amoxycillin-clavulanate (16:1, 2000/125 mg), twice daily, is at least as effective clinically and microbiologically as oral amoxycillin-clavulanate 1000/125 mg, three times daily, in the 10 day treatment of community-acquired pneumonia (CAP) in adults.

Pattern

a pharmacokinetically enhanced [INTAKE][DRUG][DOSAGE], [DOSAGE: duration] is at least as effective clinically and microbiologically as [INTAKE][DRUG][DOSAGE], [DOSAGE: duration]

Annotate this word sequence as continuum for Intervention

a pharmacokinetically enhanced formulation of oral amoxycillin-clavulanate (16:1, 2000/125 mg), twice daily

Annotate this word sequence as continuum for Comparison

oral amoxycillin-clavulanate 1000/125 mg,three times daily

Note! The design of the trial is not part of the intervention or comparsion element i.e. do not annotate double-blind, double-dummy, parallel-group

Annotation Guidelines (2)

What to do if the intervention is mentioned multiple times in the title and abstract → annotate them all

Effects of pioglitazone in combination with metformin or a sulfonylurea compared to a fixed-dose combination of metformin and glibenclamide in patients with type 2 diabetes.

Abstract

Previous Sentence 2 ▾ Next Sentence

Next Document

Current Sentence (To annotate, click an arbitrary start and end token)

Person Drug
Patients (n = 250) treated with metformin < or = 3 g/day) or an SU as monotherapy for > 3 months and with

glycosylated hemoglobin (HbA (1c)) between 7.5 % and 11 % inclusive were randomized to receive either

Drug Drug Drug
pioglitazone (15-30 mg/day) as add-on therapy to metformin or an SU or a fixed-dose combination of metformin (

Drug
400 mg) and glibenclamide (2.5 mg) (up to three tablets per day) for 6 months .

Annotations

Intervention	metformin	
Compare	SU as monotherapy	
Compare	glycosylated hemoglobin (HbA (1c))	
Intervention	pioglitazone	
Intervention	add-on therapy to metformin	
Compare	SU	
Compare	fixed-dose combination of metformin (400 mg) and glibenclamide	

Figure 4: Example of co-reference occurrence of INTERVENTION element in title and abstract.

Similar Task but different Annotators (Mechanical Turk)

- Split the task into smaller parts
 - Participants – note the choice of a more “friendly” term than population
 - Interventions
 - Outcomes
- Whole abstract is annotated
- Reduce the length of the Annotation Guidelines significantly
- Specific Mechanical Turk requirements:
 - workers require a minimum approval rate of 90% on previous tasks to participate
 - spammers are removed in a small-scale test run

Complete Annotation Guidelines (Participants)

- In medical studies, the efficacy of medical treatments is evaluated within a group of study participants.
- We present to you a sentence of a study report in which your task is to highlight the text that gives information about the participants of the study. You can highlight text in the sentence by clicking on a start and end word. If no information about the participants is mentioned, mark the corresponding checkbox.
- Relevant information about participants include:
 - gender
 - medical conditions (e.g. diseases, upcoming surgery)
 - location ("patients in Taiwanese Hospitals")
 - how many people were in the study
- Do not highlight:
 - participant mentions without relevant information ("Patients were divided into two groups." versus "**Patients with diabetes** were divided into two groups.")
- To give additional context, we show the study report in that the sentence appears. The report might be helpful, e.g., to identify that an abbreviation *AD* stands for *Alzheimer Disease*.

Examples

A study on the efficacy of recombinant human endostatin combined with apatinib mesylate in patients with middle and advanced stage non-small cell lung cancer.

To investigate the role of nicotinamide adenine dinucleotide phosphate 4 (NADPH4,NOX4) and transforming growth factor-beta (TGF- β) involve in pathogenesis of airway remodeling in chronic obstructive pulmonary disease (COPD).

A total of 270 patients with MCI were enrolled in a 24-week, multicenter, randomized, double-blind, placebo-controlled study.

Participants with mild-to-moderate AD (Mini-Mental State Examination score of 13-26) were recruited from December 1999 to November 2000 using clinic populations, referrals from community physicians, and local advertising.

A PPARA Polymorphism Influences the Cardiovascular Benefit of Fenofibrate in Type 2 Diabetes: Findings From ACCORD Lipid.

Refinement 1

- Annotate on a sentence level

amazonmturk
Worker

Text highlighting in sentences of medic... ([HIT Details](#)) Auto-accept next HIT

Requester [Markus Zlabinger](#)

HITs 5 Reward \$0.03

Time Elapsed 0:10 of 10 Min

Task Instructions (Click to expand)

Study Report (Click to expand)

Sentence (Highlight information about **participants** by clicking on a start and end word)

573 patients who had unrecognized and untreated anxiety identified from the approximately 8,000 patients who completed the waiting room screening questionnaire .

Sentence does not contain information about participants

Submit

Refinement 2

- Give annotators dynamic examples of expert annotations

Study Report (Click to expand)

Sentence (Highlight information about participants by clicking on a start and end word)

A multi-component social skills intervention for children with Asperger syndrome : the Junior Detective Training Program .

Sentence does not contain information about participants

Examples (Caution: The shown examples might contain missing highlights.)

Social skills training (SST) is a common intervention for children with autism spectrum disorders (ASDs) to improve their social and communication skills .

Teaching emotion recognition skills to young children with autism : a randomised controlled trial of an emotion training programme .

A randomized controlled study of a social skills training for preadolescent children with autism spectrum disorders : generalization of skills by training parents and teachers ?

Submit

Inter-Annotator Agreement

M. Zlabinger, M. Sabou, S. Hofstätter, A. Hanbury, *Effective Crowd-Annotation of Participants, Interventions, and Outcomes in the Text of Clinical Trial Reports*. Findings of the Association for Computational Linguistics: EMNLP 2020, <https://www.aclweb.org/anthology/2020.findings-emnlp.274>

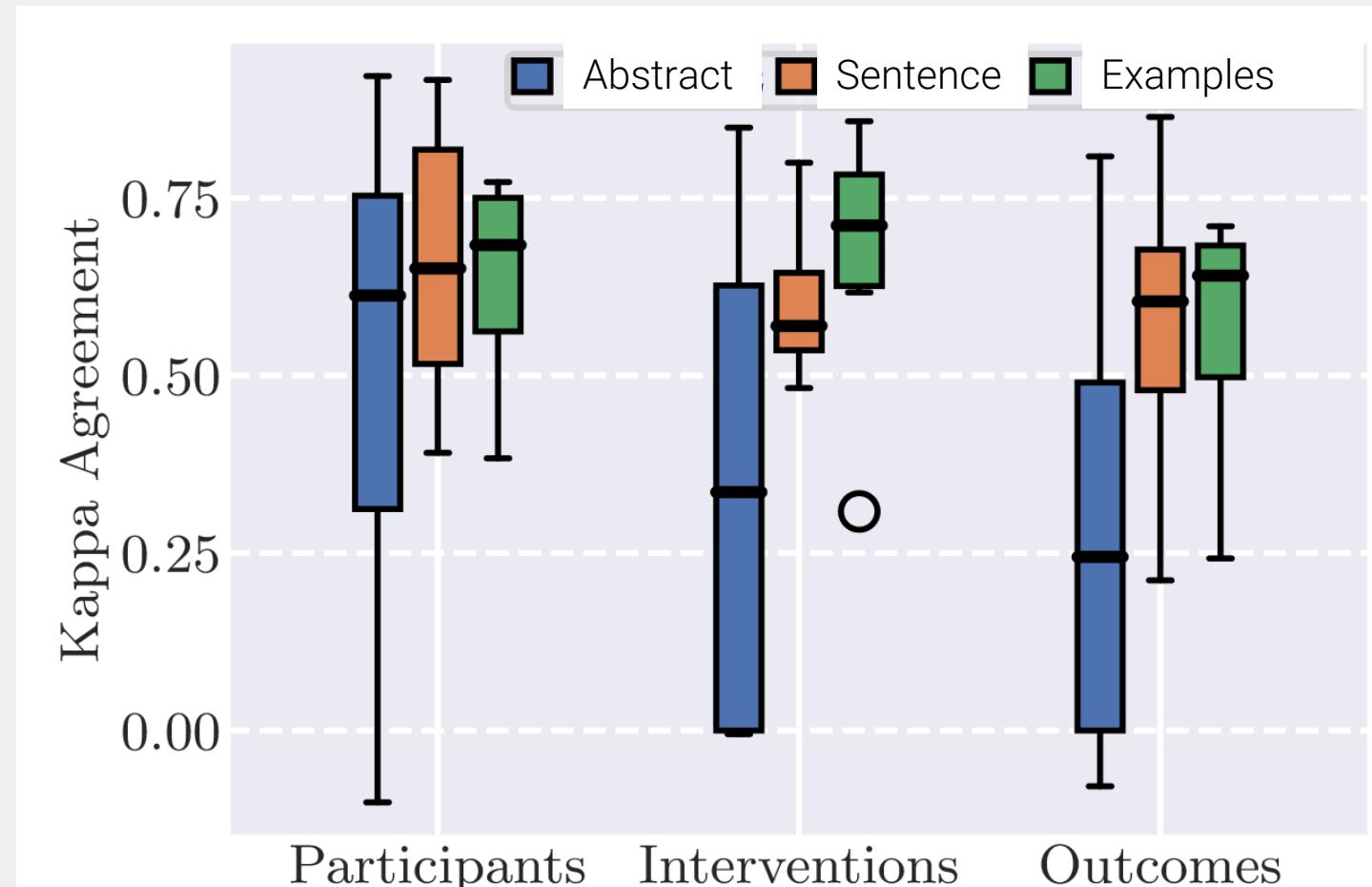


Figure 1: Kappa agreements between annotations from individual workers and the gold standard.

Agreement with the Gold Standard

	P	I	O
Abstract	0.702	0.455	0.352
Sentence	0.715	0.675	0.655
Examples	0.780	0.757	0.694

Cohen's Kappa agreements between three aggregated annotations of each approach and the gold standard. Aggregation was done by Majority Voting.

K	Agreement level
< 0	poor
0.01–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	perfect

Summary

- Annotation sounds straightforward, but usually is not
- Pay a lot of attention to the model and guidelines
 - Reduce variation in annotation by limiting annotation interface options or very specific guidelines
 - Annotators do not read the guidelines
 - If it is possible to annotate something in more than one way, then all possible ways will be used
- Improve the annotation process iteratively, with few annotators and documents in the early iterations
- Be sure that all problems in the annotation process have been ironed out before starting a massive annotation campaign