What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data? (ChaosNLI)

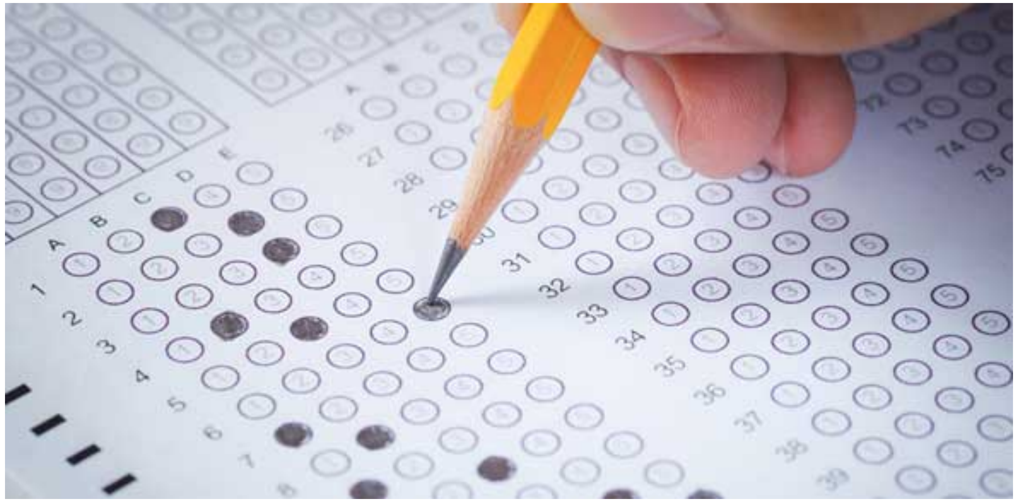
Yixin Nie, Xiang Zhou, Mohit Bansal

UNC NLP

Motivation

UNC NLP

Human Education

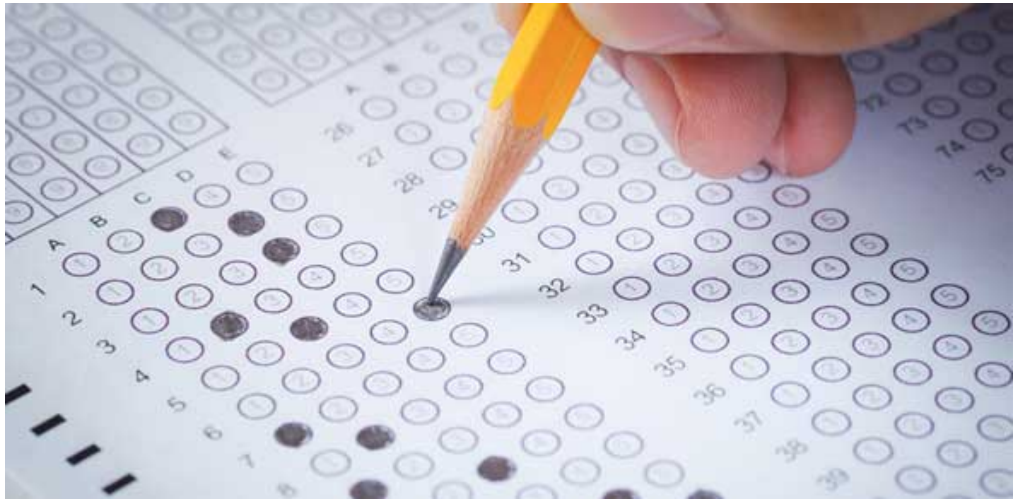


(Human education testing: SAT, GRE, etc.)

- Questions & answers are designed by educators
- Scores are used as certifications or qualifications

Motivation

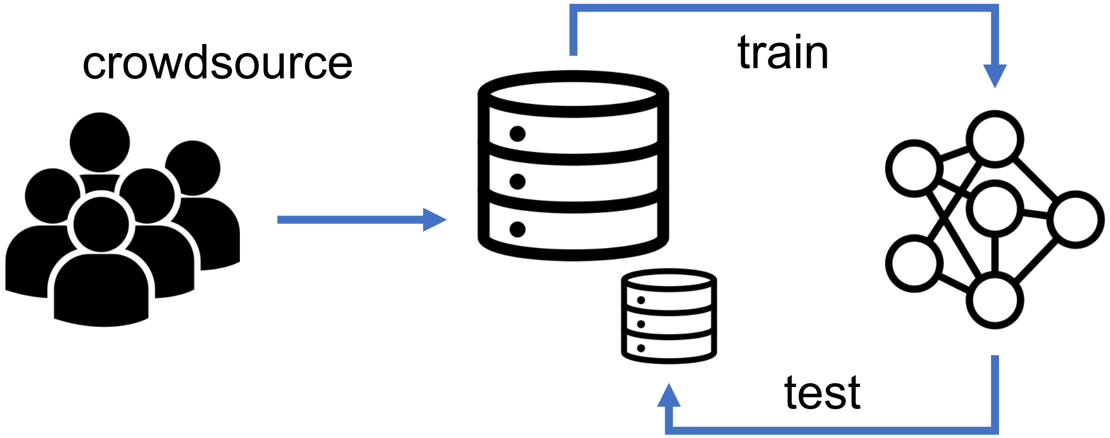
Human Education



(Human education testing: SAT, GRE, etc.)

- Questions & answers are designed by educators
- Scores are used as certifications or qualifications

Natural Language Processing



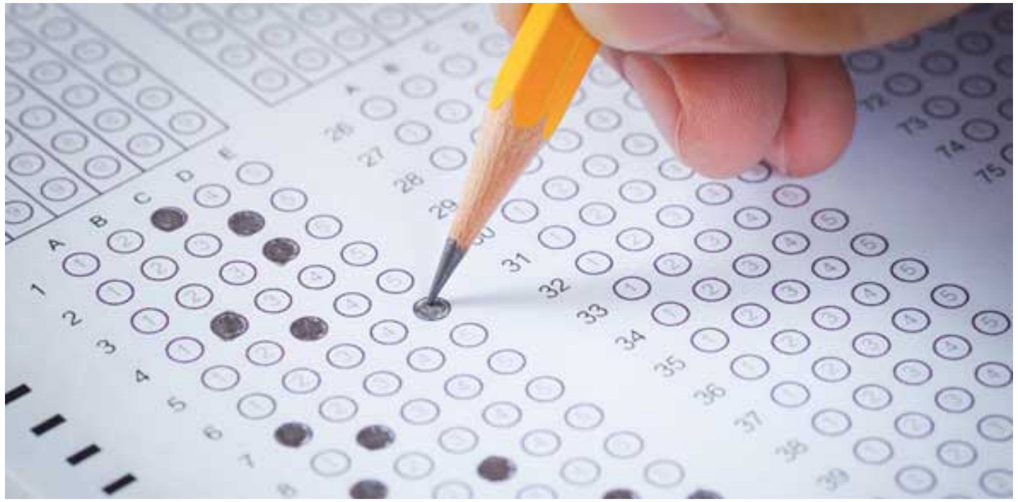
(Model evaluation & benchmarking)

- Task data are mostly collected via crowdsource
- Scores are used for model ranking

Motivation

UNC NLP

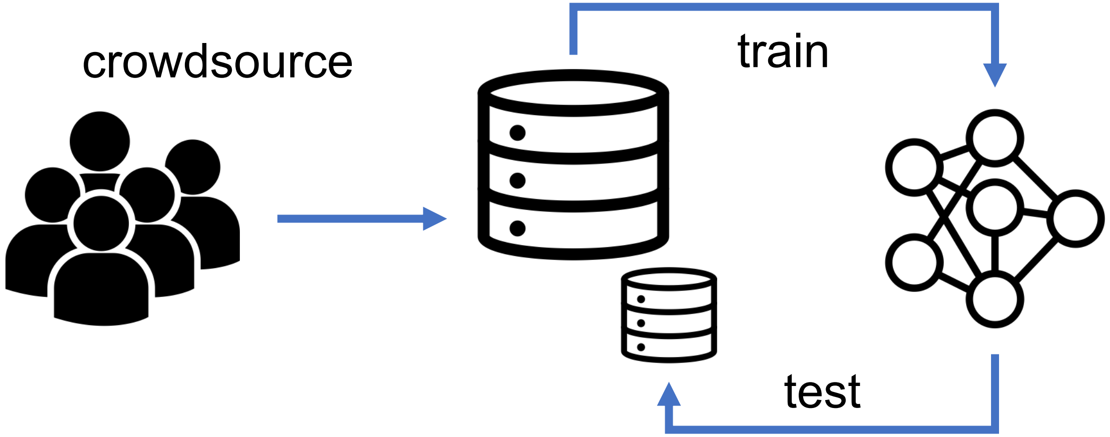
Human Education



(Human education testing: SAT, GRE, etc.)

- Questions & answers are designed by educators
- Scores are used as certifications or qualifications
- Most questions are objective

Natural Language Processing



(Model evaluation & benchmarking)

- Task data are mostly collected via crowdsource
- Scores are used for model ranking
- Many NLP tasks can be subjective

Motivation

UNC NLP

Human Education

Testing is mostly about understanding of a well-defined concept or knowledge.

Correct Labels are usually authoritative.

Natural Language Processing

Many NLP tasks depend on the unspecified pragmatic context, calculation of plausibility, etc.

Gold Label can often be debatable.

Motivation

UNC NLP

Human Education

Testing is mostly about understanding of a well-defined concept or knowledge.

Correct Labels are usually authoritative.

Natural Language Processing

Many NLP tasks depend on the unspecified pragmatic context, calculation of plausibility, etc.

Gold Label can often be debatable.

To understand collective human opinions on NLU data,
we did case studies on Natural Language Inference and Abductive Inference.

Natural Language Inference

Is the hypothesis entailed or contradicted by the premise?

Normal example in SNLI

Premise	A man inspects the uniform of a figure in some East Asian country.
Hypothesis	The man is sleeping.
Label	Entailment, Neutral, **Contradiction**

Natural Language Inference

UNC NLP

Is the hypothesis entailed or contradicted by the premise?

Normal example in SNLI

Premise	A man inspects the uniform of a figure in some East Asian country.
Hypothesis	The man is sleeping.
Label	Entailment, Neutral, **Contradiction**

Subtle example in MNLI

Premise	There are a number of expensive jewelry and other duty-free shops, all with goods priced in US dollars (duty-free goods must always be paid for in foreign currency).
Hypothesis	You can pay using the US dollar when buying goods from the duty-free shops.
Label	Entailment? Contradiction? Neutral?

Contradiction: A duty-free shop can only sell duty-free goods and you can only pay in foreign currency, assuming local is US.
Entailment: A duty-free shop can sell things other than duty-free goods for US dollar.

Abductive Commonsense Inference

UNC NLP

Which of the two hypotheses is more likely to cause Observation-Beginning to turn into Observation-Ending?

Normal example in Abductive NLI

Observation-B	It was a very hot summer day.
Hypothesis-1	He decided to run in the heat.
Hypothesis-2	He drank a glass of ice cold water.
Observation-E	He felt much better!
Label	Hypothesis-2

Abductive Commonsense Inference

Which of the two hypotheses is more likely to cause Observation-Beginning to turn into Observation-Ending?

Normal example in Abductive NLI

Observation-B	It was a very hot summer day.
Hypothesis-1	He decided to run in the heat.
Hypothesis-2	He drank a glass of ice cold water.
Observation-E	He felt much better!
Label	Hypothesis-2

Subtle example in Abductive NLI

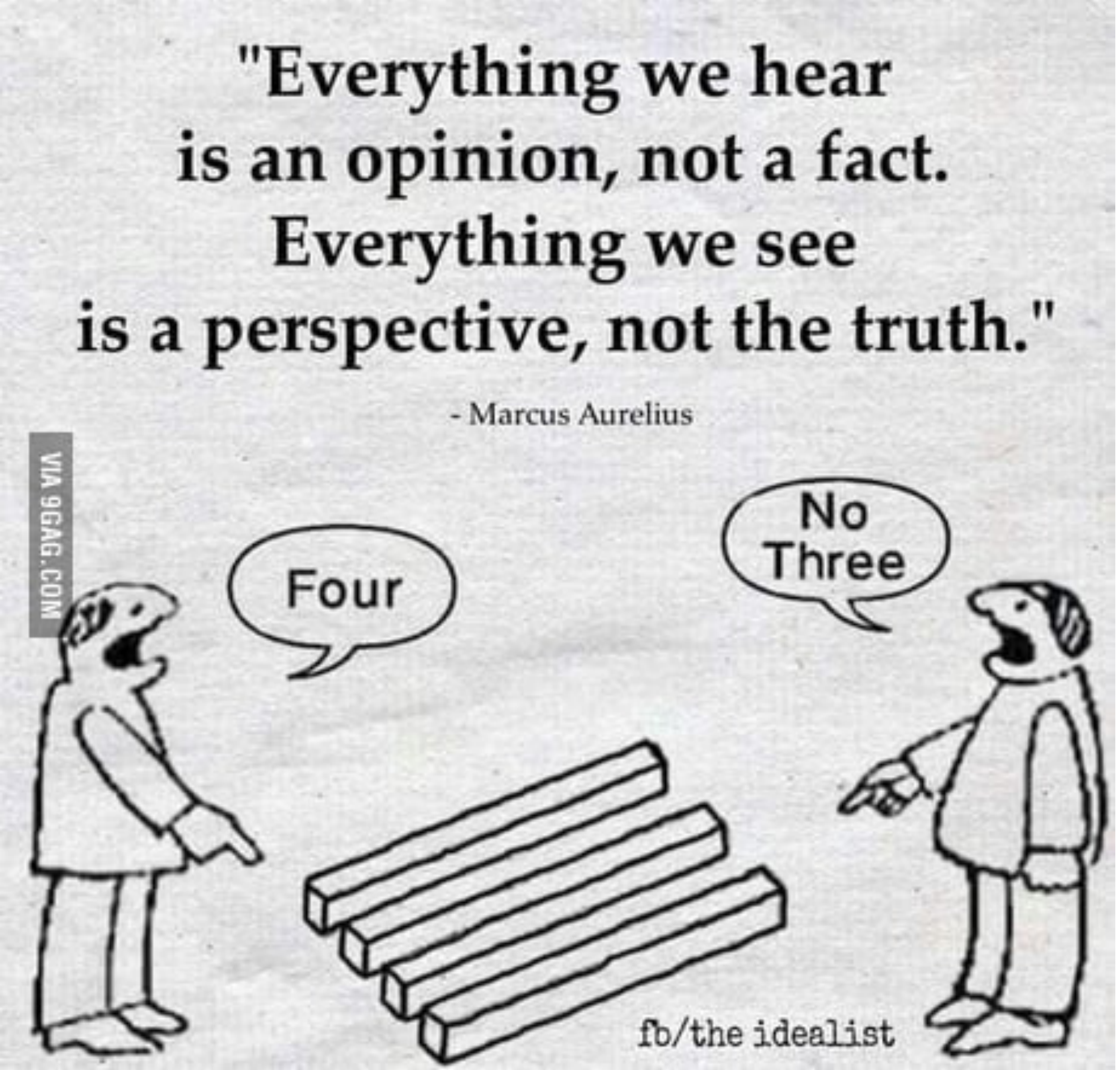
Observation-B	Amy and her friends were out at 3 AM.
Hypothesis-1	They started getting followed by a policeman, **ran**, and hid behind a building.
Hypothesis-2	The decided to break into the football field. When suddenly they **saw a flashlight** coming towards them. They all started running for the bleachers.
Observation-E	They stayed there **breathing hard**, and **praying they hadn't been seen**.
Label	Hypothesis-1 ? Hypothesis 2

Gold-label Practice is Worrisome

Gold-label Practice is Worrisome

Evaluating model ability to recover **human opinions distribution** is also (even more) important.

Collective HumAn OpinionS



Dataset

Collective **H**um**A**n **O**pinion**S**

ChaosNLI

100 Annotations per Example

	Abductive NLI (ChaosNLI-Alpha)	Stanford NLI (ChaosNLI-S)	Multi-NLI (ChaosNLI-M)
Count	1,532	1,514	1,599

A Total of 464,500 Annotations

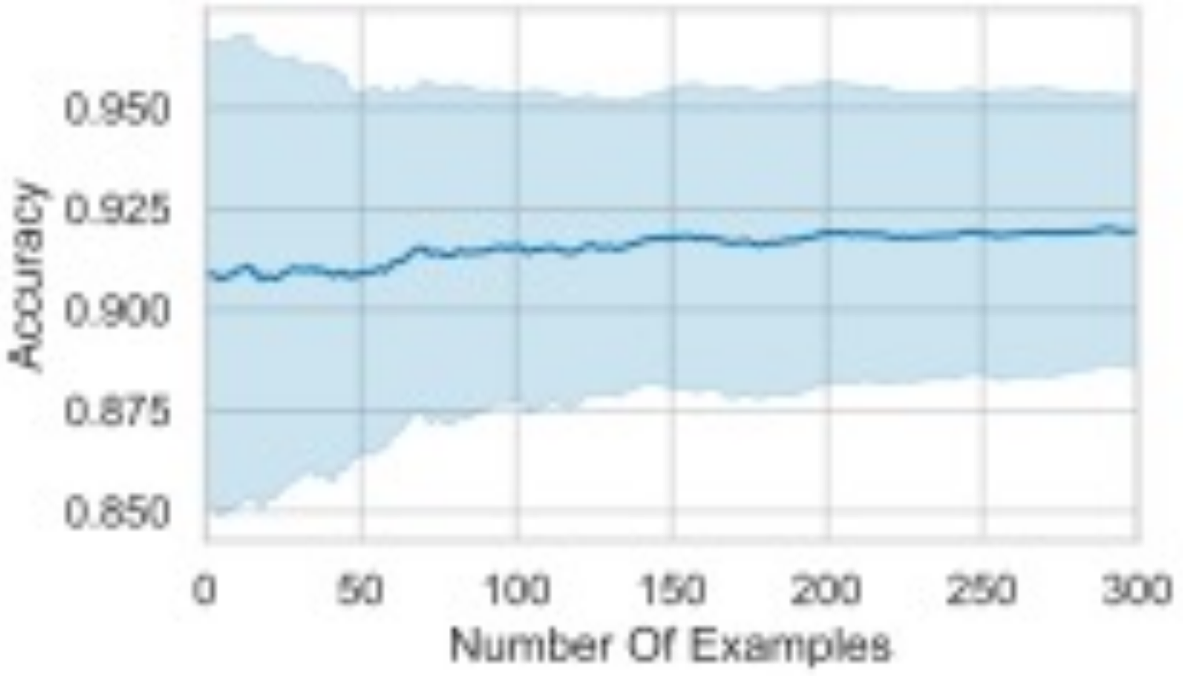
Collection

Challenge (Human Opinions)

Inter-Annotation-Agreement is not applicable

Quality Control

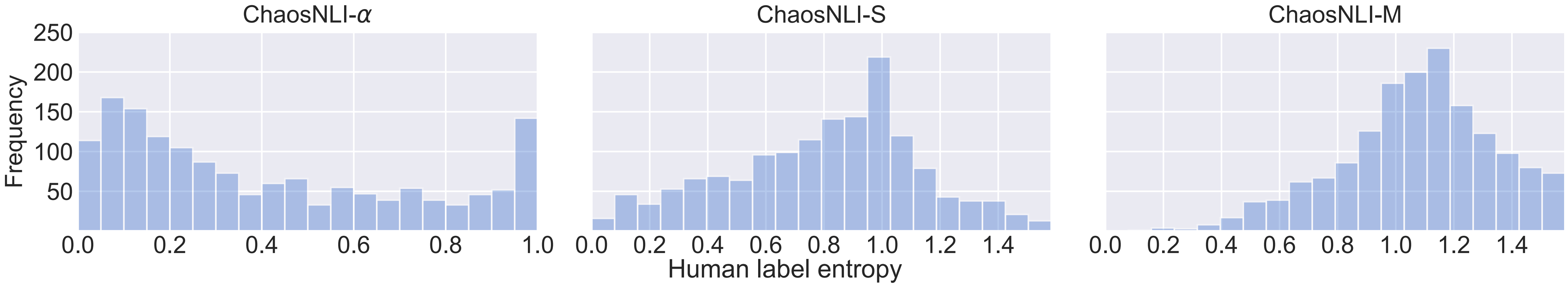
- Onboarding Test
- Training Phrase
- Performance Tracking



Human Agreement Distribution

100 annotations to calculate label distribution entropy.

$$\mathbf{H}(\mathbf{p}) = -\sum_{i \in \mathcal{C}} p_i \log(p_i) \qquad p_i = \frac{n_i}{\sum_{j \in \mathcal{C}} n_j}$$



Analysis of Model Predictions

Softmax Output vs. Human Label Distribution

$$\mathrm{KL}\left(\mathbf{p} \| \mathbf{q}\right) = \sum_{i \in \mathcal{C}} p_i \log\left(\frac{p_i}{q_i}\right)$$

$$\mathrm{JSD}\left(\mathbf{p} \| \mathbf{q}\right) = \sqrt{\frac{1}{2}\left(\mathrm{KL}\left(\mathbf{p} \| \mathbf{m}\right) + \mathrm{KL}\left(\mathbf{q} \| \mathbf{m}\right)\right)}$$

p is the estimated human distribution
q is model softmax outputs
m = (**p** + **q**) / 2

Analysis of Model Predictions

UNC NLP

Softmax Output vs. Human Label Distribution

Model	ChaosNLI-α			ChaosNLI-S			ChaosNLI-M		
	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)
Chance	0.3205	0.406	0.5098/0.5052	0.383	0.5457	0.4472/0.5370	0.3023	0.3559	0.4509/0.4634
BERT-b	0.3209	3.7981	0.6527/0.6534	0.2345	0.481	0.7008/0.7292	**0.3055**	0.7204	0.5991/0.5591
XLNet-b	0.2678	1.0209	0.6743/0.6867	0.2331	0.5121	0.7114/0.7365	0.3069	0.7927	0.6373/0.5891
RoBERTa-b	0.2394	**0.8272**	0.7154/0.7396	0.2294	0.5045	0.7272/0.7536	0.3073	0.7807	0.6391/0.5922
BERT-l	0.3055	3.7996	0.6802/0.6821	0.23	0.5017	0.7266/0.7384	0.3152	0.8449	0.6123/0.5691
XLNet-l	0.2282	1.8166	0.814/0.8133	0.2259	0.5054	0.7431/0.7807	0.3116	0.8818	**0.6742**/0.6185
RoBERTa-l	**0.2128**	1.3898	**0.8531**/0.8368	0.221	0.4937	**0.749/0.7867**	0.3112	0.8701	**0.6742/0.6354**
BART	0.2215	1.5794	0.8185/0.814	**0.2203**	0.4714	0.7424/0.7827	0.3165	0.8845	0.6635/0.5922
ALBERT	0.2208	2.9598	0.8440/**0.8473**	0.235	0.5342	0.7153/0.7814	0.3159	0.862	0.6485/0.5897
DistilBert	0.3101	1.0345	0.592/0.607	0.2439	**0.4682**	0.6711/0.7021	0.3133	**0.6652**	0.5472/0.5103
Est. Human	0.0421	0.0373	0.885/0.97	0.0614	0.0411	0.775/0.94	0.0695	0.0381	0.66/0.86

Human performance is estimated by comparing 100 human labels and another 100 human labels.

Significant difference exists between model outputs and human opinions.

Analysis of Model Predictions

Softmax Output vs. Human Label Distribution

Model	ChaosNLI-α			ChaosNLI-S			ChaosNLI-M		
	JSD$\downarrow$	KL$\downarrow$	Acc.$\uparrow$ (old/new)	JSD$\downarrow$	KL$\downarrow$	Acc.$\uparrow$ (old/new)	JSD$\downarrow$	KL$\downarrow$	Acc.$\uparrow$ (old/new)
Chance	0.3205	0.406	0.5098/0.5052	0.383	0.5457	0.4472/0.5370	0.3023	0.3559	0.4509/0.4634
BERT-b	0.3209	3.7981	0.6527/0.6534	0.2345	0.481	0.7008/0.7292	**0.3055**	0.7204	0.5991/0.5591
XLNet-b	0.2678	1.0209	0.6743/0.6867	0.2331	0.5121	0.7114/0.7365	0.3069	0.7927	0.6373/0.5891
RoBERTa-b	0.2394	**0.8272**	0.7154/0.7396	0.2294	0.5045	0.7272/0.7536	0.3073	0.7807	0.6391/0.5922
BERT-l	0.3055	3.7996	0.6802/0.6821	0.23	0.5017	0.7266/0.7384	0.3152	0.8449	0.6123/0.5691
XLNet-l	0.2282	1.8166	0.814/0.8133	0.2259	0.5054	0.7431/0.7807	0.3116	0.8818	**0.6742**/0.6185
RoBERTa-l	**0.2128**	1.3898	**0.8531**/0.8368	0.221	0.4937	**0.749/0.7867**	0.3112	0.8701	**0.6742/0.6354**
BART	0.2215	1.5794	0.8185/0.814	**0.2203**	0.4714	0.7424/0.7827	0.3165	0.8845	0.6635/0.5922
ALBERT	0.2208	2.9598	0.8440/**0.8473**	0.235	0.5342	0.7153/0.7814	0.3159	0.862	0.6485/0.5897
DistilBert	0.3101	1.0345	0.592/0.607	0.2439	**0.4682**	0.6711/0.7021	0.3133	**0.6652**	0.5472/0.5103
Est. Human	0.0421	0.0373	0.885/0.97	0.0614	0.0411	0.775/0.94	0.0695	0.0381	0.66/0.86

Chance baseline is using uniform distribution on the labels.

Even chance baseline is hard to beat.

Analysis of Model Predictions

Softmax Output vs. Human Label Distribution

Model	ChaosNLI-α			ChaosNLI-S			ChaosNLI-M		
	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)	JSD↓	KL↓	Acc.↑ (old/new)
Chance	0.3205	0.406	0.5098/0.5052	0.383	0.5457	0.4472/0.5370	0.3023	0.3559	0.4509/0.4634
BERT-b	0.3209	3.7981	0.6527/0.6534	0.2345	0.481	0.7008/0.7292	**0.3055**	0.7204	0.5991/0.5591
XLNet-b	0.2678	1.0209	0.6743/0.6867	0.2331	0.5121	0.7114/0.7365	0.3069	0.7927	0.6373/0.5891
RoBERTa-b	0.2394	**0.8272**	0.7154/0.7396	0.2294	0.5045	0.7272/0.7536	0.3073	0.7807	0.6391/0.5922
BERT-l	0.3055	3.7996	0.6802/0.6821	0.23	0.5017	0.7266/0.7384	0.3152	0.8449	0.6123/0.5691
XLNet-l	0.2282	1.8166	0.814/0.8133	0.2259	0.5054	0.7431/0.7807	0.3116	0.8818	**0.6742**/0.6185
RoBERTa-l	**0.2128**	1.3898	**0.8531**/0.8368	0.221	0.4937	**0.749/0.7867**	0.3112	0.8701	**0.6742/0.6354**
BART	0.2215	1.5794	0.8185/0.814	**0.2203**	0.4714	0.7424/0.7827	0.3165	0.8845	0.6635/0.5922
ALBERT	0.2208	2.9598	0.8440/**0.8473**	0.235	0.5342	0.7153/0.7814	0.3159	0.862	0.6485/0.5897
DistilBert	0.3101	1.0345	0.592/0.607	0.2439	**0.4682**	0.6711/0.7021	0.3133	**0.6652**	0.5472/0.5103
Est. Human	0.0421	0.0373	0.885/0.97	0.0614	0.0411	0.775/0.94	0.0695	0.0381	0.66/0.86

Large models are not always better.

Examples (NLI)

Premise	There are a number of expensive jewelry and other duty-free shops, all with goods priced in US dollars (duty-free goods must always be paid for in foreign currency).
Hypothesis	You can pay using the US dollar when buying goods from the duty-free shops.
Old Labels	C, C, E, N, C
New Labels	E(51), N(3), C(46)

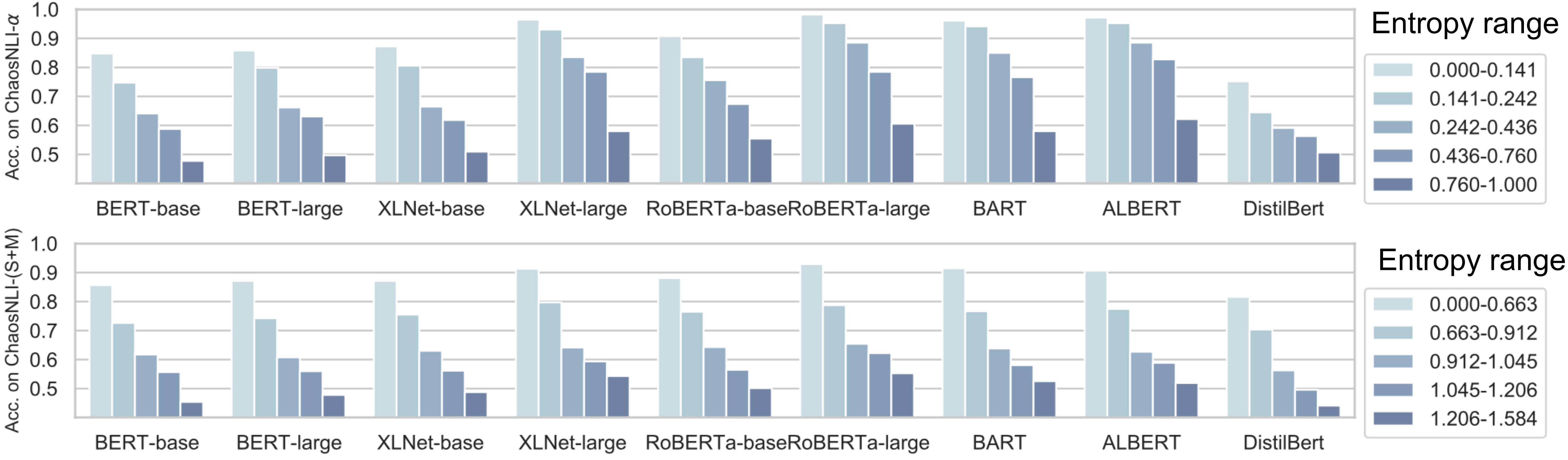
	BERT-large	RoBERTa-large	XLNet-large	BART	ALBERT	DistilBERT
Entailment	50.03%	95.04%	91.80%	95.16%	38.16%	46.33%
Neutral	5.33%	3.63%	1.59%	3.97%	6.33%	32.69%
Contradiction	44.63%	1.33%	6.61%	0.87%	55.50%	20.98%

Examples (Abductive NLI)

Observation-B	A scientist discovers that there is a disease beginning to spread.
Hypothesis-1	The scientist warns everyone then realizes he was wrong.
Hypothesis-2	They accidentally contaminated themselves with the spread.
Observation-E	They feel foolish for having done so.
Old Label	Hypothesis-1
New Label	Hypothesis-1(41), Hypothesis-2(59)

	BERT-large	RoBERTa-large	XLNet-large	BART	ALBERT	DistilBERT
Hypothesis-1	0.01%	4.50%	12.25%	0.67%	97.67%	4.92%
Hypothesis-2	99.9%	95.50%	87.75%	99.33%	2.33%	95.08%

The effect of human agreement



Models achieve near perfect accuracy on data with high agreement while cannot beat random guess on data with low agreement.

Related Work

UNC NLP

This work is inspired by previous work on "Inherent Disagreements in Human Textual Inferences". (Pavlick and Kwiatkowski, 2019)

(We stick to the 3-way NLI labeling schema while Pavlick&Kwiatkowski2019 use a continuous labeling schema)

Human annotation disagreements are also studied on other tasks including:

- word sense disambiguation (Erk and McCarthy, 2009; Jurgens, 2013), coreference (Versley, 2008),
- frame corpus collection (Dumitrache et al., 2019),
- anaphora resolution (Poesio and Artstein, 2005; Poesio et al., 2019), entity linking (Reidsma and op den Akker, 2008),
- tagging and parsing (Plank et al., 2014; Alonso et al., 2015),
- veridicality (De Marneffe et al., 2012; Karttunen et al., 2014).

Take away

UNC NLP

- NLU evaluation should consider evaluating collective human opinions;
- We present **ChaosNLI**; (100 annotations per example for examples in SNLI, MNLI and AbductiveNLI)
- **High human disagreement** exists in a **noticeable** amount of examples;
- The models **lack the ability** to recover the distribution over human labels;
- The models achieve **near-perfect accuracy** on the data with **high agreement**, whereas they **can barely beat a random guess** on the data with **low agreement**.

Future work

Explicit design on both evaluating and training models for human opinions distribution, especially on NLP tasks with a descriptive nature.

Thanks

Contact: yixin1@cs.unc.edu
GitHub: https://github.com/easonnie/ChaosNLI
Website: https://cs.unc.edu/~yixin1

Acknowledgment:
Thanks to the authors of SNLI, MNLI, and αNLI. This work was supported by NSF, DARPA, ONR.

UNC NLP