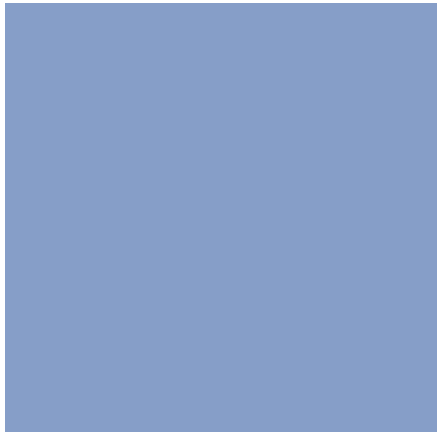


January 16, 2018
Professor Meteer

Models and Specifications

Example: The SWBD Dysfluency Corpus

CS140b NL Annotation for ML

First steps towards your annotation projects

- Select your annotation goal
  - 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?
- What will the annotation be used for?
  - Database population, linguistic analysis, summarization, timeline creation…
- How in-depth will your classification be?
- What aspects of the text will help you with your classification?
  - All the text in the document?
  - Individual words?
  - Relations between words? Between documents?

Switchboard Dysfluencies

- What were the goals of the annotation?
- How was the corpus selected?
- Why was it an appropriate corpus for the goals?

Goal → Model

- Model = <T,R,I>
  - Terms, relations, interpretations
- Terms = classifications applied to the data
  - Ex: “Spam”, “not-spam”, parts-of-speech
- Relations = connections between terms
  - Ex: link between a word and its definition
- Interpretation = what the metadata means
  - How to interpret the annotation

+ Elements

- Sentence breaks
  - Complete sentences
  - Incomplete sentences
  - Interrupted sentences

A: You interested in woodworking? /
A: we did get the wedge cut out by building some kind of--
B: A cradle for it. /
A: -- a cradle for it. /

A: Right / Right /

B: what I've seen of this kind before is you have the, -/ if you're looking at adding on you have, -/

Nonsentence elements

- Filled pause
- Editing term
- Discourse marker
- Conjunction
- Aside

- {F Oh },yeah. / Uh_huh. /
- {F Oh,} yeah,/ { F uh,} the whole thing was small and, [you, + {E I mean, } you] actually put it on
- {P Well }, we have a cat who's also about four years old. /
- {C and then } I painted, {F uh }, about eight different, {F uh }, colors, /
- I, {Fuh }, talked about how a lot of the problems they have to [come, + overcome] [to, + {F uh, } {A it's a very complex, {F uh, } situation } to] go into space. /

+ Restarts

- Restarts
  - Reparandum
  - interregnum
  - Interruption point
  - Repair
  - Note: Complex restarts involved embedding

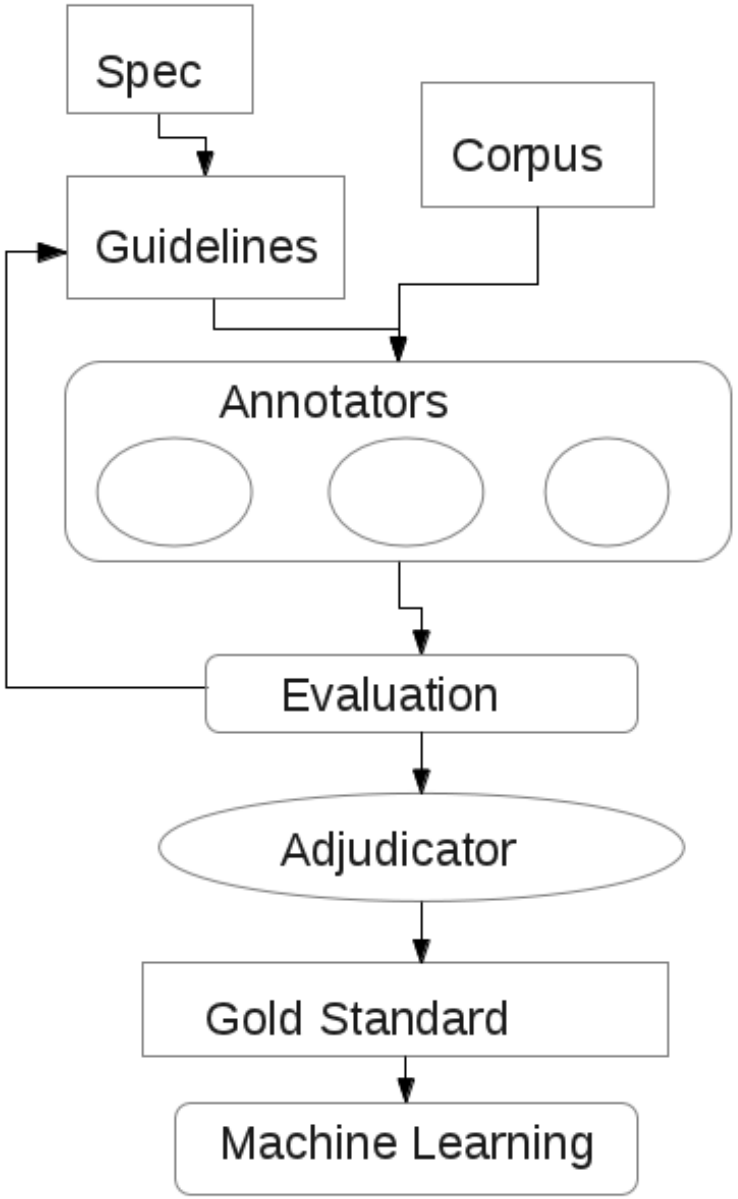
[it, + the instructions] in the book I had said use a coping saw but there's no coping saw big enough [to, + for] a fourteen inch wide watermelon /

B: Yeah, / [[they're, + Um you know they're] like Ber- ,+
A: Dress shorts. /
B: they're like black corduroy [Ber-, + Bermuda]] shorts.

Model → Specification

- Specification = full description of the tags and attributes in your annotation
- Ex: TLINKs can indicate different temporal orderings:
  - <!ATTLIST TLINK relType (
  - BEFORE | AFTER | INCLUDES | IS_INCLUDED | DURING | DURING_INV | SIMULTANEOUS | IAFTER | IBEFORE | IDENTITY | BEGINS | ENDS | BEGUN_BY | ENDED_BY) #REQUIRED >

+ Annotation Process



Groups Contract

- Select your annotation goal
- Broadly define tasks
  - Use the outline of deliverables
  - choose task leads
- Group Contract
  - Can be simple and point to a project plan
    - https://sites.google.com/site/cs216group2anaphora/contract
  - Can include a broad outline of all the steps
    - https://docs.google.com/a/brandeis.edu/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxicmFuZGVpc3NhcmNhc21sfGd4OjRjYjliZDBlMWQ4NDEyZGU
- Submit via Latte (one per group)
  - Annotation goal
  - Group contract