11830 Report: Biases in Crowdsourced Annotations

Jiyang Tang
jiyangta@andrew.cmu.edu

1 Introduction

The Stanford Natural Language Inference (SNLI) corpus[1] is a large crowdsourced natural language inference dataset (Bowman et al., 2015). For each premise sentence, annotators were asked to write a hypothesis that is either a contradiction, a neutral statement, or entailment of the premise.

The data was collected from Amazon Mechanical Turk (MTurk) crowdsourcing service. Previous studies have shown that MTurk crowd-workers tend to have lower income, higher education levels, and lower average ages (Levay et al., 2016). We believe that biases are propagated from the dataset authors to the premise text, and from the crowd-workers to the hypothesis text. Therefore, we will analyze the biases in the data by performing word association tests in this report.

2 Method

2.1 Pointwise Mutual Information

Pointwise Mutual Information (PMI) is used to measure how much word w_i is associated with the word w_j (Church and Hanks, 1990; Jurafsky and Martin, 2009). PMI is calculated as follows:

$$\text{PMI}(w_i, w_j) = log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i)c(w_j)}$$

where N is the total number of sentences in the corpus, $c(w_i, w_j)$ is number of times w_i and w_j co-occur in a sentence, $c(w_i)$ is the number of times w_i occurs in the corpus, and $c(w_j)$ is the number of times w_j occurs in the corpus.

Note that if a pair of words w_i and w_j occurs multiple times in the same sentence, $c(w_i, w_j)$ is counted as 1.

PMI ranges from negative infinity to positive infinity. Large PMI suggests high word association. Negative PMI implies two words co-occur less often than by chance and is unreliable in practice (Jurafsky and Martin, 2009).

3 Experiments

3.1 Data Preprocessing

As mentioned before, each data sample contains a premise and a hypothesis. Note that multiple hypotheses might be generated from the same premise, therefore duplicated sentences are removed. All words are converted into their lowercase form and stop words are removed. Then we use `spaCy` (Honnibal et al., 2020) `en_core_web_sm` model to tokenize raw strings into lists of words and remove all punctuations.

Note that words that occur less than 10 times in the corpus are removed.

3.2 Unigram PMI

We individually perform unigram PMI analysis on the premise text and the hypothesis text. For each analysis experiment, we focus on the most associated words with a set of identity labels (Rudinger et al., 2017).

3.3 Bigram PMI

We extend the list of identity labels by combining existing ones into bigrams. We also extract all bigrams from the text and add them to the vocabulary for PMI calculation. Note that we skip the calculation if two bigrams share a unigram since we are not interested in the association between "he" and "he doesn't".

4 Results and Discussion

4.1 Unigram PMI

Table 1 lists the top associated unigrams with some of the identity labels in the premises and in the hypotheses. We can easily spot some alarmingly biased word associations. For example, `muslims` is highly associated with

[1] https://nlp.stanford.edu/projects/snli/

Identity	Premise	Hypothesis
women	saris, headscarves, bikinis, headdresses, coverings	burkas, husbands, saris, kimonos, bikinis
men	turbans, tuxedos, ladders, jumpsuits, wetsuits,	turbans, rickshaws, wives, cigars, tuxedos,
africans	tribe, hearts, tap, huts	armed, source, die, cloths, tribal
caucasian	lockers, handsome, explains, contemplates, straddling	slender, fleece, non, zip, festive
muslims[1]	channel, news, sponsored, celebrate, speech	christians, terrorists, celebrate, opening, phones
christians[1]	praising, lord, crazy, fun, woods,	muslims, gospel, impressed, pork, villagers
gay	pride, marriage, attendees, protester, participants,	pride, rights, marriage, experimenting, abraham
straight	razor, ahead, stony, sketch, crack	razor, tambourine, ahead, stared, lanes
israeli	desolate, nuts, pirates, cigarettes, u.s.	problems, eastern, cashier, cigarettes, counter
american	footballer, african, native, patriotic, south	idol, native, african, latin, drapes
indian	sari, headdresses, saris, ritual, chief	southeast, style, boot, muffins, descent

[1] Words that don't have associated words that occur more than 10 times in premise text. The threshold is ignored for these words.

Table 1: Top associated unigrams with identity labels

terrorists, africans co-occur with armed and die frequently, while caucasian is often associated with handsome. There are also many implicit biases or stereotypes in the data. women is most associated with words related to fashion, while men with words about tools and work clothes.

Meanwhile, we can spot more heavily biased word associations in the hypothesis data compared to the premises. For example, associations between muslims and terrorists and between women and gossip are only present in the hypotheses.

However, there are yet some cases where the bias is more prevalent on the premises. For example, israeli is most associated with desolate and pirates in premises, compared to problems, eastern, cashier and so on in hypotheses.

We also find it interesting that south is among the top 5 most associated words with american while north is not found even in the top 20s, as if people think american's are from North America by default.

4.2 Bigram PMI

Table 2 lists the top associated unigrams or bigrams with bigram identity labels. This result corresponds to our previous finding that hypothesis text contains more biases and stereotypes. For example, old men and grumpy, black people and rioting, and asian teenagers and electronic devices.

Table 3 shows the top associated unigrams or bigrams with unigram identity labels. We can spot new stereotypical phrases not seen in previous experiments. For example, indian and works hard, and africans and shooting guns.

4.3 Qualitative Analysis

In this section, we will present some examples that contain biases or stereotypes.

In the example below, the annotator somehow chooses Africans shooting guns as a contradictory event of the premise. In addition, they use "shooting guns" instead of "hunting".

Premise: Africans in tribe clothes, walking pass a green.
Hypothesis: Africans are shooting guns at a bear.
Label: contradiction

In this example, the hypothesis also contains stereotypes against African people.

Premise: Africans working in a mine digging.
Hypothesis: Africans are being forced to mine for diamonds.
Label: neutral

In the next example, we believe that the annotator is misled by the premise. Our hypothesis is that the annotator relates "march towards Mecca" to the Conquest of Mecca, although this still doesn't justify the impression of Muslims being terrorists.

Premise: Several Muslim worshipers march towards Mecca.
Hypothesis: The Muslims are terrorists.
Label: neutral

This example is a more appropriate neutral hypothesis.

Premise: Muslim women talking in a marketplace.
Hypothesis: The Muslims are talking on their phones.
Label: neutral

Identity	Premise	Hypothesis
old men	park benches, stock, rock concert, swim caps, concrete bench	reminisce, straw hats, inappropriate, raincoats, grumpy
old women	selling vegetables, wearing hat, cases, walker, fresh produce	intense game, vases, earring, carring, doorstep
black people	scarf walks, couches, looking forward, boston, dilapidated	corvette, wedding ceremony, rioting, felt, robbed
black male	puma, free throw, grinder, shirt hanging, waffle	green t-shirt, notebooks, backyard pool, building looking, unfinished building
asian teenagers	NA	red chairs, electronic devices, carrying flags, atm, piercings

Table 2: Top associated unigrams or bigrams with bigram identity labels

Identity	Premise	Hypothesis
indian	sari, headdresses, saris, ritual, chief	southeast asian, southeast, wearing traditional, works hard, asian descent
africans	green vegetables, tribe, hearts, tap, huts	shooting guns, taking shelter, armed, source, getting water

Table 3: Top associated unigrams or bigrams with unigram identity labels

Below is an example of stereotypes against Israeli people.

Premise: Five women wearing uniforms and carrying guns are spending time together while a man in a black robe talks on a cellphone in the background.
Hypothesis: The women are Israeli soldiers.
Label: neutral

And in other cases the hypothesis is fine.

Premise: Protesters marching against Israeli actions.
Hypothesis: People are have a protesting problems in Israeli.
Label: entailment

And there seems to be a common association between female identities and gossiping.

Premise: Two women sit outside on a bench along a street, looking across the street.
Hypothesis: Two women sit outside on a bench along a street, looking across the street as they gossip with each other.
Label: neutral

Premise: Two men in orange vests work in road construction.
Hypothesis: Three women gossip in the break room.
Label: contradiction

Premise: Three girls wear straw hats and dance on a stage.
Hypothesis: 3 girls are painting their nails and gossiping in the locker room.
Label: contradiction

Premise: Four girls are sitting in a hot tub.
Hypothesis: The girls are gossiping.
Label: neutral

4.4 Crowdsourcing Setup

As we have seen in previous sections, the premises sometimes contain biases. Meanwhile, they can also indirectly mislead crowd-workers into adding their own biases to the hypotheses by containing concepts, which are harm-free by themselves, that resonate with crowd-workers biases.

A common characteristic of such cases is that they are mostly unrelated to the premises except for the main subject. This phenomenon is particularly obvious in `neutral` data samples, in which the crowd-workers have the most freedom for creative imagination.

We believe with adequate instructions, a free-form generated data collection approach can be mostly free from biases. For example, specific warnings should be given to crowd-workers so that they do not imagine completely unrelated scenarios when writing hypotheses. In addition, they should be careful about introducing gender, race, nationality, and other identities that are unseen in the premises. Perhaps asking them for a short paragraph to justify their reasoning in such cases can discourage them from doing that. Meanwhile, this extra data can be used for building reasoning models.

5 Conclusion

In this report, we have analyzed the biases in SNLI dataset both quantitatively and qualitatively. In addition to biases in premises, we found much more in the hypotheses. Sometimes such biases are indirectly generated based on concepts in the premises, which encourages us to develop a good crowdsourcing paradigm that discourages crowd-workers from generating biases.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed edition. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J. OCLC: 213375806.

Kevin E. Levay, Jeremy Freese, and James N. Druckman. 2016. The demographic and political composition of mechanical turk samples. *SAGE Open*, 6(1):2158244016636433.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.