Human Annotation in NLP
Data annotation - a major part of NLP

- In research and industrial development
- Annotation is a major part of AI, for example:
  - Object recognition in images, videos
    - Autonomous/assisted driving
  - Speech recognition
- NLP annotations of
  - Linguistic notions
    - POS, parsing, semantic roles
  - Applied tasks - generic or specific for (industrial) domains/settings
    - NER, question answering, relation/information extraction, dialog
- Annotations needed since supervised methods typically work best
- Needed whenever defining new tasks, or applying tasks to specific variants and domains
Nature of NLP annotation tasks

● In NLP - annotations are often “soft” - there may be some subjectivity
  ○ In task definition
    ■ NER - which name types to define
    ■ RE - which relations are of interest
  ○ In annotation
    ■ People may disagree on linguistic notions (parsing) and semantic/applied ones
      ● E.g. PP attachment parsing, which role is the right one, whether a relation is inferred from the text

● In applied task - a gap between the applied “need” and the annotation definition

● Annotation guidelines - a crucial step to bridge this gap
Examples for guidelines issues

- “I want to detect company acquisitions events”
  - document or sentence level?
  - what is an acquisition?
  - Trigger mention-level vs. whole event-level - with arguments, boundaries/spans

- Text classification: what are the labels? are all distinctions meaningful / easy to do?
- Generally - need to define the right set classes
Annotator types: “Expert”

- Traditionally, NLP annotations were done by “expert” annotators
- For example - POS, Syntactic Tree-bank, SRL
  - Hiring students with linguistic expertise
  - Training and testing them over the annotation guidelines
- In applied tasks, like QA/RE - good language speakers may suffice
  - No need in linguistic expertise
  - Training still required, to comprehend guidelines
- Running expert annotation is usually an expensive project, requiring substantial management
  - Recruit, train, manage, payment administration
Annotator types: Crowdsourcing

- The general notion of crowdsourcing
- Crowdsourcing human intelligence tasks
  - Most famous service: Amazon Mechanical Turk
  - Additional services exist

Figure Eight Inc.

From Wikipedia, the free encyclopedia

Figure Eight (formerly known as Dolores Lab, CrowdFlower) is a human-in-the-loop machine learning and artificial intelligence company based in San Francisco. The company raised $58 million in venture capital and was acquired by Appen and Five River Group of Industries\(^1\) in March 2019 for $300 million\(^2\).
Crowdsourcing for NLP

- Limited to simple annotation tasks
  - Task is typically explained via help/instructions within the annotation interface
  - Platform provides somewhat simple interface options, or can link to a targeted application
- If needed: can break annotation to a pipeline of simple tasks
- May pose qualification criteria
  - Language, country, level of experience, number of “approved tasks” (by employer)
- Risk of spammers - may apply mechanisms to filter them
- Projects are much easier to manage and scale
  - Leverage a large pool of workers
  - Fast project duration
  - Easy to manage - leveraging the crowdsourcing platform
“Controlled Crowdsourcing” - a midpoint

● “Trap task” to recruit high-quality workers
● Contact workers individually, to conduct an “expert-like” annotation
  ○ More elaborate guidelines (but not too long)
  ○ Training and testing phase
  ○ Individual feedback
  ○ Paying for each step
● Allows raising the complexity level of crowdsourced annotation
Addressing annotation quality

- Often obtain multiple annotation for each instance
- Address disagreements by majority vote, possibly weighted (e.g. MACE), or arbitration.
- Measure quality by inter-annotator agreement
  - Between individual annotators - indicates subjectivity level of task
  - Between final annotations by “teams” - across teams