Annotation Basics and Challenges

Allan Hanbury

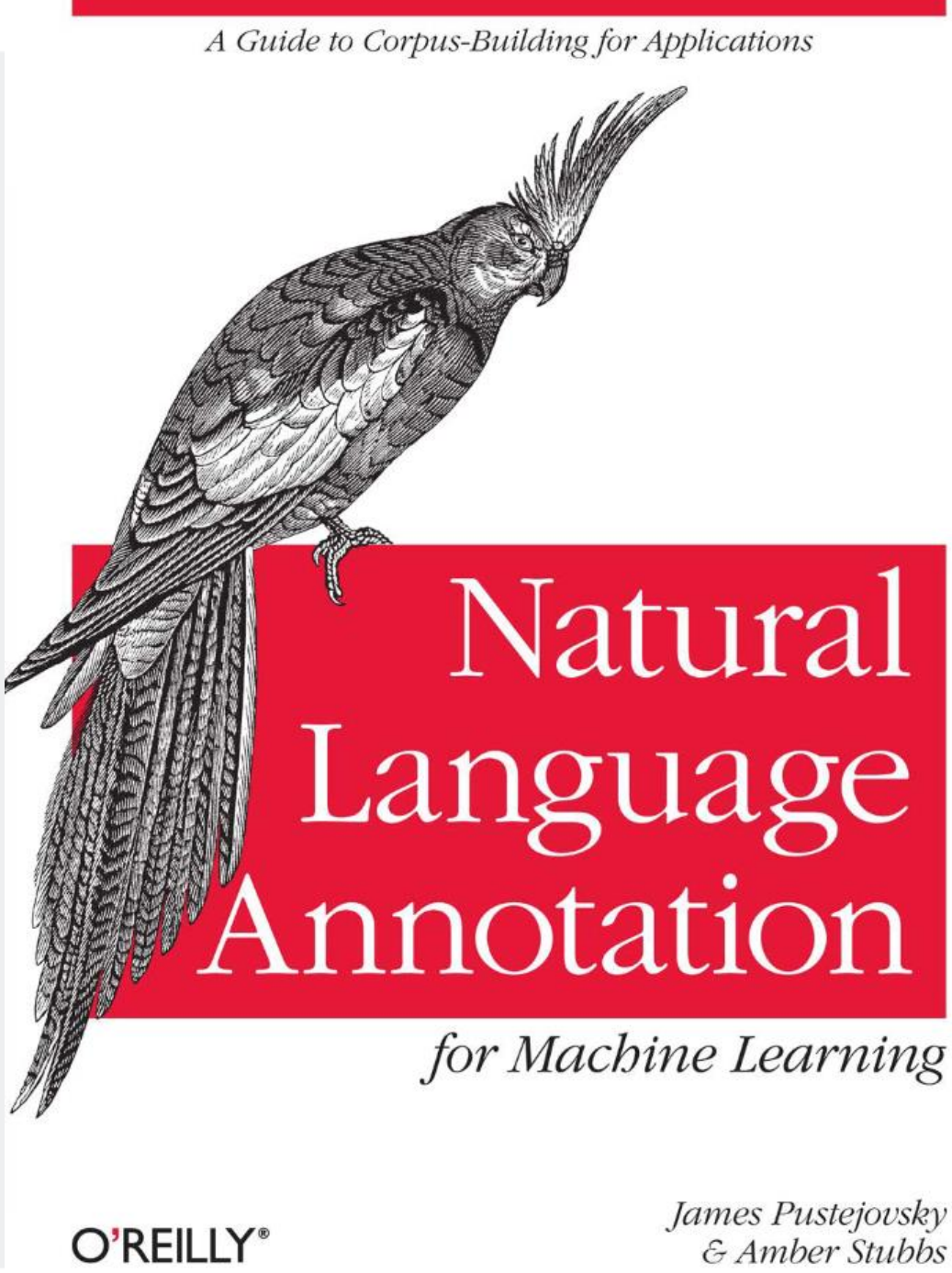
TU WIEN Informatics

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/brandeisnlaml/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

IT'S NOT ABOUT THE ALGORITHM

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.

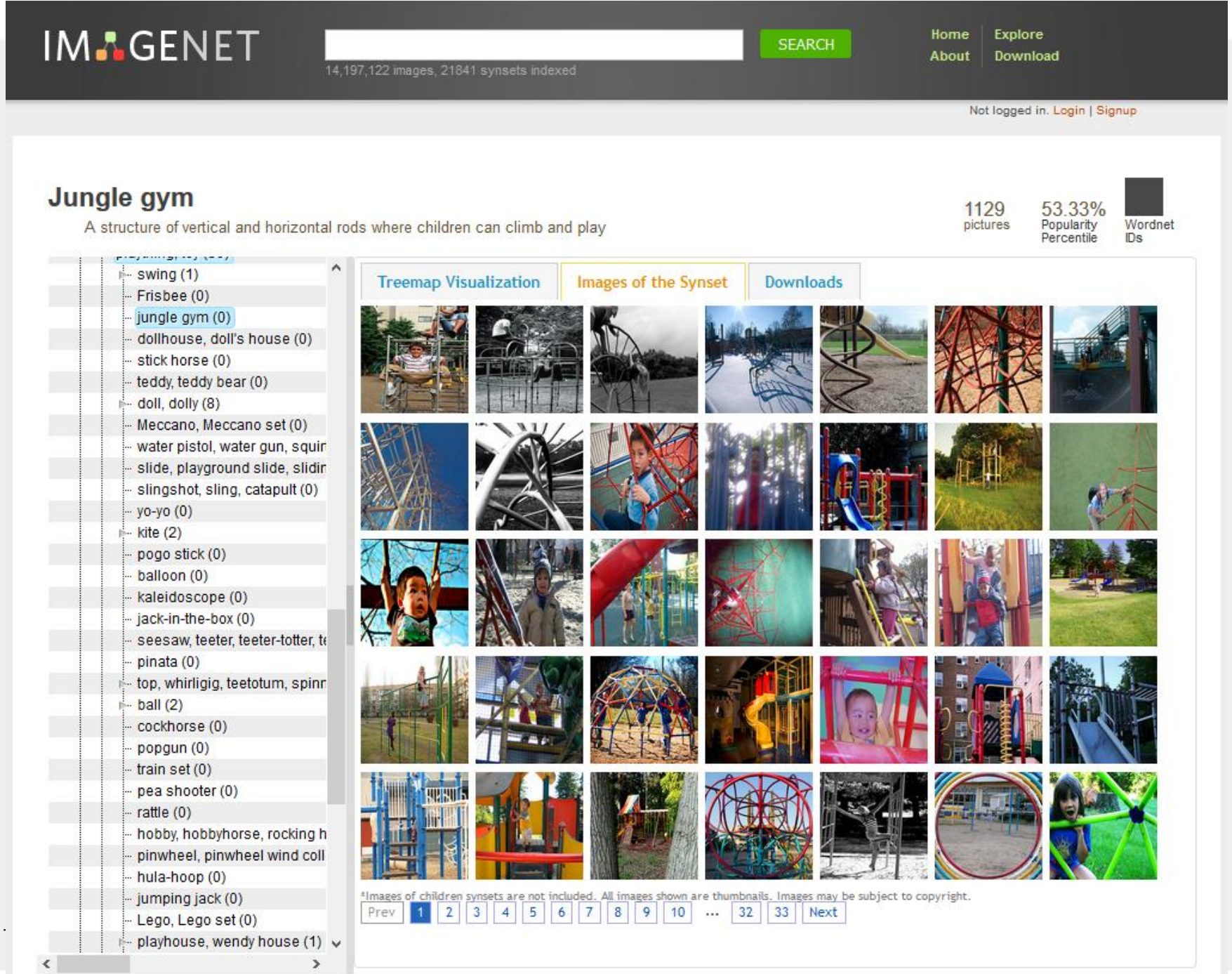
https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

IMAGENET

- 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)

https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/


IMAGENET
SEARCH
14,197,122 images, 21841 synsets indexed
Home
Explore
About
Download
Not logged in. Login | Signup
Jungle gym
A structure of vertical and horizontal rods where children can climb and play
1129
pictures
53.33%
Popularity
Percentile
Wordnet
IDs
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, t
pinata (0)
top, whirligig, teetotum, spinr
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

IMAGENET

SEARCH

14,197,122 images, 21841 synsets indexed

Home | Explore | About | Download

Not logged in. Login | Signup

Bean sprout

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

1135 pictures | 65.06% Popularity Percentile | Wordnet IDs

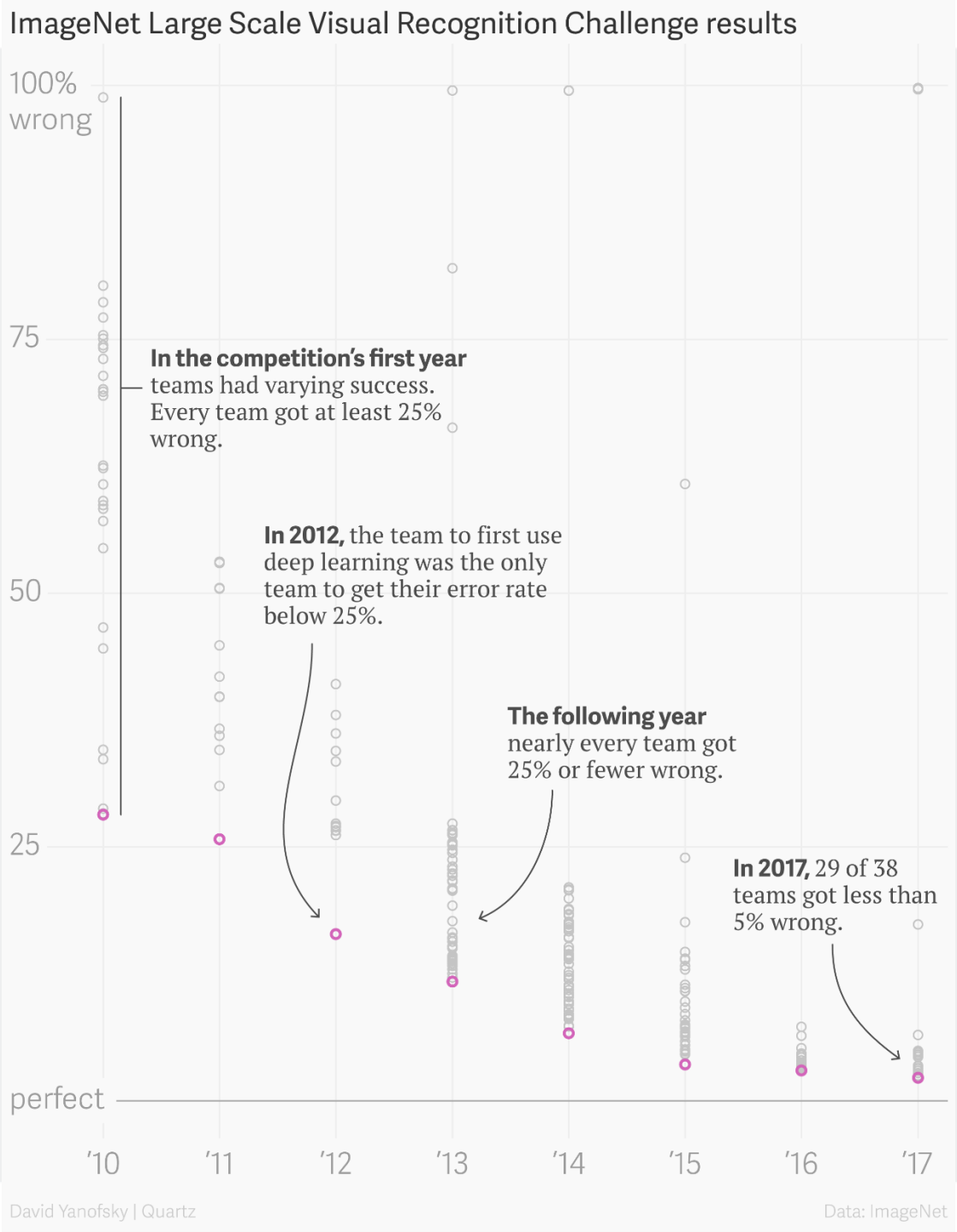
- wax bean, yellow bean (0)
- Fordhooks (0)
- lima bean (1)
- sieva bean, butter bean, butterl
- fava bean, broad bean (0)
- green soybean (0)
- shell bean (5)
- fresh bean (13)
- flageolet, haricot (0)
- common bean (20)
- field soybean (0)
- soy, soybean, soya, soya bean
- bean, edible bean (24)
- 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 b
- buttercrunch (0)

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

TU WIEN Informa



https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

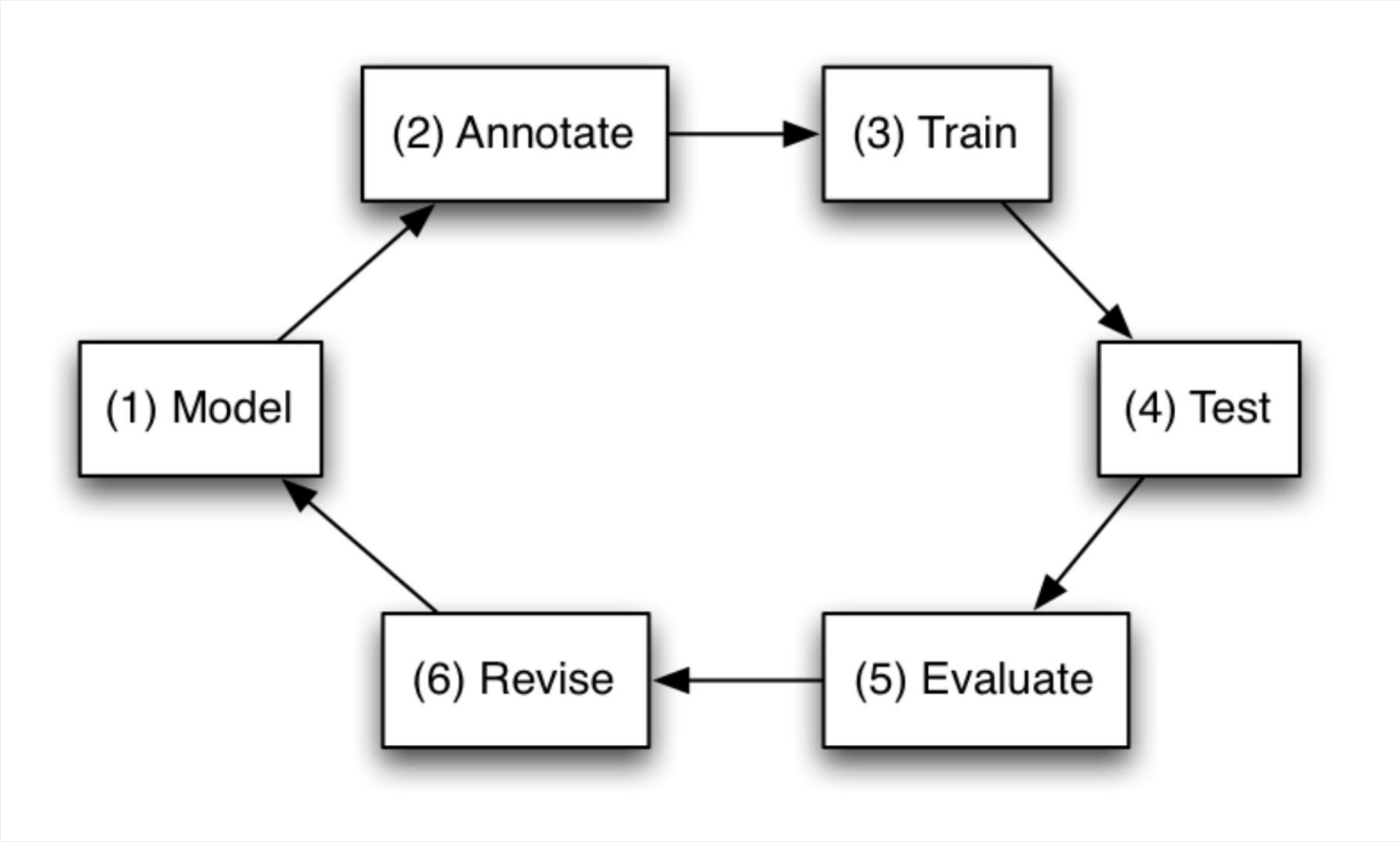
Aspects of Text Annotation

- 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)

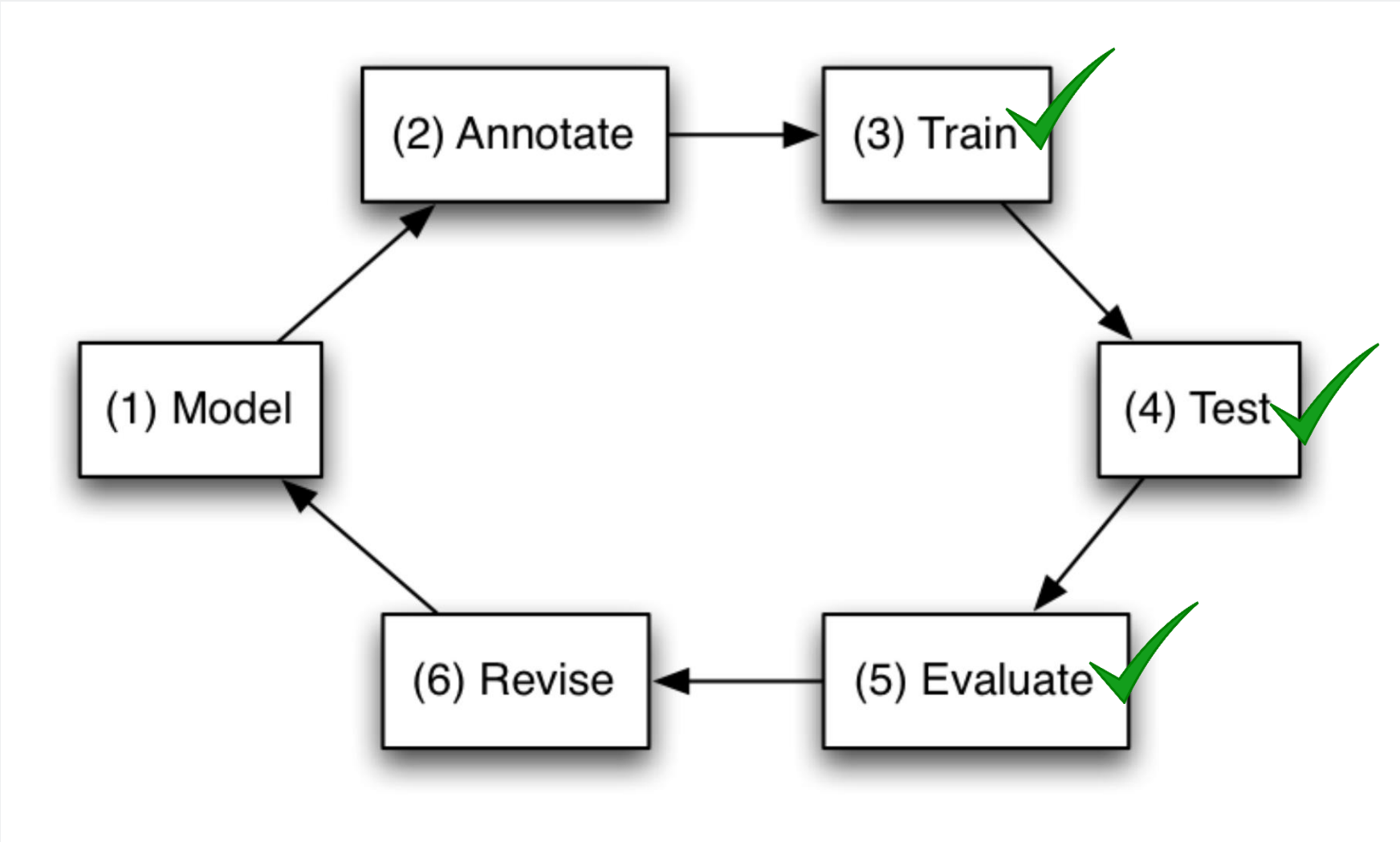
Creating Annotations

- 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



https://pixabay.com/de/vectors/tick-marke-okay-perfekt-%C3%BCberpr%C3%BCfen-305245/

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

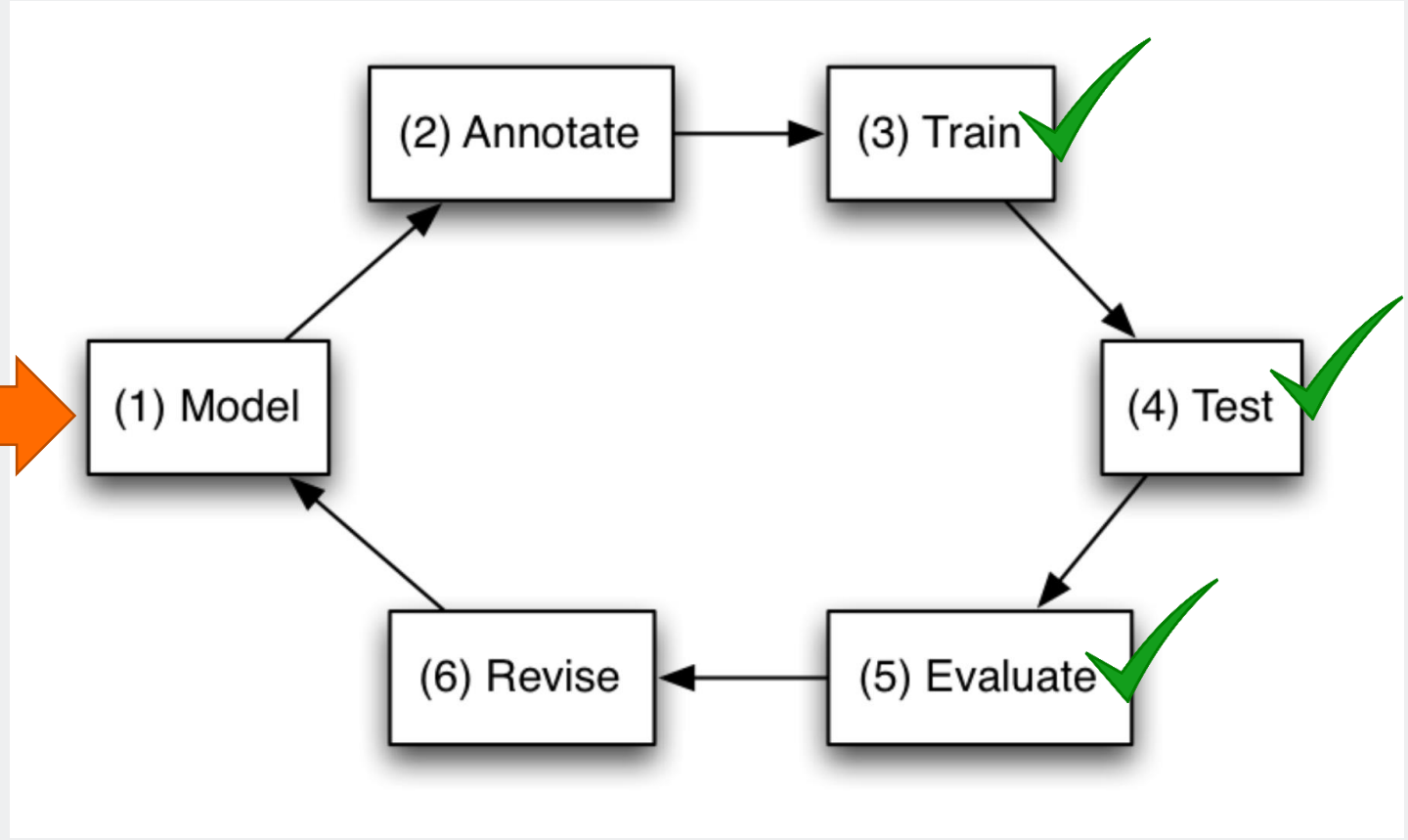
Achieving the Goal

- 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

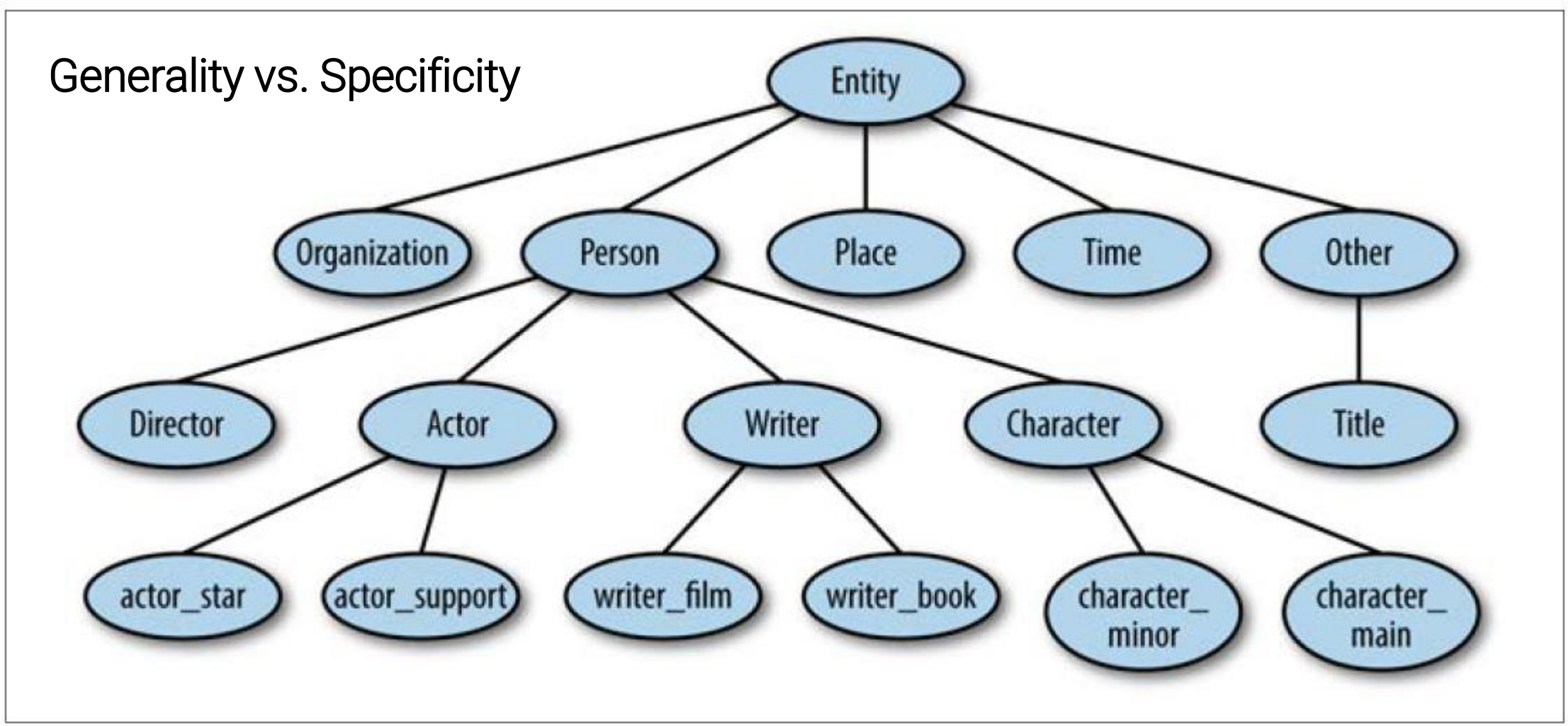
Example: Film Summaries – Named Entities

- 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

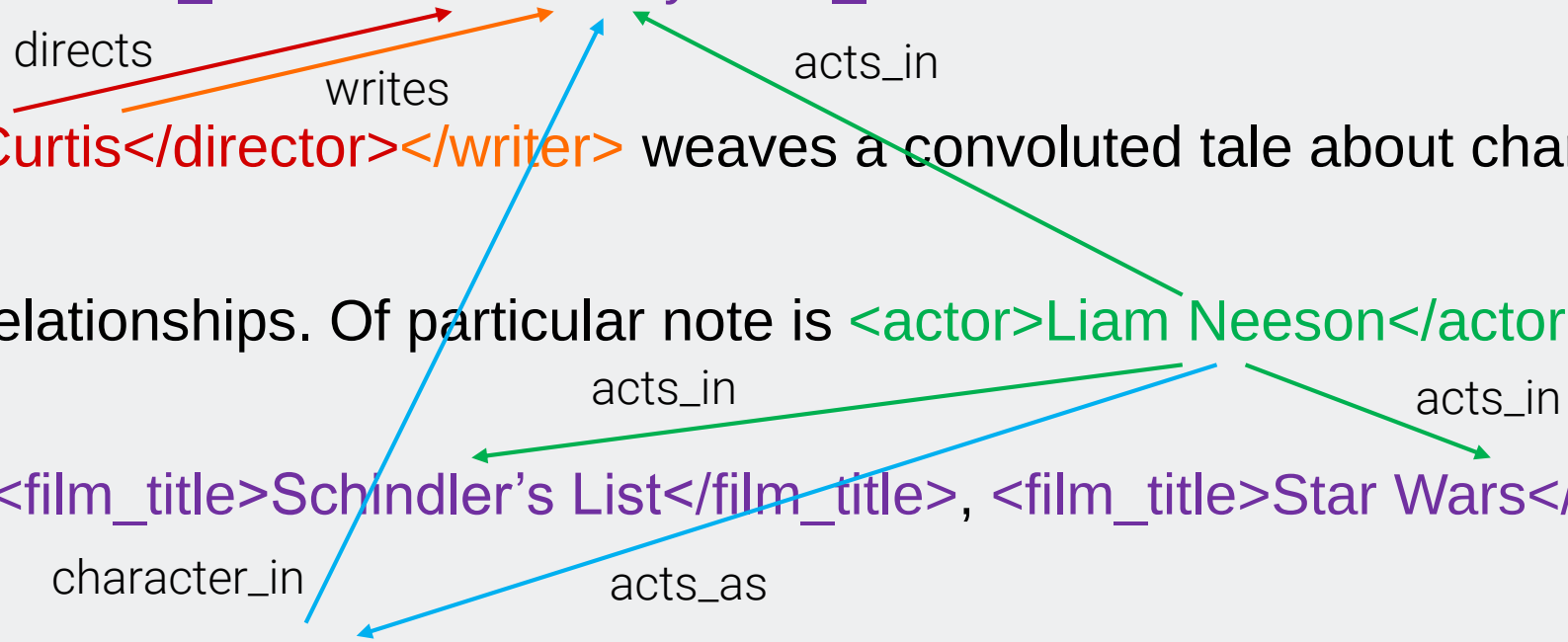


Example: Film Summaries – Semantic Roles

- 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> …”



Another Example: Time Stamping Events

April 25, 2010

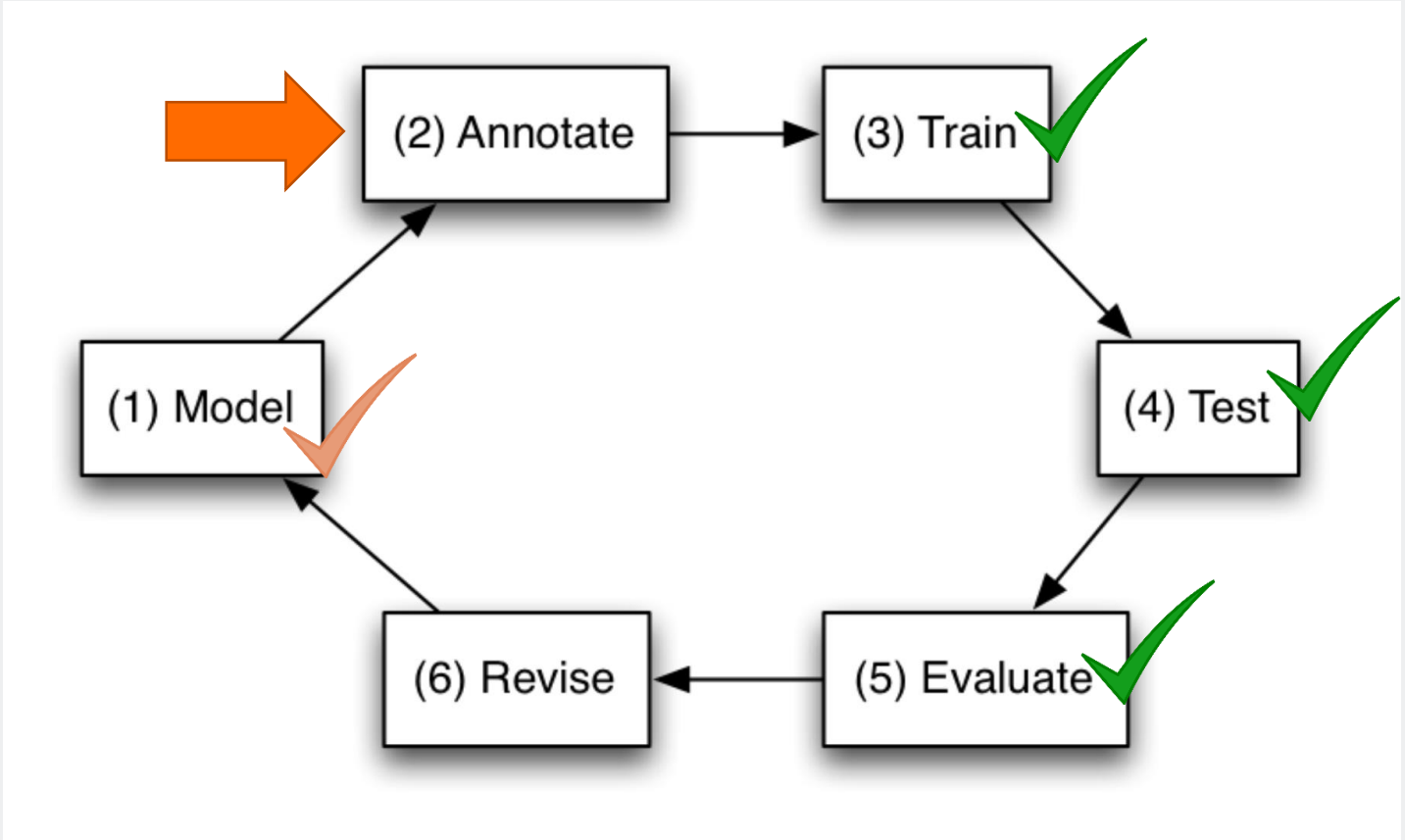
- President Obama paid tribute Sunday to 29 workers killed in an explosion at a West Virginia coal mine earlier this month, saying they died "in pursuit of the American dream." The blast at the Upper Big Branch Mine was the worst U.S. mine disaster in nearly 40 years.Obama ordered a review earlier this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

Another Example: Temporal Ordering of Events

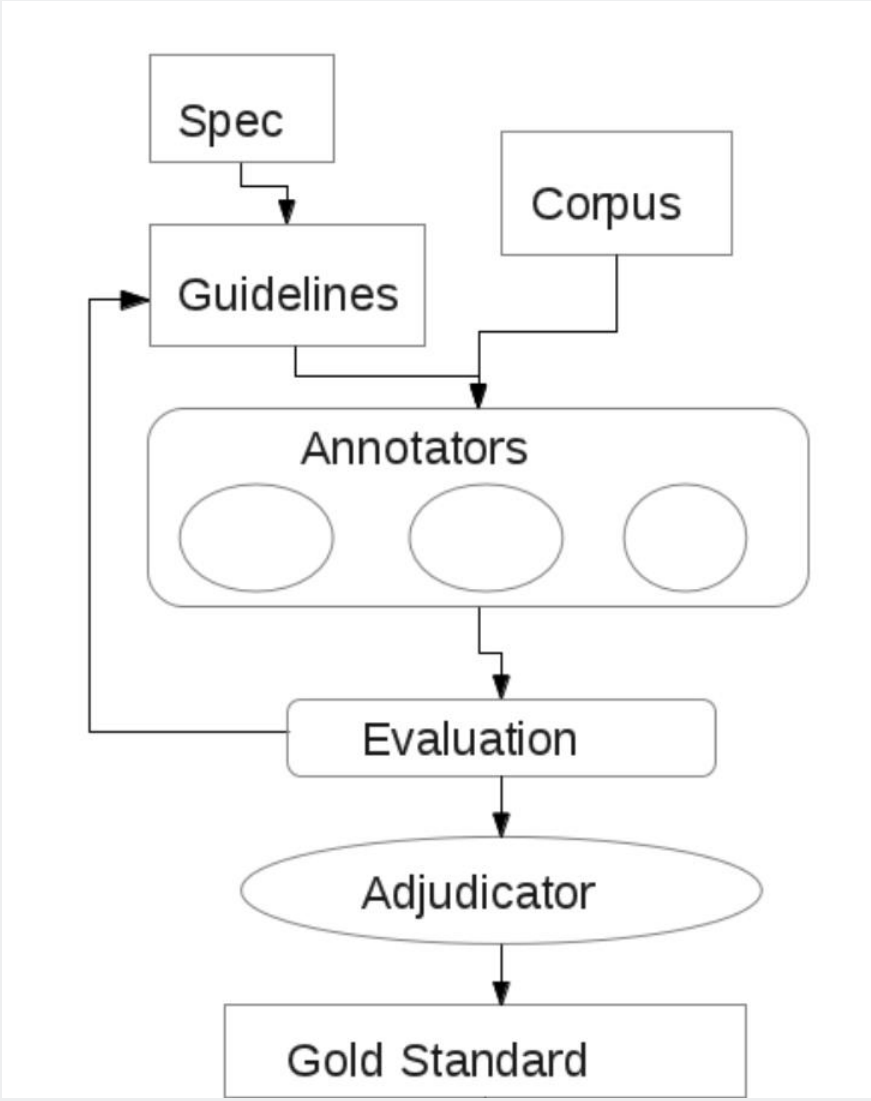
April 25, 2010

- 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

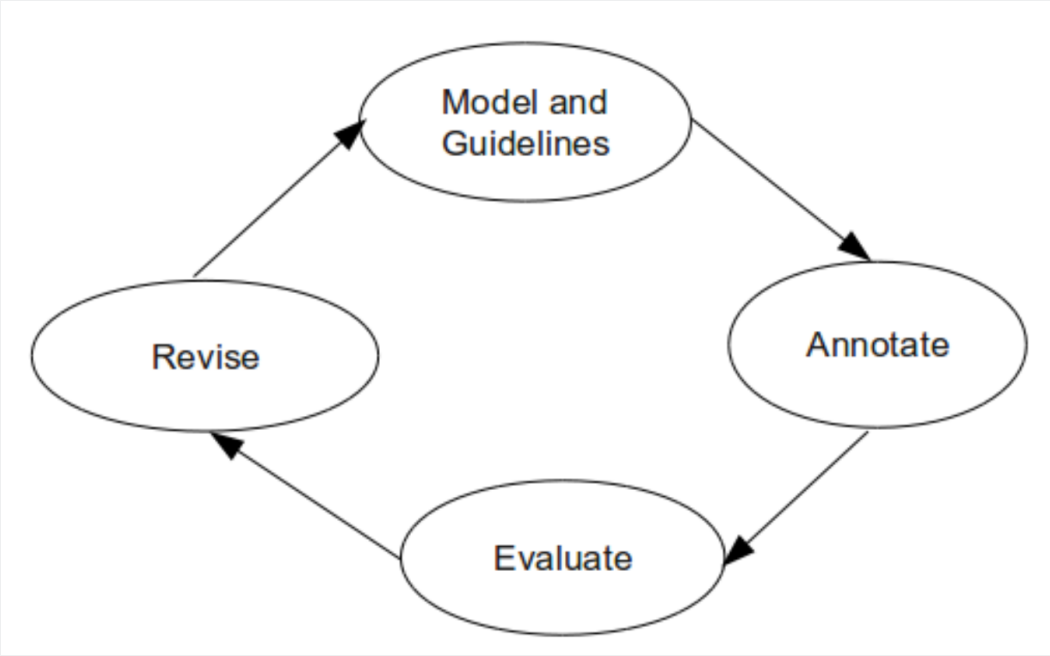
- Directions for the annotators
  - What, where, why, how
- Reusable
- Different from specification
  - Focus on how, why
- Balance between length and level of detail

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]

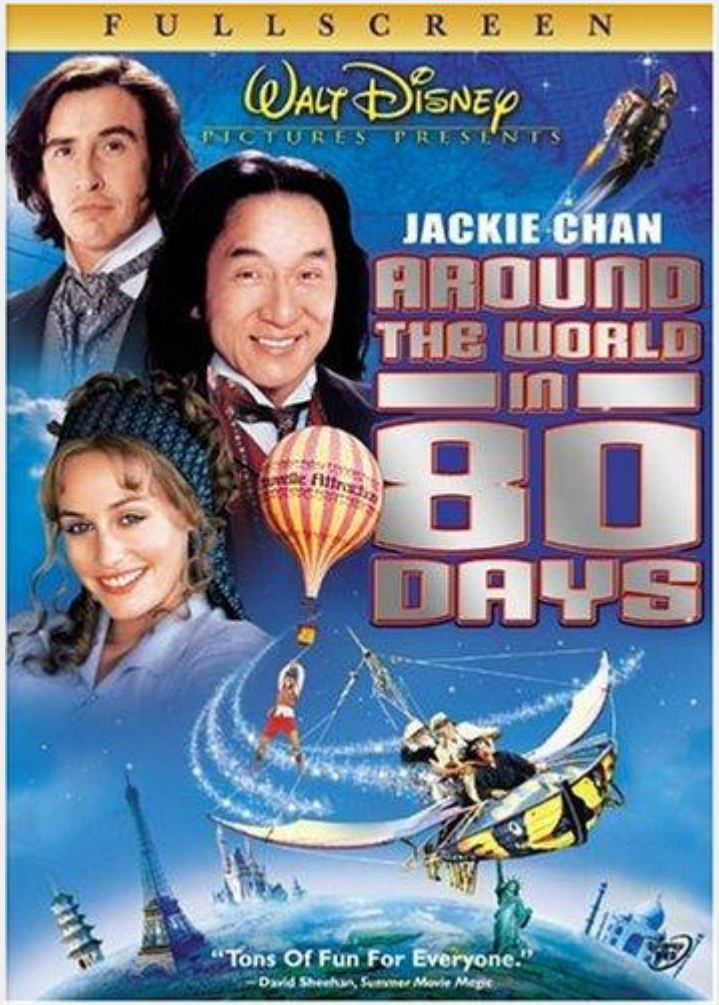
Kindle Customer

★★★★★ **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.

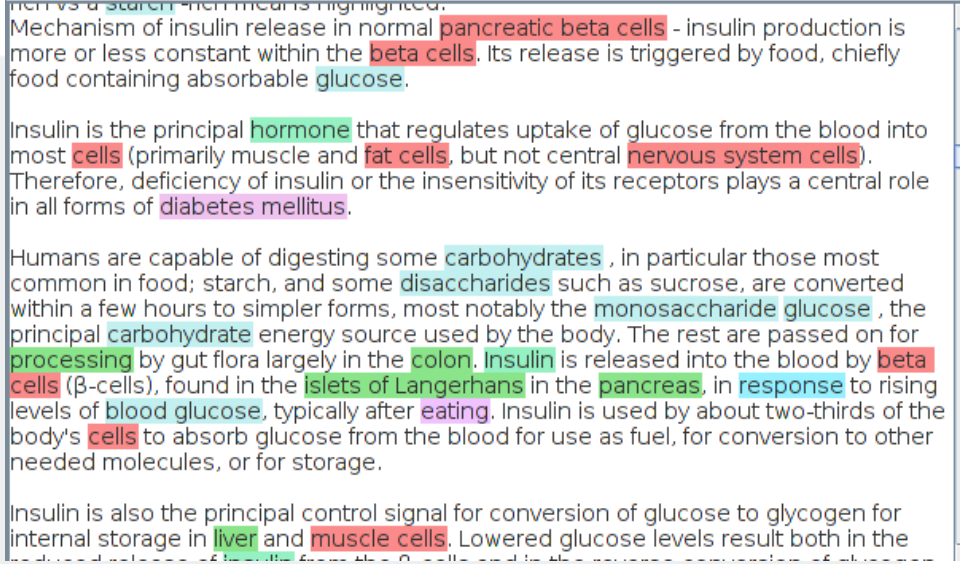
2.2.1 `BBBebaubareFlaecheGesamterBauplatz`

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


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

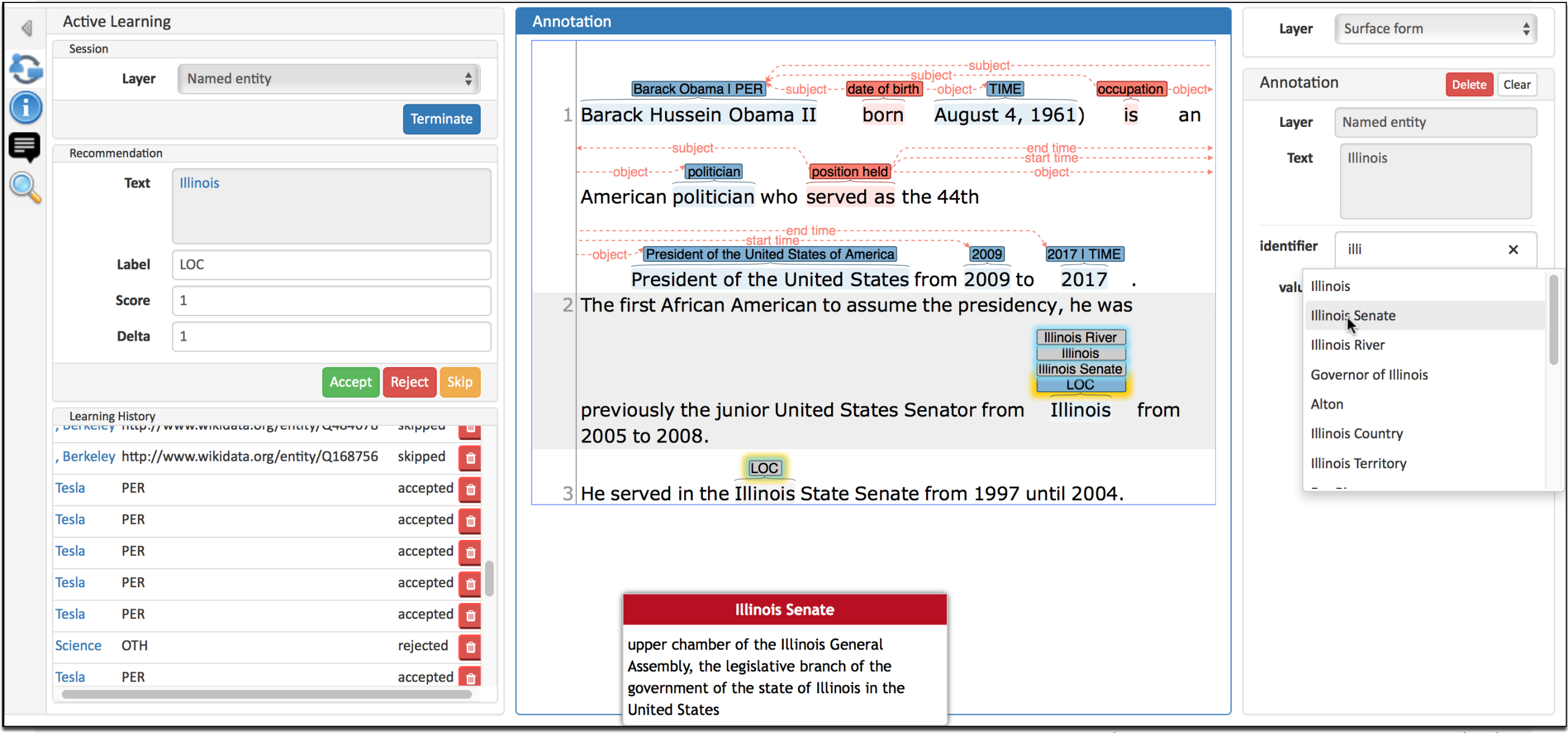
- How should the annotators decide how long each tagged span should be?
- How should complex cases of people's names be handled?
  E.g. "Stefan and Anna Warchalowski"
- Should every mention of a Named Entity be annotated or only the first mention?
- What about an entity playing two different roles?
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

Choosing an Annotation Environment

- Standalone or online?
- Existing tool or write your own?

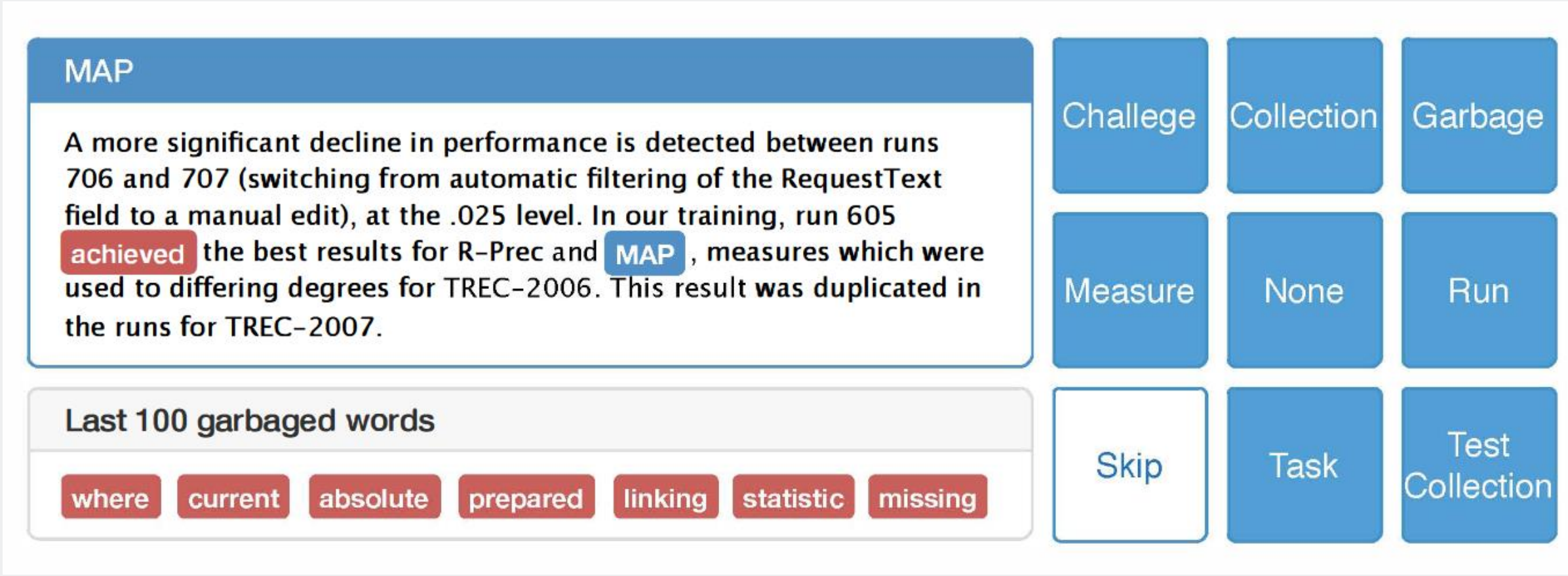
- Questions specific to the annotation task:
  - Does the learning curve of the tool match the capabilities of the annotators?
  - What parts of the task must the annotation software support?
  - What are the units of the annotation task? (characters, entities, documents, …)
  - Should the annotation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

Example: INCEpTION



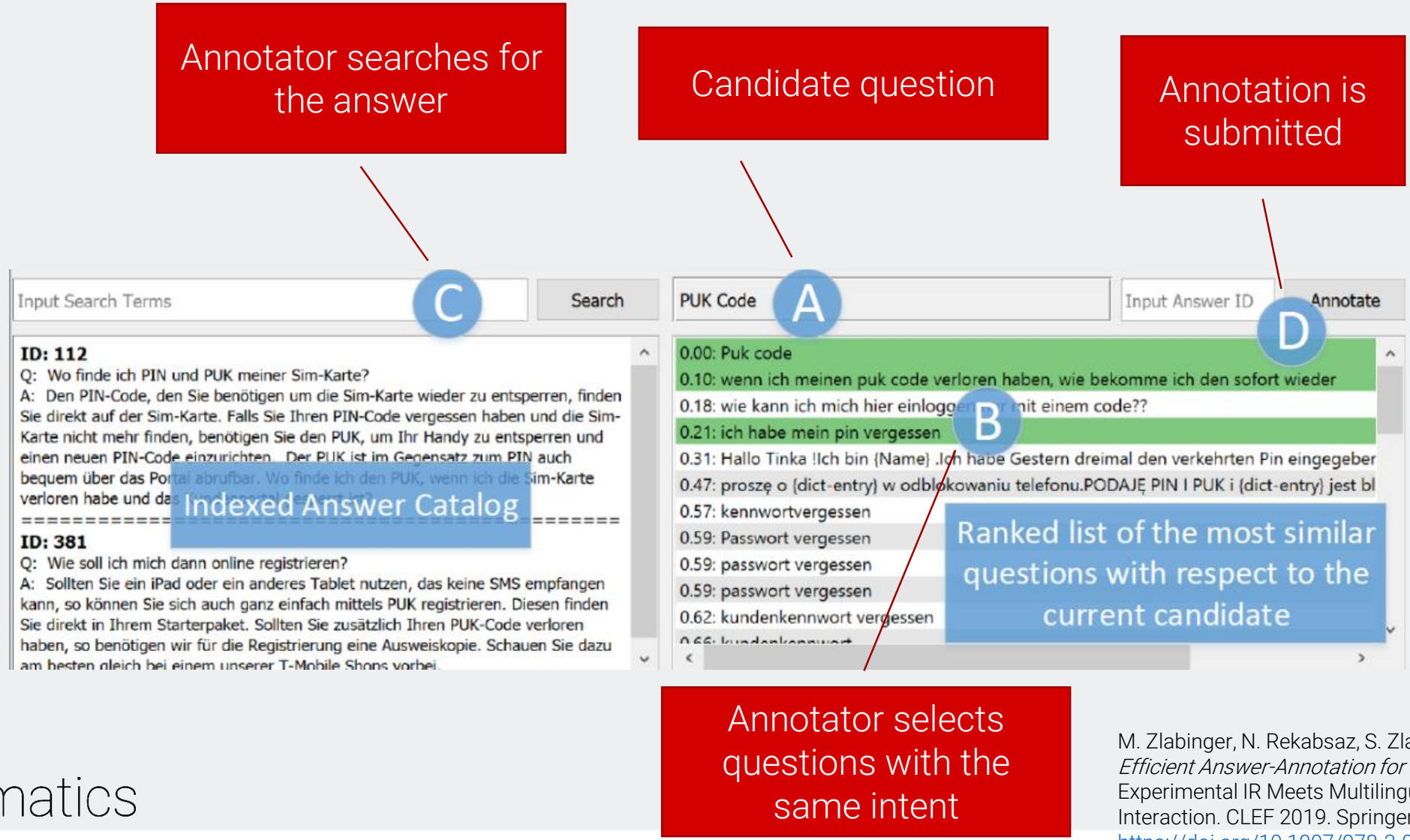
https://inception-project.github.io/

Example: Annotation for few concepts



F. Piroi, A. Lipani, M. Lupu, and A. Hanbury. *DASyR(IR) - document analysis system for systematic reviews (in Information Retrieval)*. Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR 2015), https://doi.org/10.1109/ICDAR.2015.7333830

Example: Question-Answering Interface



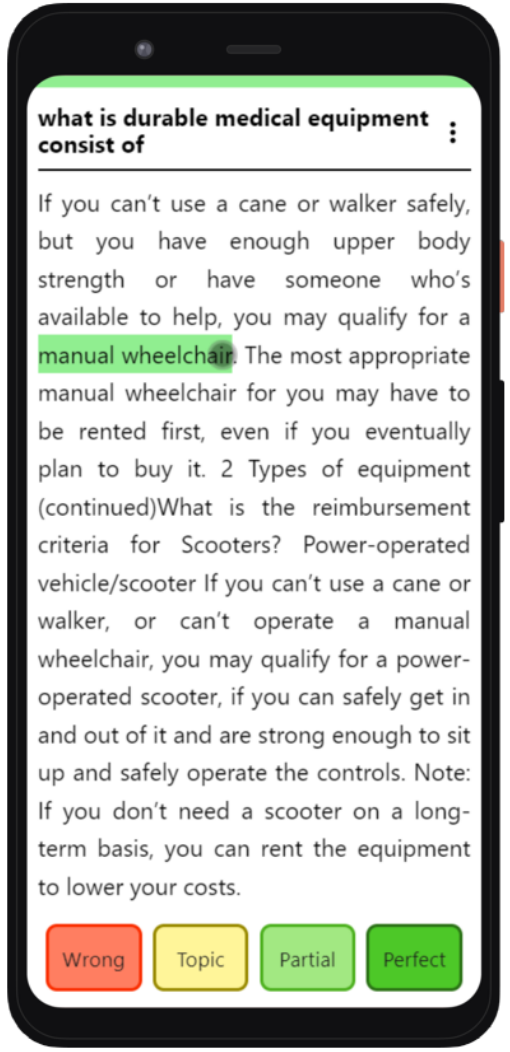
M. Zlabinger, N. Rekabsaz, S. Zlabinger, A. Hanbury (2019) *Efficient Answer-Annotation for Frequent Questions*. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. Springer LNCS 11696. https://doi.org/10.1007/978-3-030-28577-7_8

Example: Marking relevant text of relevant document extracts

what is durable medical equipment consist of

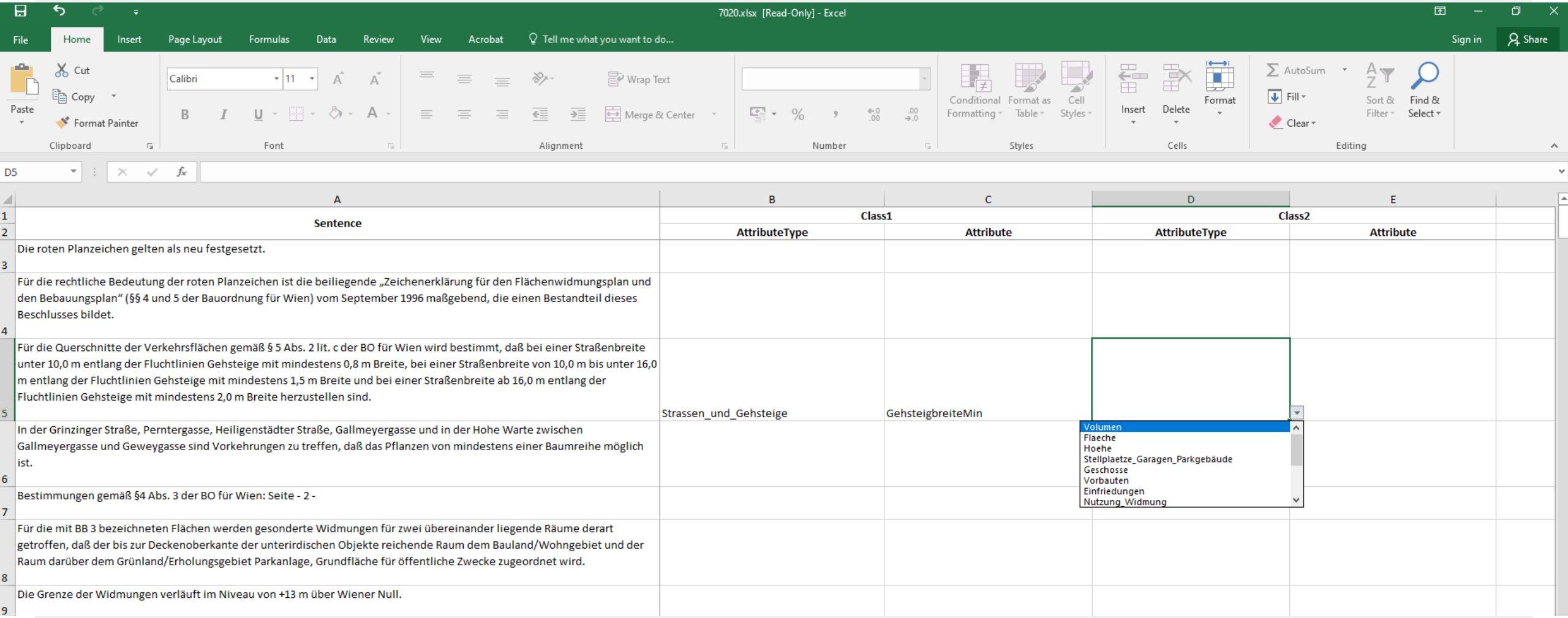
If you can't use a cane or walker safely, but you have enough upper body strength or have someone who's available to help, you may qualify for a manual wheelchair. The most appropriate manual wheelchair for you may have to be rented first, even if you eventually plan to buy it. 2 Types of equipment (continued)What 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



https://github.com/pkerschbaum/fira

Example: Sentence Annotation in Excel



	A	B	C	D	E
1	Sentence	Class1		Class2	
2		AttributeType	Attribute	AttributeType	Attribute
3	Die roten Planzeichen gelten als neu festgesetzt.				
4	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.				
5	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 Fluchtlinien Gehsteige mit mindestens 0,8 m Breite, bei einer Straßenbreite von 10,0 m bis unter 16,0 m entlang der Fluchtlinien Gehsteige mit mindestens 1,5 m Breite und bei einer Straßenbreite ab 16,0 m entlang der Fluchtlinien Gehsteige mit mindestens 2,0 m Breite herzustellen sind.	Strassen_und_Gehsteige	GehsteigbreiteMin		
6	In der Grinzinger 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.				
7	Bestimmungen gemäß §4 Abs. 3 der BO für Wien: Seite - 2 -				
8	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.				
9	Die Grenze der Widmungen verläuft im Niveau von +13 m über Wiener Null.				

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)

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

- Two annotators, A and B, assigned the labels "positive," "neutral," and "negative" to a set of 250 movie reviews
- $\Pr(a) = \frac{(54+18+72)}{250} = 0.576$
- 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
- $\Pr(e) = 0.116 + 0.077 + 0.146 = 0.339$
- $\kappa = \frac{(0.576-0.339)}{(1-0.339)} = 0.359$

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

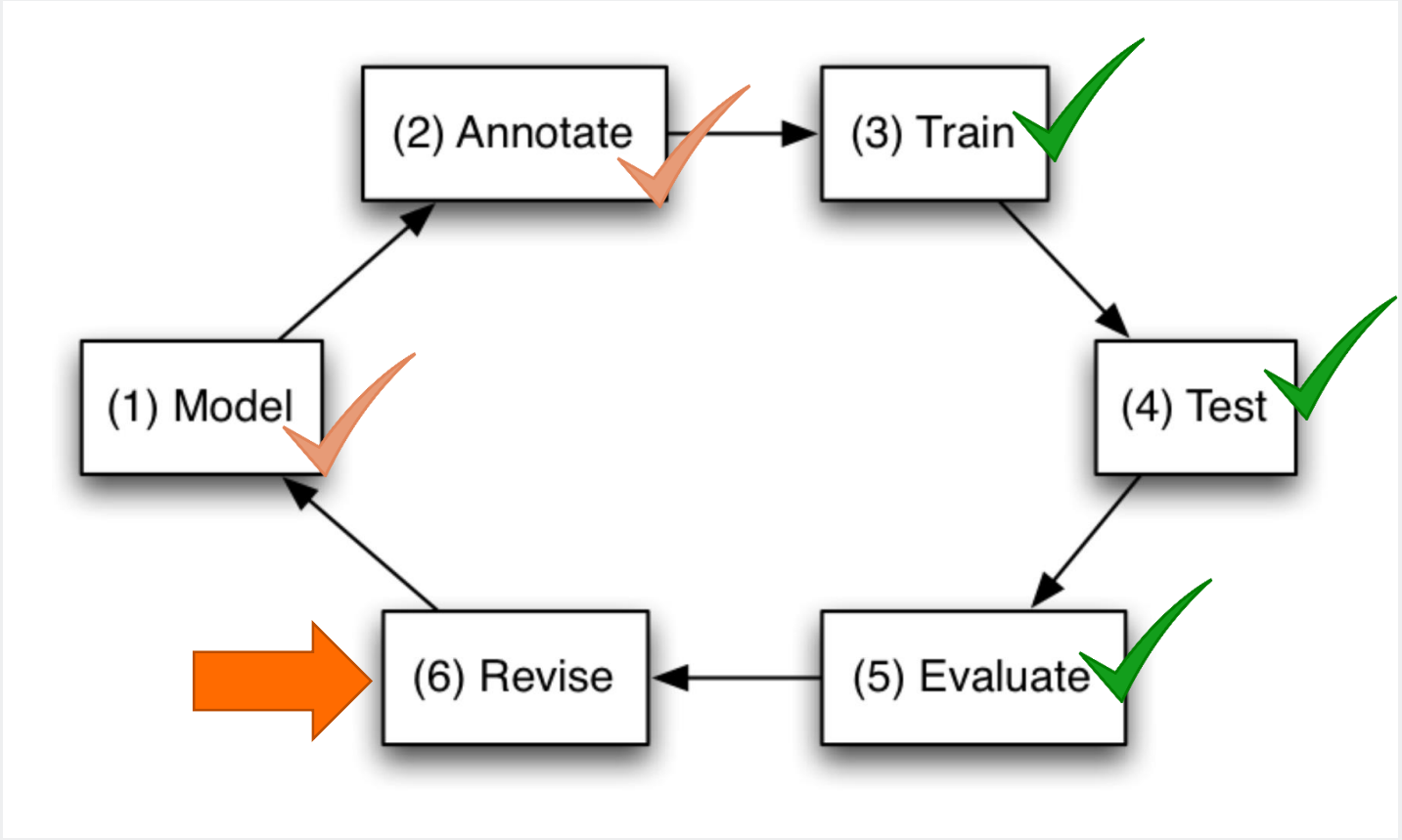
		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

κ	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.

M. Zlabinger, L. Andersson, A. Hanbury, M. Andersson, V. Quasnik, J. Brassey, *Medical Entity Corpus with PICO elements and Sentiment Analysis*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), https://www.aclweb.org/anthology/L18-1044.pdf

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

Pedersen S[1], Potter P, Dachev S, Bosheva M, Kaczmarek J, Sprin

Author information

1. Population

2. Intervention

3. Comparison

4. Outcome
"Is the I better than the C?"

cacy and safety of three d asthma.

PATIENTS AND METHODS: This was a multicentre, double-blind, placebo-controlled, 12-week study of ciclesonide 40, 6-11 years) were randomised 1:1 to treatment via a meter er. The primary variable was change from baseline in F). Secondary variables included: time to first lack of effica ory volume in 1 s (FEV(1)), asthma symptom score nents included: adverse events (AEs), urinary cortisol excretion and body height.

RESULTS: In total, 1073 childr ificantly improved with all doses of cicl n time to first LOE, but ciclesonide was (1) and QoL. There were no difference ces of AEs were comparable betwee between-group differences in body height or urinary cortisol.

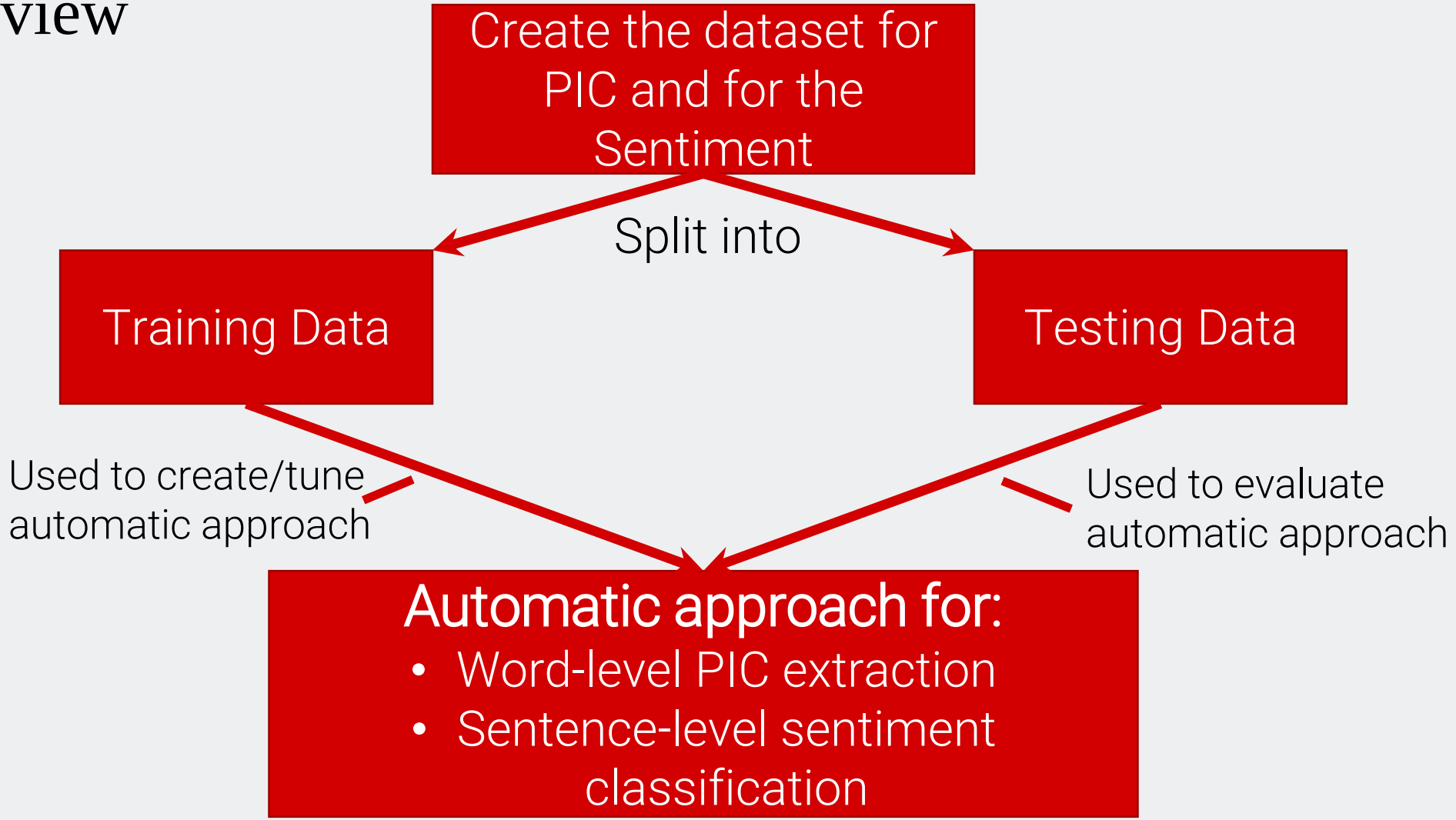
CONCLUSIONS: Ciclesonide 40-160 µg once daily is effective and well tolerated in children with persistent asthma; its efficacy and safety are unaffected by the use of a spacer. clinicaltrials.gov registration number: NCT00384189.

Use case

Query: 🔍 *children with asthma*

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

Overview


Create the dataset for PIC and for the Sentiment
Split into
Training Data
Testing Data
Used to create/tune automatic approach
Used to evaluate automatic approach
Automatic approach for:
• Word-level PIC extraction
• Sentence-level sentiment classification

Machine Learning Prototype

PubMed ID: 20619624 Submit

Legend: Population Intervention Comparison Conclusion Sentence Positive 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.

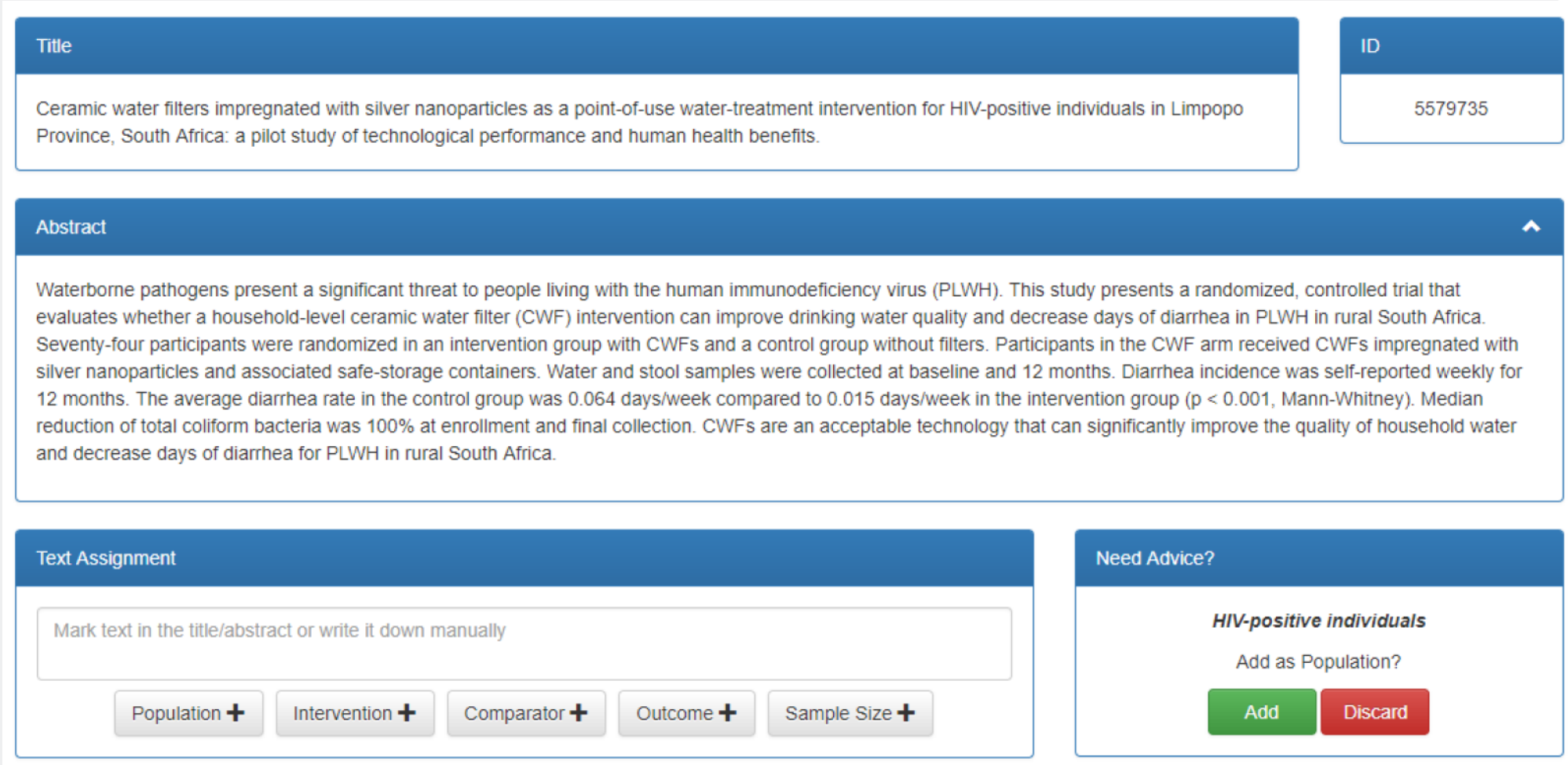
Data creation

- 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%

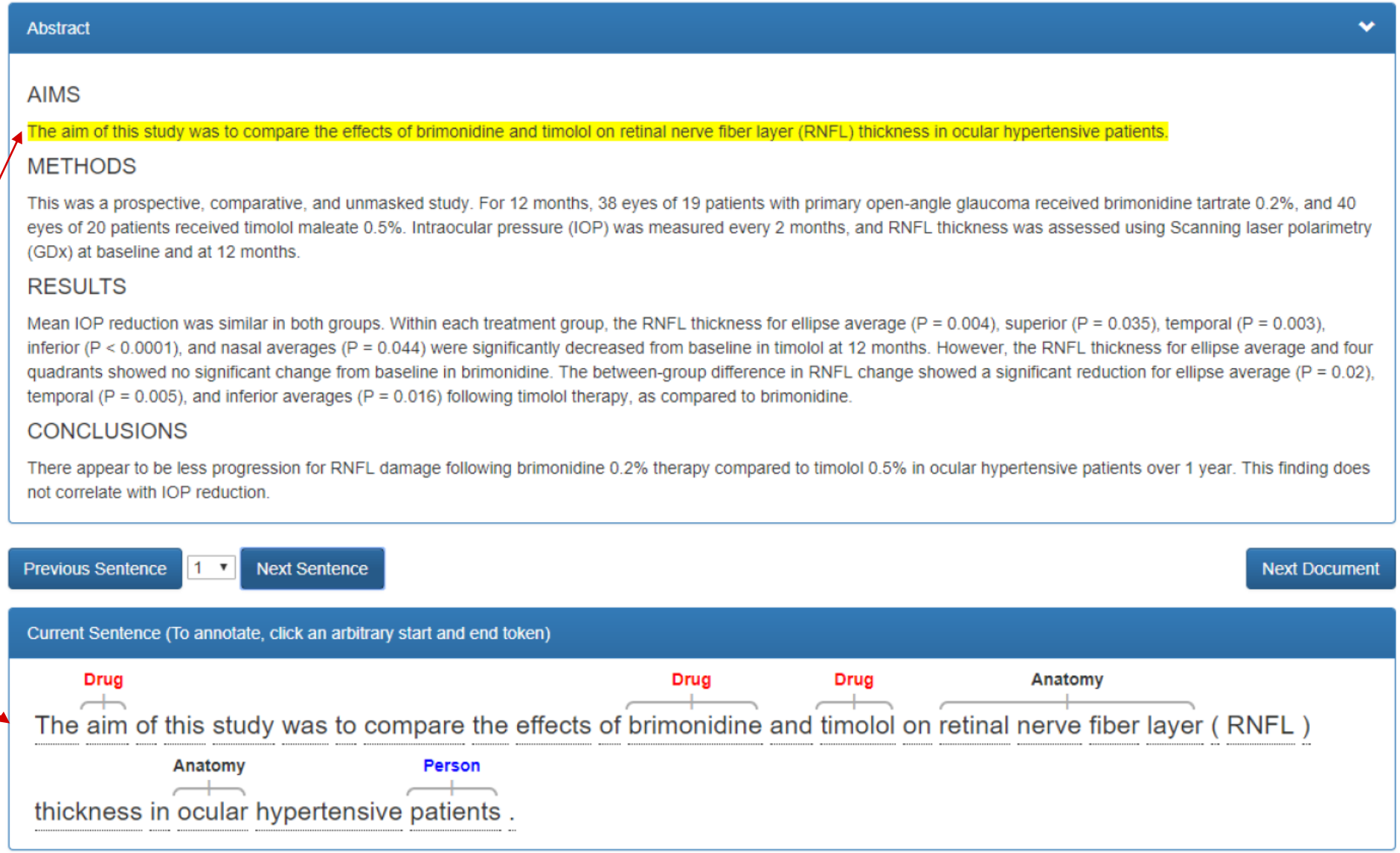
Why?
Too much freedom through free text input



Second Annotation Prototype

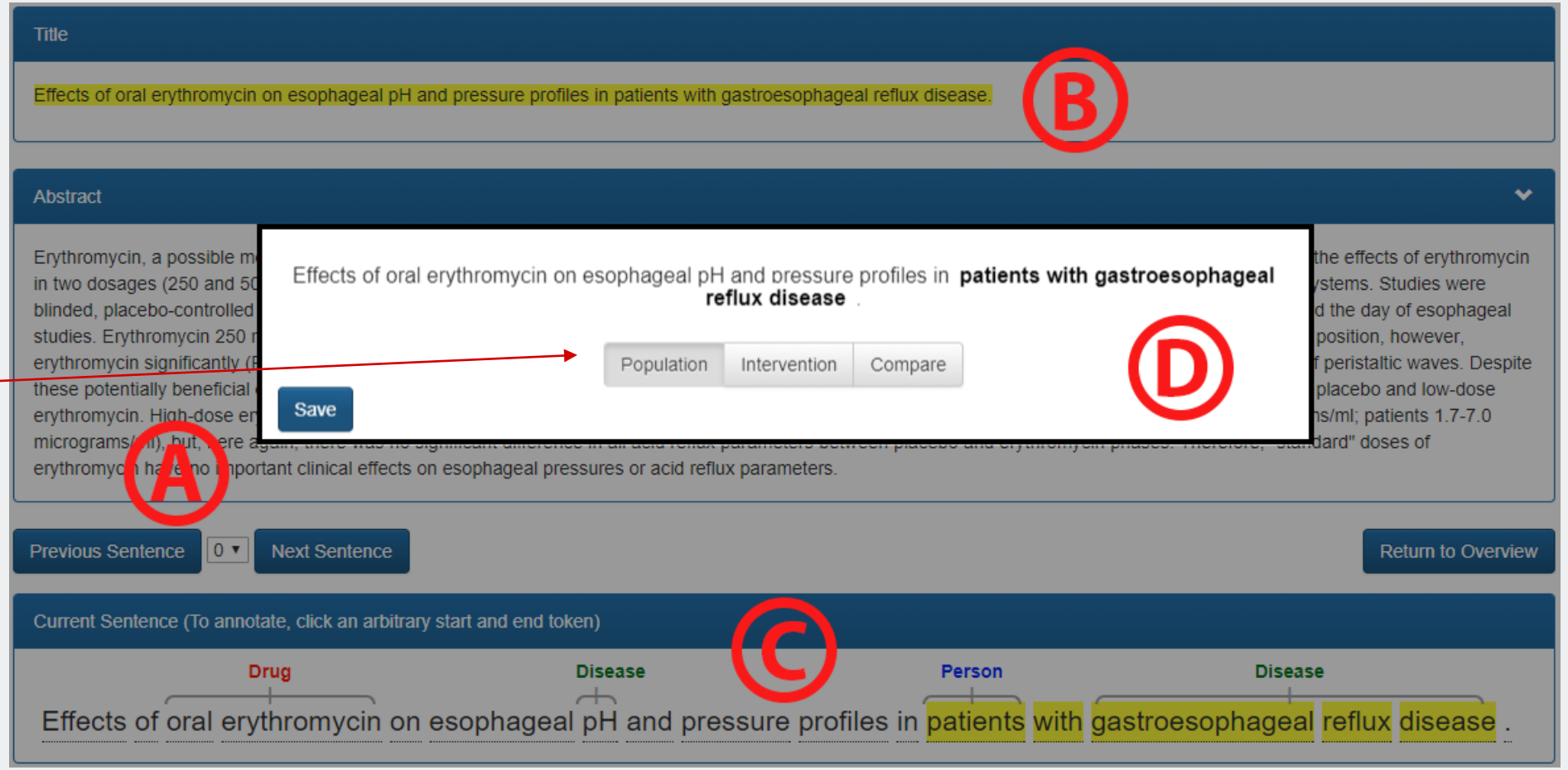
1. Select a sentence. The sentence is shown below with an annotation from a medical pipeline (disease, drug, anatomy, person, …)

2. Click on the start and end of a concept



TU WIEN Informatics

3. Select the correct class and “Save”



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)

Person Disease Disease

Eleven children with end stage renal failure and anaemia (haemoglobin concentration < 90 g/l) were enrolled into a

Anatomy Disease

single blind , placebo controlled , crossover study to assess the cardiovascular effects of reversing anaemia using

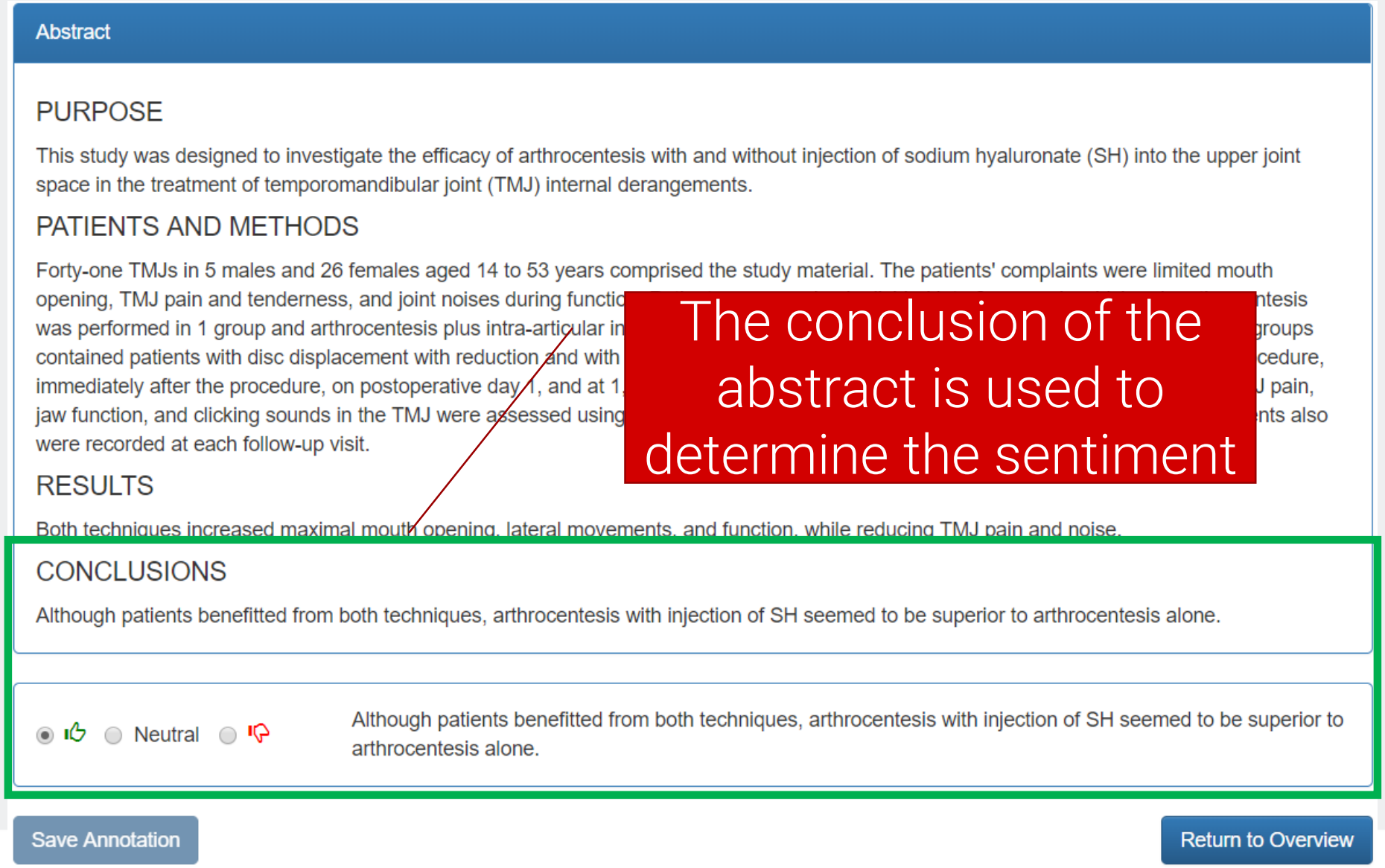
Person Drug

subcutaneous human recombinant erythropoietin (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)



TU WIEN Informatics

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 p neumonia (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

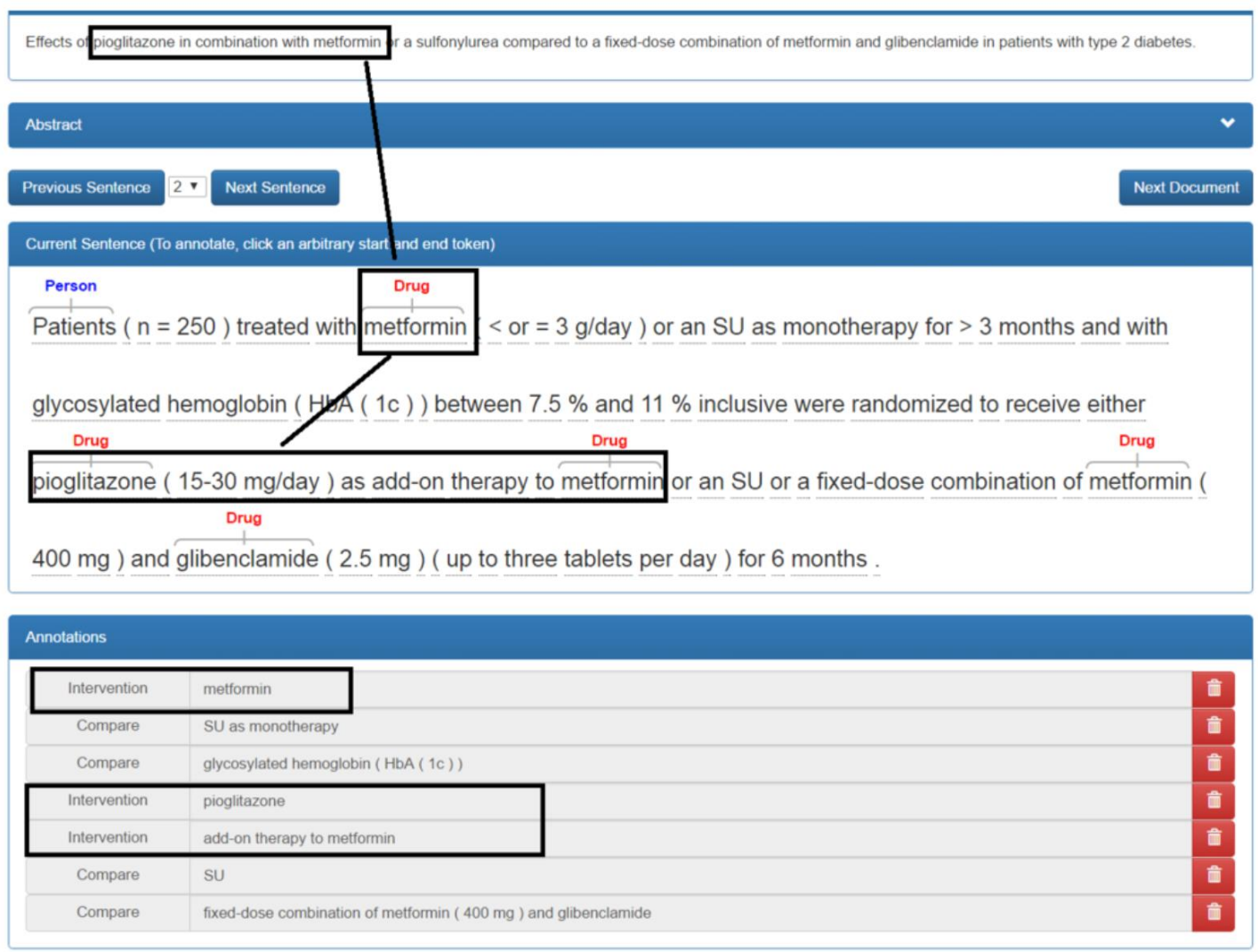


Figure 4: Example of co-reference occurrence of INTERVENTION element in title and abstract.

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)**.*

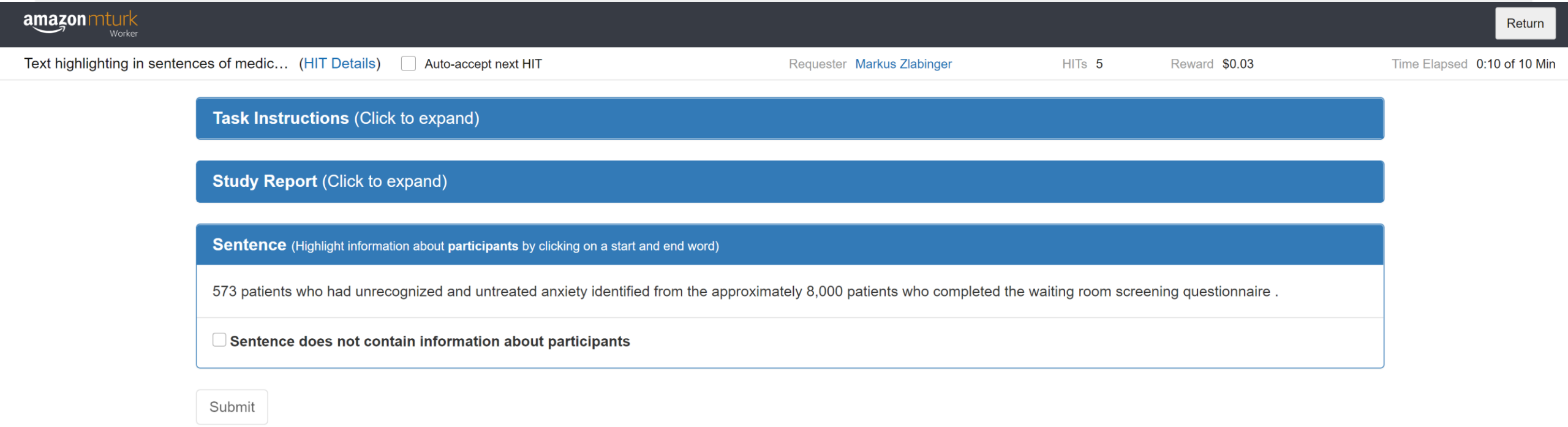
***A total of 270 patients with MCI** were enrolled in a 24-week, multicenter, randomized, double-blind, placebo-controlled study.*

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.*

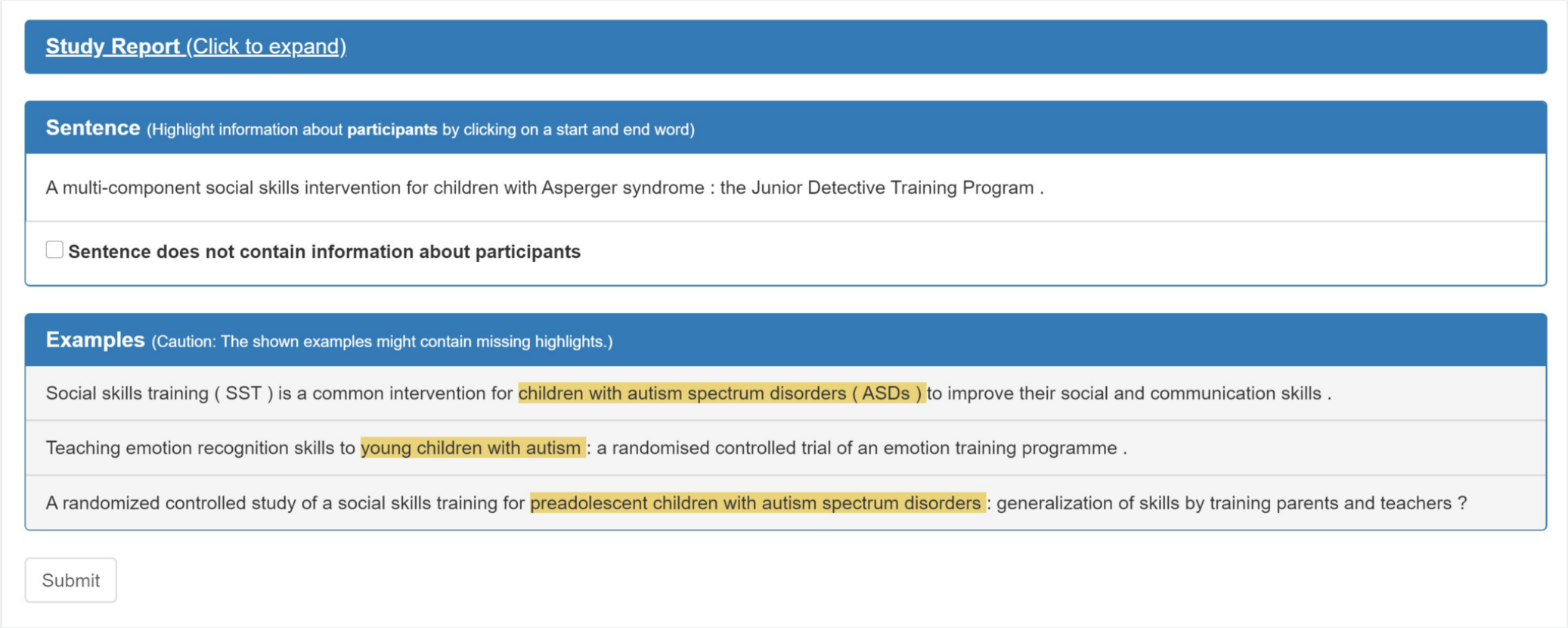
Refinement 1

- Annotate on a sentence level



Refinement 2

- Give annotators dynamic examples of expert annotations



Inter-Annotator Agreement

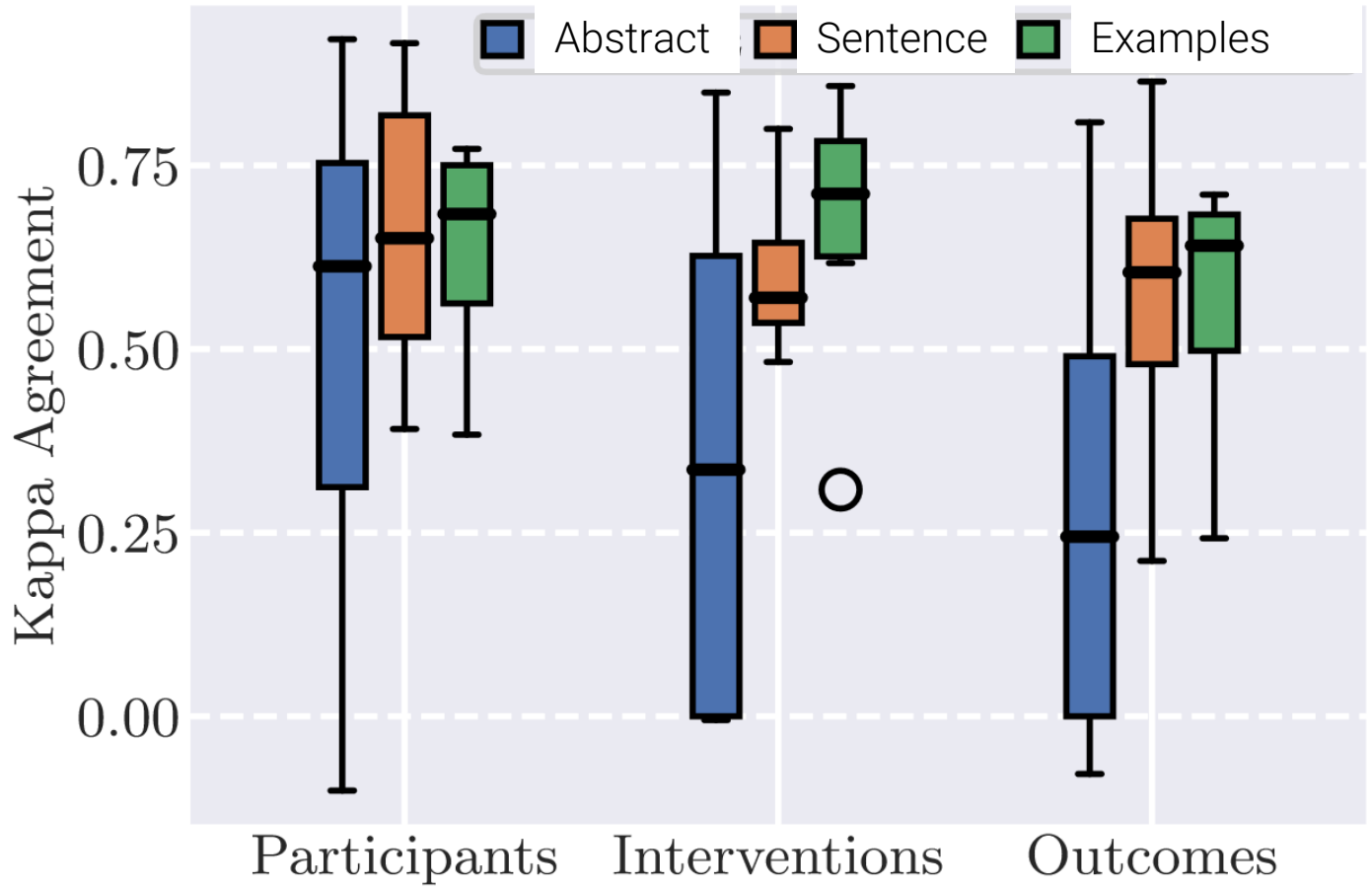


Figure 1: Kappa agreements between annotations from individual workers and the gold standard.

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

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.

κ	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