11830 Report: Biases in Crowdsourced Annotations

Jiyang Tang
jiyangta@andrew.cmu.edu

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

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, 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 inevitably 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 word w_j (Church and Hanks, 1990; Jurafsky and Martin, 2009). PMI is calculated as follows:

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

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

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 premise text is removed. All words are converted into its lower case 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 occurs less than 10 times in the corpus are removed.

3.2 Unigram PMI

We first perform unigram PMI analysis on the entire corpus, and then on the premise text and the hypothesis text individually.

For each analysis experiment, we focus on the most associated words with a set of identity labels (Rudinger et al., 2017).

4 Results and Discussion

4.1 Unigram PMI

Table 1 lists top associated unigrams with some of the identity labels in the entire corpus, in

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

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

Table 1: Top associated unigrams with identity labels

premise text, and in hypothesis text. We can easily spot some alarmingly biased word association. For example, `muslims` is highly associated with `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. This implies that women spend more time on fancy clothes while men work on labor jobs and don't care about their looks, which is not true.

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

However, there are yet some cases where the bias is more prevalent in 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` are from North America by default.

4.2 Qualitative Analysis

In this section, we will present some examples that contains 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 implies that Africans are slaves. Moreover, the premise itself also contains stereotypes of Africans working labor jobs in harsh conditions.

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 annotator is misled by the premise although the label is `neutral`. 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

4.3 Crowdsourcing Setup

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.