Analysing and Using Crowd Annotations

Silviu Paun

Friday 24th September 2021

DALI Project

My role in the project

- Help make the Phrase Detectives resources more accessible
- Help inform the NLP community about best practices for annotation analysis
- Help the team with machine learning expertise

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio
TACL 2018

Example of an annotation task

- Recognizing textual entailment
  - Coders are presented with two sentences and asked to judge whether the second sentence, called a hypothesis, can be inferred from the first
- A positive case of textual entailment:
  - Premise: “Crude Oil Prices Slump.”
  - Hypothesis: “Oil prices drop.”
- A case of false entailment:
  - Premise: “The government announced last week that it plans to raise oil prices.”
  - Hypothesis: “Oil prices drop.”

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio

TACL 2018

Number of positive annotations	Number of items	Number of items per gold class	
		Negative class	Positive class
0	5	5	0
1	25	25	0
2	90	90	0
3	105	102	3
4	103	92	11
5	65	50	15
6	62	18	44
7	59	11	48
8	108	3	105
9	105	4	101
10	73	0	73
Total	800	400	400

A summary of the annotation patterns from the Recognizing Textual Entailment (RTE) dataset (Snow et al., 2008)

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio
TACL 2018

Probabilistic models of annotation

- Simply put, a model of annotation is a probabilistic framework for distilling the disagreement between coders from noisy interpretations
- We can specify our assumptions about the annotation process, e.g., the interactions between the items and the coders
- Our assumptions are then considered when inferring the corpus labels

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio
TACL 2018

Dawid and Skene (1979)

- Each item has a true class whose prior probability is given by the prevalence of the true classes in the corpus:

$$c_i \sim Categorical(\pi)$$

- The model assumes that each annotation is produced according to its coder's annotation behavior associated with the true class:

$$y_{i,n} \sim Categorical(\zeta_{jj[i,n],c_i})$$

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio

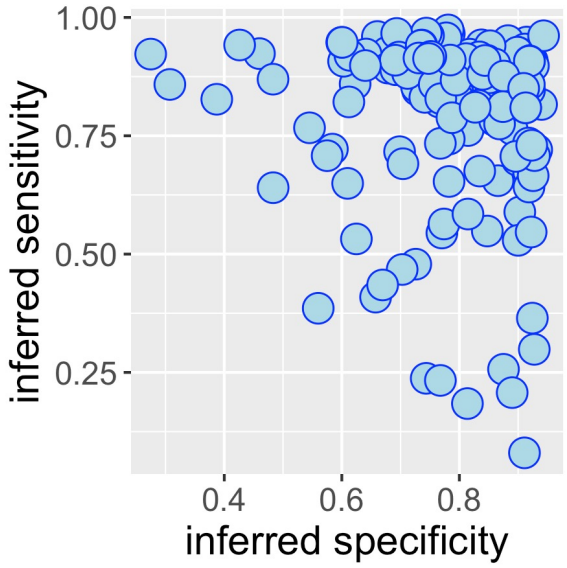
TACL 2018

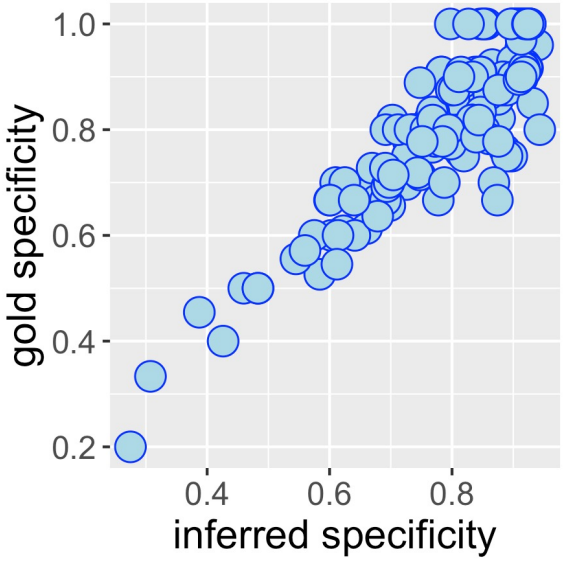
Example of a typically found annotator	Gold* confusion	Dawid and Skene (1979)
the good annotator	$\begin{pmatrix} 0.88 & 0.12 \\ 0.25 & 0.75 \end{pmatrix}$	$\beta_{17} = \begin{pmatrix} 0.81 & 0.19 \\ 0.24 & 0.76 \end{pmatrix}$
the highly accurate annotator	$\begin{pmatrix} 0.80 & 0.20 \\ 0.00 & 1.00 \end{pmatrix}$	$\beta_{119} = \begin{pmatrix} 0.94 & 0.06 \\ 0.04 & 0.96 \end{pmatrix}$
the annotator biased towards the first class	$\begin{pmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \end{pmatrix}$	$\beta_{88} = \begin{pmatrix} 0.91 & 0.09 \\ 0.92 & 0.08 \end{pmatrix}$
the annotator biased towards the second class	$\begin{pmatrix} 0.20 & 0.80 \\ 0.00 & 1.00 \end{pmatrix}$	$\beta_{65} = \begin{pmatrix} 0.28 & 0.72 \\ 0.08 & 0.92 \end{pmatrix}$

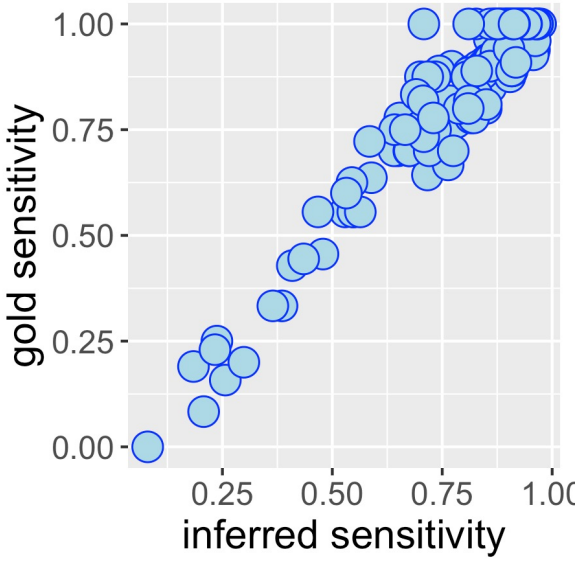
Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio

TACL 2018







Coder estimates for the model of Dawid and Skene (1979) fitted on the RTE dataset

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio
TACL 2018

Disagreements in annotation: item difficulty

- Example of an ‘easy’ judgement:
  - Premise: “The three-day G8 summit will take place in Scotland.”
  - Hypothesis: “The G8 summit will last three days.”
- Example of a ‘difficult’ judgement:
  - Premise: “EU membership is a strategic necessity for Turkey, as Ankara will inevitably suffer greater foreign policy problems in the future unless it makes it into the Union.”
  - Hypothesis: “Turkey to join the EU.”

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio
TACL 2018

Carpenter, 2008

- The probability of an annotator being correct on an item is based on a subtractive relationship between their ability and the difficulty of the item:

$$p(y_{i,n} = c_i) = logistic(\alpha_{jj[i,n]} - \theta_i)$$

- If the item is easy ($\theta_i < 0$) the accuracy of the coders is increased
- In case of a hard item ($\theta_i > 0$) , the difficulty parameter reduces the accuracy of the coders

Comparing Bayesian Models of Annotation

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio

TACL 2018

Positive annotations	No. items	Positive items (gold)	Carpenter (2008)	
			avg. θ_i	pos. items
0	5	0	-1.55	0
1	25	0	-1.03	0
2	90	0	-0.58	0
3	105	3	-0.14	4
4	103	11	0.27	7
5	65	15	0.65	13
6	62	44	0.78	40
7	59	48	0.58	47
8	108	105	0.10	106
9	105	101	-0.36	105
10	73	73	-0.91	73

The average difficulty estimated by the models of Carpenter (2008) for items with a certain amount of disagreement

logistic(-1.55) = 0.18

logistic(0.78) = 0.69

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

The anaphoric annotation task

- In standard anaphoric annotation projects the mentions are predefined for better agreement
- The annotation scheme allows coders to mark a mention as discourse new or as discourse old
- In the latter case the annotators also must specify the entity in question
- Richer annotation schemes allow annotators to also mark, e.g., expletives and predicative noun phrases

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

An example of anaphoric annotation

- [John], [a colleague from work], said [it] will rain later today. [He] was right.
- The annotators should mark:
  - “John” as discourse new
  - “a colleague from work” as a predicative noun phrase
  - “it” as an expletive
  - and the pronoun “he" as a discourse old mention further selecting “John” as its most recent antecedent
    - Mentions “John” and “he” form a coreference chain

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

What is the challenge?

- This looks like a classification task
- Unlike in standard classification tasks, the set of classes the coders can choose from changes depending on the mentions they annotate
- For this reason, standard models of annotation are not immediately applicable to aggregate anaphoric judgements

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

- The annotations are transformed into mention-pairs:

Mention-pair	Type	Coder 1	Coder 2	Coder 3
(“he”, “John”)	discourse old	1	1	0
(“he”, discourse new)	discourse new	0	0	1

- MPA models these mention-pair judgements as the result of the sensitivity and the specificity of the annotators

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

MPA's generative process

- For every mention-pair (i,m) an indicator encodes whether it is correct or not:

$$c_{i,m} \sim Bernoulli(\pi_{z_{i,m}})$$

- If the mention pair is believed to be correct ($c_{i,m} = 1$), then the associated binary judgements are assumed to be the result of the annotators' sensitivity for that type of mention pairs:

$$y_{i,m,n} \sim Bernoulli(\alpha_{jj[i,m,n],z_{i,m}})$$

- When the mention pair is incorrect ($c_{i,m} = 0$) the binary judgements are modelled according to the specificity of the coders:

$$y_{i,m,n} \sim Bernoulli(1 - \beta_{jj[i,m,n],z_{i,m}})$$

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

Observations

- The annotators have a sensitivity and a specificity associated with each type of mention-pairs
- After the parameters have been estimated each mention is assigned the most likely interpretation based on the posterior of the mention-pair indicators
- The coreference chains (entities) can then be built by simply following the link path from the aggregated mention pairs
- The model can also be used in an analysis of anaphoric ambiguity

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

MPA at work

- We used MPA to adjudicate the anaphoric interpretations collected by the *Phrase Detectives* game with a purpose
- The latest version of the released corpus (Poesio et al., 2019) contains at least 8 anaphoric judgements for over 100 thousand mentions from about 540 documents covering 2 main genres, Wikipedia articles and fiction from the Gutenberg collection
- 45 of those documents, containing over 6 thousand mentions, were annotated by linguists to provide a reliable gold standard for evaluation

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

- Estimated mention-pairs evaluated against a gold standard built by linguists

	Majority Vote			MPA (Paun et al., 2018b)		
	Precision	Recall	F1	Precision	Recall	F1
discourse old	**0.94**	0.63	0.75	0.90	**0.87**	**0.89**
discourse new	0.79	**0.99**	0.88	**0.95**	0.96	**0.95**
predicative NPs	0.54	0.10	0.16	**0.64**	**0.72**	**0.68**
expletives	**0.97**	0.71	0.82	0.94	**0.98**	**0.96**
Accuracy	0.83			**0.92**		
Avg. F1	0.66			**0.87**		

A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio
EMNLP 2018

- The quality of various coreference chains evaluated using standard coreference metrics against expert-annotated chains

	Method	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Singletons included	MV	**96.0**	63.9	76.7	**95.7**	78.7	86.4	77.1	**94.9**	85.1	82.7
	MPA	91.6	**82.4**	**86.8**	94.8	**87.8**	**91.2**	**92.4**	93.8	**93.1**	**90.3**
	Stanford	65.4	62.4	63.8	78.9	76.1	77.5	78.4	85.2	81.7	74.3
Singletons excluded	MV	**96.1**	64.8	77.4	**93.8**	45.0	60.8	66.3	48.5	56.1	64.8
	MPA	92.2	**89.2**	**90.7**	88.1	**77.8**	**82.6**	**79.5**	**80.2**	**79.8**	**84.4**
	Stanford	65.7	62.1	63.9	50.3	42.5	46.1	42.7	49.8	46.0	52.0

In conclusion…

- A workshop on “Aggregating and Analysing Crowdsourced Research Annotations for NLP”
  - Silviu Paun, Dirk Hovy
- A tutorial on “Aggregating and Learning from Multiple Annotators”
  - Silviu Paun, Edwin Simpson
- A book on “Statistical Methods for Annotation Analysis”
  - Silviu Paun, Ron Artstein, Massimo Poesio

Thank you!