Applied Natural Language Processing

Info 256

Lecture 15: Annotating data (March 14, 2019)

David Bamman, UC Berkeley

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995;1999); morphosyntactic annotations of WSJ

- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference

- FrameNet (1998–): frame-semantic lexica/annotations

- MPQA (2005): opinion/sentiment

- SQuAD (2016): annotated questions + spans of answers in Wikipedia

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of human judgments.
  - What's the correct part of speech tag?
  - Syntactic structure?
  - Sentiment?

Ambiguity

"One morning I shot
an elephant in my pajamas"

Animal Crackers

Dogmatism

Fast and Horvitz (2016),
"Identifying Dogmatism in Social Media: Signals and Models"

Given a comment, imagine you hold a well-informed, different opinion from the commenter in question. We'd like you to tell us how likely that commenter would be to engage you in a constructive conversation about your disagreement, where you each are able to explore the other's beliefs. The options are:

(5): It's unlikely you'll be able to engage in any substantive conversation. When you respectfully express your disagreement, they are likely to ignore you or insult you or otherwise lower the level of discourse.

(4): They are deeply rooted in their opinion, but you are able to exchange your views without the conversation degenerating too much.

(3): It's not likely you'll be able to change their mind, but you're easily able to talk and understand each other's point of view.

(2): They may have a clear opinion about the subject, but would likely be open to discussing alternative viewpoints.

(1): They are not set in their opinion, and it's possible you might change their mind. If the comment does not convey an opinion of any kind, you may also select this option.

Sarcasm

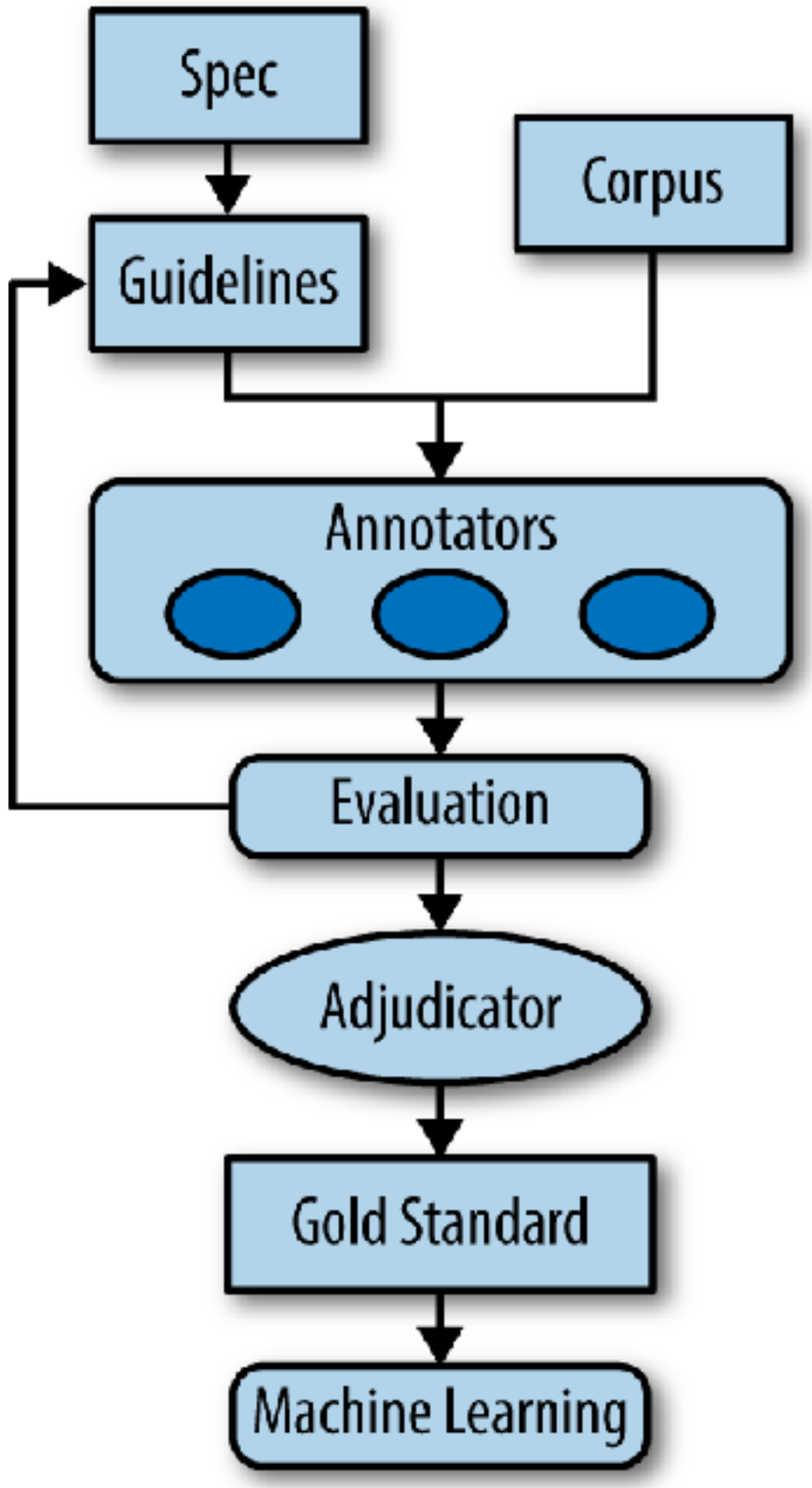
"In many respects you know they honor President Obama. ISIS is honoring President Obama! He is the founder of ISIS. He's the founder of ISIS, O.K.! He's the founder, he founded ISIS and I would say the co-founder would be crooked Hillary Clinton. Co-founder, crooked Hillary Clinton. And that's what it's about."

https://www.nytimes.com/2016/08/12/opinion/an-even-stranger-donald-trump.html?ref=opinion

Fake News

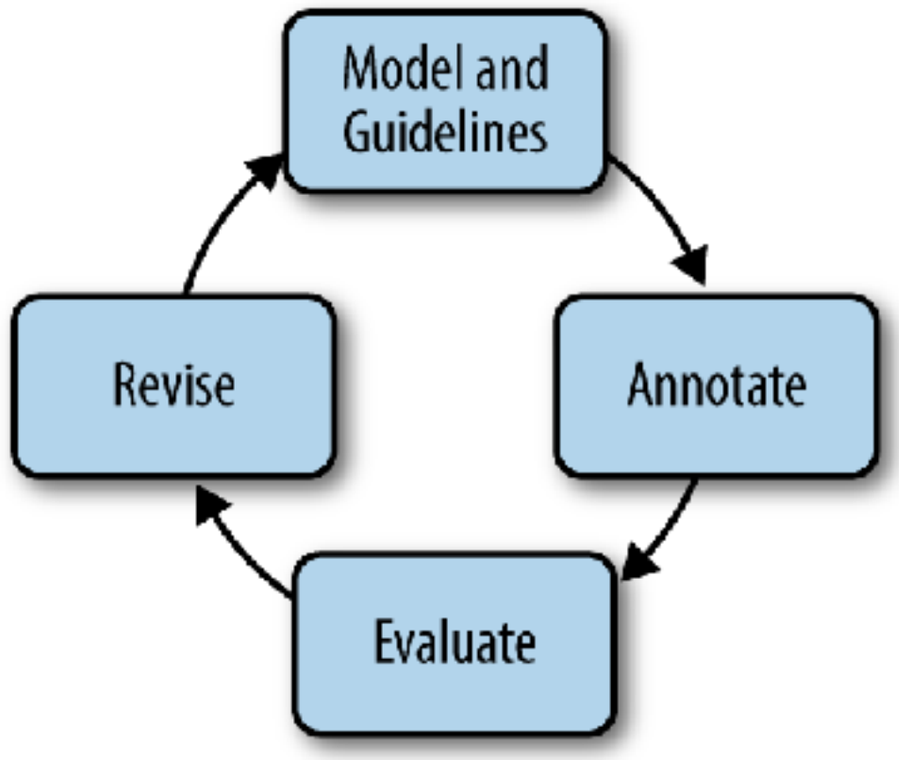
http://www.fakenewschallenge.org

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation Guidelines

- Our goal: given the constraints of our problem, how can we formalize our description of the annotation process to encourage multiple annotators to provide the same judgment?

Annotation guidelines

- What is the goal of the project?
- What is each tag called and how is it used? (Be specific: provide examples, and discuss gray areas.)
- What parts of the text do you want annotated, and what should be left alone?
- How will the annotation be created? (For example, explain which tags or documents to annotate first, how to use the annotation tools, etc.)

Pustejovsky and Stubbs (2012), Natural Language Annotation for Machine Learning

Practicalities

- Annotation takes time, concentration (can't do it 8 hours a day)

- Annotators get better as they annotate (earlier annotations not as good as later ones)

Why not do it yourself?

- Expensive/time-consuming
- Multiple people provide a measure of consistency: is the task well enough defined?
- Low agreement = not enough training, guidelines not well enough defined, task is bad

Adjudication

- Adjudication is the process of deciding on a single annotation for a piece of text, using information about the independent annotations.

- Can be as time-consuming (or more so) as a primary annotation.

- Does not need to be identical with a primary annotation (both annotators can be wrong by chance)

Interannotator agreement

annotator A

annotator B

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

https://twitter.com/teenybiscuit/status/705232709220769792/photo/1

Cohen’s kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

annotator B \ annotator A	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen’s kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

annotator A

annotator B

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen's kappa

- Expected probability of agreement is how often we would expect two annotators to agree assuming independent annotations

$$p_e = P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken})$$

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

Cohen's kappa

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

P(A=puppy)	15/100 = 0.15
P(B=puppy)	11/100 = 0.11
P(A=chicken)	85/100 = 0.85
P(B=chicken)	89/100 = 0.89

$$= 0.15 \times 0.11 + 0.85 \times 0.89$$
$$= 0.773$$

annotator A

annotator B

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773}$$

$$= 0.471$$

annotator A

annotator B

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen’s kappa

- “Good” values are subject to interpretation, but rule of thumb:

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement

annotator A

annotator B

	puppy	fried chicken
puppy	0	0
fried chicken	0	100

annotator A

annotator B

	puppy	fried chicken
puppy	50	0
fried chicken	0	50

annotator A

annotator B

	puppy	fried chicken
puppy	0	50
fried chicken	50	0

Interannotator agreement

- Cohen's kappa can be used for any number of classes.

- Still requires two annotators who evaluate the same items.

- Fleiss' kappa generalizes to multiple annotators, each of whom may evaluate different items (e.g., crowdsourcing)

Fleiss' kappa

- Same fundamental idea of measuring the observed agreement compared to the agreement we would expect by chance.

- With N > 2, we calculate agreement among pairs of annotators

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Fleiss' kappa

Number of annotators who assign category *j* to item *i*

$$n_{ij}$$

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Fleiss' kappa

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Annotator

A	B	C	D
+	+	+	-

Label	n_{ij}
+	3
-	1

agreeing pairs of annotators →

A-B
B-A
A-C
C-A
B-C
C-B

$$P_i = \frac{1}{4(3)}(3(2) + 1(0))$$

Fleiss' kappa

Average agreement among all items

$$P_o = \frac{1}{N} \sum_{i=1}^{N} P_i$$

Probability of category *j*

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

Expected agreement by chance — joint probability two raters pick the same label is the product of their independent probabilities of picking that label

$$P_e = \sum_{j=1}^{K} p_j^2$$

Krippendorf's alpha

- Kappa values still require categorical labels

- What about real-valued labels (e.g., Likert ratings, ordinal values)?

Krippendorf's alpha

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

- We'll use the same principle that we used before: how much do our observed labels for a document differ from what we'd expect given the ratings we see?

- For real-valued ratings, we will also use distance metric to quantify how different two ratings are.

Observed

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

when one annotator gives this label…

… how often do we see another with this label for that same item?

	1	3	4	5
1		1		
3	1			
4				2
5			2	2

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

rating	count
1	1
3	1
4	2
5	4

Given this distribution of ratings overall, how often would we expect to see a pair of ratings together?

$$P(r_1 = 5) = 4/8$$

$$P(r_2 = 1) = 1/7$$

normalize over 7 now instead of 8 because we already selected one

$$P(r_1 = 5, r_2 = 1) = 4/8 \times 1/7 = 1/14$$

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

rating	count
1	1
3	1
4	2
5	4

$$P(r_1 = 1, r_2 = 1) = ?$$

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

when one annotator gives this label…

… what's the probability of seeing another with this label for that same item?

	1	3	4	5
1	0	1/56	3/56	1/14
3	1/56	0	1/28	1/14
4	1/28	1/28	1/28	1/7
5	1/14	1/14	1/7	3/14

Expected

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

Transform these into expected counts by multiplying by the total number of annotations (8)

when one annotator gives this label…

… how often do we expect to see another with this label for that same item?

	1	3	4	5
1	0	1/7	3/7	4/7
3	1/7	0	2/7	4/7
4	2/7	2/7	2/7	8/7
5	4/7	4/7	8/7	12/7

Distance

	doc 1	doc 2	doc 3	doc 4
annotator A	5	5	5	1
annotator B	4	5	4	3

For real labels, we can use the squared distance as a measure of cost.

$$(r_1 - r_2)^2$$

	1	3	4	5
1	0	4	9	16
3	4	0	1	4
4	9	1	0	1
5	16	4	1	0

Krippendorf's alpha

observed

distance

$$1 - \frac{\text{sum}\left[\begin{array}{|c|c|c|c|c|}\hline & 1 & 3 & 4 & 5 \\ \hline 1 & & 1 & & \\ \hline 3 & 1 & & & \\ \hline 4 & & & & 2 \\ \hline 5 & & & 2 & 2 \\ \hline\end{array} \times \begin{array}{|c|c|c|c|c|}\hline & 1 & 3 & 4 & 5 \\ \hline 1 & 0 & 4 & 9 & 16 \\ \hline 3 & 4 & 0 & 1 & 4 \\ \hline 4 & 9 & 1 & 0 & 1 \\ \hline 5 & 16 & 4 & 1 & 0 \\ \hline\end{array}\right]}{\text{sum}\left[\begin{array}{|c|c|c|c|c|}\hline & 1 & 3 & 4 & 5 \\ \hline 1 & 0 & 1/7 & 3/7 & 4/7 \\ \hline 3 & 1/7 & 0 & 2/7 & 4/7 \\ \hline 4 & 2/7 & 2/7 & 2/7 & 8/7 \\ \hline 5 & 4/7 & 4/7 & 8/7 & 12/7 \\ \hline\end{array} \times \begin{array}{|c|c|c|c|c|}\hline & 1 & 3 & 4 & 5 \\ \hline 1 & 0 & 4 & 9 & 16 \\ \hline 3 & 4 & 0 & 1 & 4 \\ \hline 4 & 9 & 1 & 0 & 1 \\ \hline 5 & 16 & 4 & 1 & 0 \\ \hline\end{array}\right]}$$

expected

distance

Implementation

- https://www.nltk.org/api/nltk.metrics.html

Activity

- Form groups of 3
- You'll be assigned an annotation task (suspense or subjectivity)
- Your task:
    - Annotate the data we provided (~150 sentences)
    - Do so independently, and find where you disagree.
    - Create a set of annotation guidelines that you can pass off to another group to best recreate your judgments.

Activity

- Suspense
  - Label how suspenseful a sentence is in the context of a narrative. Assign each sentence an interger rating from 0 (not suspenseful) to 4 (most suspenseful).
- Subjectivity
  - Label whether each sentence is objective (presenting a disinterested set of facts) or subjective (presenting opinion or sentiment).