Inter-Annotation Agreement

COSI 140 – Natural Language Annotation for Machine Learning

James Pustejovsky

February 23, 2016

Brandeis University

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Corpus Reliability

- Supervised techniques depend on annotated corpus.
- For appropriate modeling of a natural phenomena the annotated corpus should be reliable.
- The recent trend is to annotate corpus with more than one annotator and measure agreement.
- Agreement measure/coefficient of reliability.

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Existing Reliability Measures

- Cohen's Kappa (Cohen, 1960)
- Scott's π (Scott, 1955)
- Krippendorff's α (Krippendorff, 1980)
- Rosenberg and Binkowski, 2004
  - Annotation limited to two categories

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Motivation

- Affect corpus: Annotation may be fuzzy and one text segment may belong to multiple categories

DISGUST

SAD

"A young married woman was burnt to death allegedly by her in-laws for dowry."

- The existing measures are applicable to single class annotation.

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Affective Text Corpus and Annotation

- Consists of 1000 sentences collected from news headlines and articles in *Times of India (TOI)* archive.
- Affect classes → Set of basic emotions [P. Ekman]
  - Anger, disgust, fear, happiness, sadness, surprise

	Anger	Disgust	Fear	Happy	Sad	Surprise
U1	0	1	0	0	0	1
U2	0	0	0	1	0	1

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

A_m Agreement Measure and Reliability

- Features of A_m
  - Handles multi-class annotation
  - Non-inclusion in a category is also considered as agreement.
  - Inspired by Cohen's Kappa and is formulated as

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

where P_o is the *observed agreement* and P_e is the *expected agreement.*

  - Considers category pairs while computing P_o and P_e.

Notion of Paired Agreement

- For an item, two annotators *U1* and *U2* are said to agree on category pair <*C1, C2*> if

$$U1.C1 = U2.C1$$

$$U1.C2 = U2.C2$$

where Ui.Cj signifies that the value for Cj for annotator Ui and the value may either be 1 or 0.

	Anger	Fear
UI	0	I
U2	0	I

Example Annotation

Sen	Judge	A	D	S	H
1	U1	0	1	1	0
	U2	0	1	1	1
2	U1	1	0	1	0
	U2	0	1	1	0
3	U1	0	0	1	0
	U2	1	0	1	0
4	U1	1	0	1	1
	U2	1	0	1	0

A → Anger
D → Disgust
F → Sadness
H → Happiness

Computation of P_o

- U = 2, C = 4, I = 4
- The total agreement on a category pair p for an item i is n_{ip}, the number of annotator pairs who agree on p for i.

	A-D	A-S	A-H	D-S	D-H	S-H
n_{Ip}	I	I	0	I	0	0

- The average agreement on a category pair p for an item i is

	A-D	A-S	A-H	D-S	D-H	S-H
P_{Ip}	1.0	1.0	0.0	1.0	0.0	0.0

$$P_{ip} = \frac{1}{\binom{U}{2}} n_{ip}$$

Computation of P_o (Cont…)

- The average agreement for the item i is

$P_1 = 0.5$

$$P_i = \frac{1}{\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{p \in \mathbf{S}} n_{ip}$$

- Similarly, $P_2 = 0.57$, $P_3 = 0.5$, $P_4 = 1$
- The observed agreement is

Po = 0.64

$$\begin{aligned} P_o &= \frac{1}{\mathbf{I}} \sum_{i=1}^{I} P_i \\ &= \frac{1}{\mathbf{I}\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \\ &= \frac{4}{\mathbf{IC}(\mathbf{C}-1)\mathbf{U}(\mathbf{U}-1)} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \end{aligned}$$

Computation of P_e

- Expected agreement is the expectation that the annotators agree on a category pair.
- For a category pair, possible assignment combinations

  G = {[0 0], [0 1], [1 0], [1 1]}

Computation of P_e (Cont….)

- Overall proportion of items assigned with assignment combination g $\in$ G to category pair p by annotator u is

$$\hat{P}(p_g|u) = \frac{n_{p_g u}}{\mathbf{I}}$$

	0-0	0-1	1-0	1-1
A-D (U1)	¼ = 0.25	¼ = 0.25	2/4 = 0.5	0/4 = 0.0
A-D (U2)	0/4 = 0.0	2/4 = 0.5	2/4 = 0.5	0/4 = 0.0

Computation of P_e (Cont….)

- The probability that two arbitrary coders agree with the same assignment combination in a category pair is

$$\hat{P}(p_g) = \frac{1}{\binom{\mathbf{U}}{2}} \sum_{(u_x,u_y)\in W} \hat{P}(p_g|u_x)\hat{P}(p_g|u_y)$$

	0-0	0-1	1-0	1-1
A-D	0.0	0.125	0.25	0.0

Computation of P_e (Cont….)

- The probability that two arbitrary annotators agree on a category pair for all assignment combinations is

A-D	A-S	A-H	D-S	D-H	S-H
0.375	0.5	0.25	0.5	0.375	0.623

$$\hat{P}(p) = \sum_{g \in G} \hat{P}(p_g)$$

- The chance agreement is

$P_e = 0.46$

$$P_e = \frac{1}{\binom{C}{2}} \sum_{p \in S} \hat{P}(p)$$

- $A_m = 0.33$

Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Gold Standard Determination

- Majority decision label is assigned to an item.
- Expert Coder Index of one annotator indicates how often he agrees with others.
- Expert Coder Index is used when there is no majority of any class for an item.

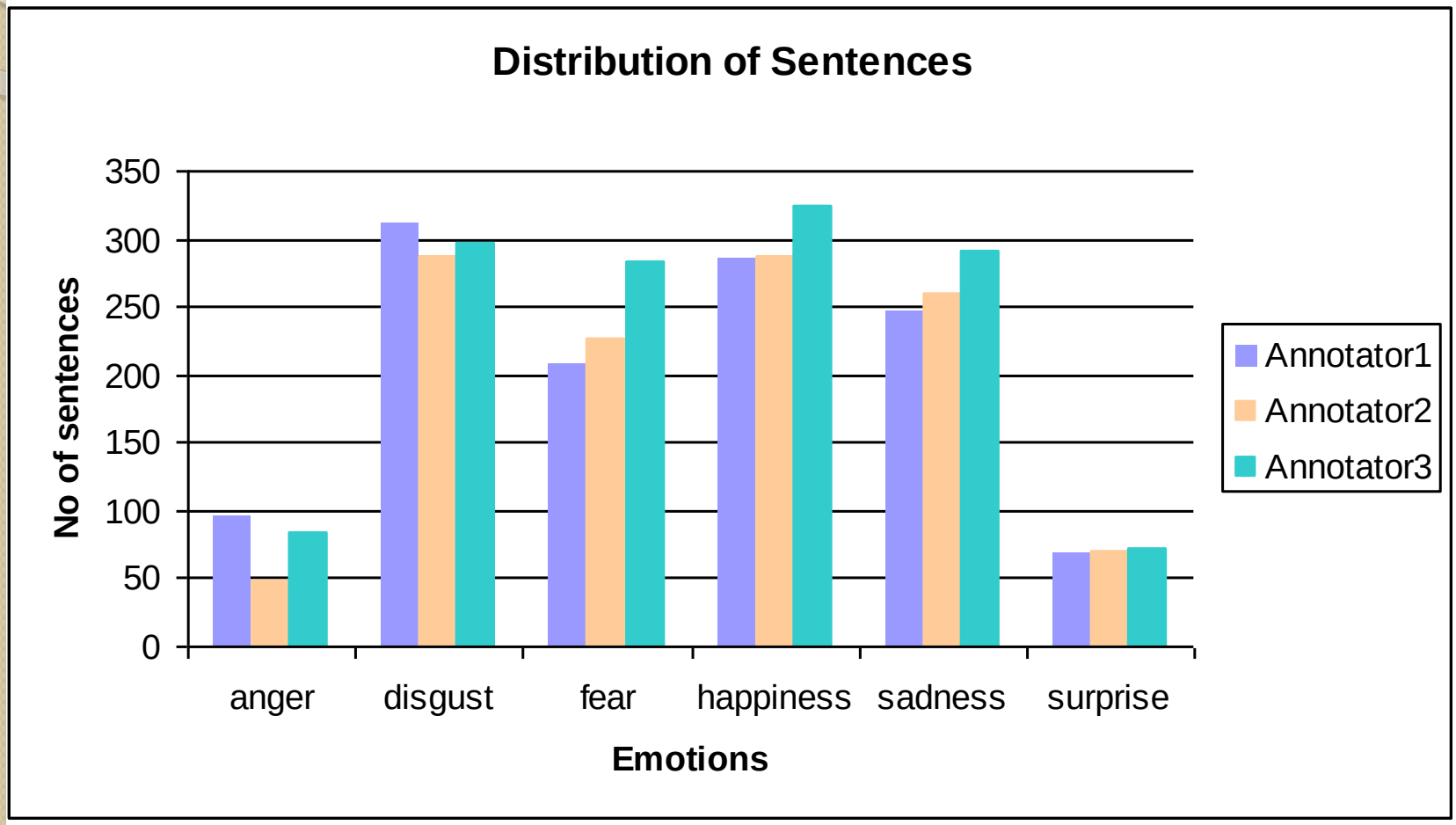
Outline

- Corpus Reliability
- Existing Reliability Measures
- Motivation
- Affective Text Corpus and Annotation
- A_m Agreement Measure and Reliability
- Gold Standard Determination
- Experimental Results
- Conclusion

Annotation Experiment

- Participants: 3 human judges
- Corpus: 1000 sentences from TOI archive
- Task: annotate sentences with affect categories.
- Outcome: Three human judges were able to finish within 20 days.
- We report results based on data provided by three annotators.

Annotation Experiment (Cont….)



Analysis of Corpus Quality

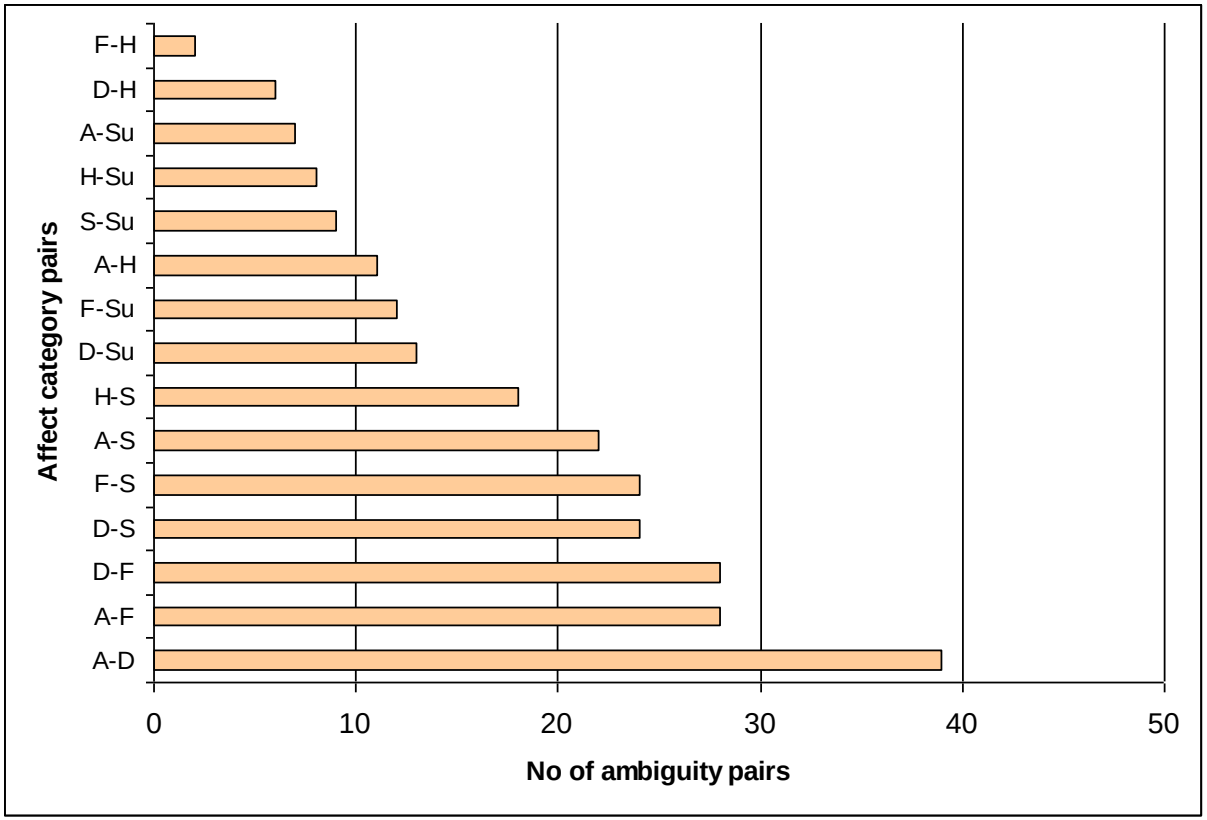
- Agreement Value

Agreement	A_m **Value**
Observed Agreement(P_o)	0.878
Chance Agreement(P_e)	0.534
A_m	**0.738**

- Agreement study
  - 71.5% of the corpus belongs to [0.7 1.0] range of observed agreement and among this portion, the annotators assign 78.6% of the sentences into a single category.
  - For the non-dominant emotions in a sentence, ambiguity has been found while decoding.

Analysis of Corpus Quality (Cont…)

- Disagreement study



A → anger

D → disgust

F → fear

H → happiness

S → sadness

Su → surprise

	X	Y
UI	0	I
U2	I	0

Analysis of Corpus Quality (Cont…)

- Category pair with maximum confusion is [anger disgust]
- Anger and disgust are close to each other in the evaluation-activation model of emotion.
- *anger*, *disgust* and *fear* are associated with three topmost ambiguous pairs.

Gold Standard Data