- NLI models and datasets have been tested against their syntactic (McCoy, 2019, Right for the Wrong Reasons) and semantic (Glockner, 2018, Breaking NLI Systems) biases and brittleness. It has also been probed by logic fragments Probing Natural Language Inference Models through Semantic Fragments Kyle Richardson
-
- https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf refer Table 3 to make Dataset comparison table
    - Add FTFY: joke, correction, other
    - esnli
    - snli
    - fever
    - mnli
    - multinli
    - vitc
    - and more
- NLI majorly solves a 3 way classification with a variety of approaches, datasets, and domains etc. By other works in the domain, the 3 way classification has to be changed to adapt to the data or the way to generate explanations etc. where they expand to more classes than 3. Thus these are not very generalizable for all kinds of NLI tasks, use Knowledge Graphs or WordNet etc. as resources, or do not detail out on where and what the inconsistency in the text pair was etc.
    - Explaining Text Matching on Neural Natural Language Inference YOUNGWOO KIM slightly departs from this and makes two parallel tagset in a way, including mismatch in one of them
    - TaxiNLI: Taking a Ride up the NLU Hill Pratik Joshi and Alok Sathe takes an entirely different direction by proposing a new and fine-gained taxonomy for NLI tasks as a whole
    - We focus specifically on inconsistencies, which could include bits of entailment and contradiction from the classical NLI divisions but finding where inconsistencies could lie and what kind of inconsistency is useful as fact checking etc. etc. require it
- It is clear that we need more information. It could be external like Chen? CHECK, or from LMs themselves like Thorne (Evidence-based Factual Error Correction), or by crowdsourced annotations like eSNLI etc. The kind of information also differs from WK like Wiki, Politifact etc. to linguistic like Manning etc. but it needs to do 2 things, help the model learn + help the readers understand what is wrong. => Motivation for our system which benefits both together, it is light enough for a Jointly learning model to digest and make use of, while it is detailed enough to show why and what is wrong to the user.
    - We do both
- Our mismatch span is what some other papers consider as explanations like Thorne (Generating Token-Level Explanations for Natural Language Inference) model must generate an explanation e defined as a subset of zero or more tokens from both the premise and hypothesis sentences. But our explanation gives the span with more information.
    - Here also, the annotators were asked to use tokens from mismatching spans to generate an explanation: eSNLI is also a 3 way classification, but it brings in a crowdsourced and organic explanation dataset. That is difficult to

generate because of NLG progress and difficult to read in a Grammarly like manner, we need something lighter and more targeted in terms of only a handful of labels

- We need to say two things for acl refer (https://dl.acm.org/doi/pdf/10.1145/3485127 pg. 4) if needed:
    - Usage CC copyright stuff for FEVER & FICE
    - What FEVER was meant for
    - What FICE is meant for
    - Has FICE done anything outside of FEVER's scope?
-
- Future work
    - Automating Annotations
        - The SRT, Head-Modifier tags anyway bring in structural constraints which were shown as an important robustness check by UnNatural Language Inference Koustuv Sinha since this still requires expert annotators, we could explore adding dependency level information to the data to help identify such triples instead, as shown by Evaluating Factuality in Generation with Dependency-level Entailment Tanya Goyal these relation arcs are definitely helpful to nli modelling anyway.
            - Correcting Contradictions Aikaterini-Lida Kalouli
            - Shows that even human Turkers cannot deal with improving the SICK corpus, because of linguistic complexity in predicate-argument structure. Thus expert annotation or dep parser is necessary
        - Secondly we could use commonly occurring logical frames like negation etc to rely lesser on annotators marking mismatch location.
            - Like Towards Semantic Modeling of Contradictions and Disagreements: A Case Study of Medical Guidelines Wlodek Zadrozny
            - However this is easier in med domain rn, not generalizable yet
        - Thirdly, given precise mismatch location identification, we could possibly use synsets to automatically mark mismatch type as well
            - Will have the issue of exhaustability and will need to be updated to current lexicon + NER issues
        - Finally, we could also use Wordnet in a slightly more involved manner to delexicalise the mismatching entity types to taxonomic heads.
            - Deciding how far up to go and making an algorithm for that is problematic currently
    - is to make our output fluent and intelligible as a sentence or something, which can be done using Frames or some other NLI NLG method
    - Debais claim set against things like huge amount negation for making it harder for the model to ignore context in case it was already, as shown by Towards Debiasing Fact Verification Models Tal Schuster and Darsh J Shah, FICE inherits these from the FEVER dataset which can be carefully looked into. Also Annotation Artifacts in Natural Language Inference Data Suchin Gururangan says this too, both important citations, this is more important in fact.
    - Ext issues to work with multiple sentence or para level contexts

- Conclusion
    - We can use our model/scheme to boost:
        - NLI (or fake news) models which use meta data to compare context claim pairs as well like LIAR (Yang, “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection)
        - for nli models focused on natural or more common sense explanations like: Explain Yourself! Leveraging Language Models for Commonsense Reasoning Nazneen Fatema Rajani , which currently purely rely on & are restricted by Amazon Turkers to provide explanations.
        - Or even for sentiment class based opinion contradiction nli systems Towards a Framework for Detecting and Managing Opinion Contradictions Mikalai Tsytsarau, as FICE explanations annotation scheme is **very generalized**
    - Fact checking domains can be greatly benefited by an approach like this as well. As A Survey on Automated Fact-Checking Zhijiang Guo notes, fine-grained veracity ratings (like “half-truth”) can be noisy and unreliable. Thus an approach like FICE can help provide resources for explaining the “why?” or a more useful explanation, as the survey mentions to be the way ahead for the fact checking domain.
    - Like Grammarly Use it for correction like Thorne (Evidence-based Factual Error Correction)

---

Emotion, Motivation, and Text Comprehension:The Detection of Contradictions in Passages:

- Many of the participants failed to detect the contradictions
- in the absence of instructions about potential contradictions, the depressed-mood group identified only one third as many contradictions as the neutral-mood group, and both groups identified relatively few of the six possible contra-dictions. In contrast, when instructions about the possible presence of a contradiction were administered, both neutral-and depressed-mood groups accurately identified more contradictions than those who received no instructions. Nevertheless, the depressed-mood participants who were motivated, by way of instructions, still identified fewer correct contradictions than neutral-mood participants.
- depressed participants again reliably identified fewer correct contra-dictions than the neutral-mood participants, replicating the findings of Experiment 1. Even though participants no longer had to maintain the location of the contradictions in working memory and were allowed more time to search for contradictions, depressed individuals continued to be less effective at identifying contradictions.
- only neutral-mood participants' judgements of task difficulty reliably predicted their contradiction-identification performance, indicating that depressed individuals were less accurate in their metacomprehension judgements.

Papers left:

Left things:

- Summarisation and Dialog papers

- Model papers:

-- https://aclanthology.org/2020.acl-main.656.pdf

-- https://aclanthology.org/2020.acl-main.771.pdf

-- https://aclanthology.org/P19-1085.pdf

-- https://arxiv.org/pdf/2104.08142.pdf

Generating Fact Checking Explanations:

- Built on the LIAR-PLUS Dataset
- For every corresponding claim in the dataset, it had comments for that particular case(in our case context), veracity ratings(pants on fire, false, mostly false, half true, mostly true, true) and corresponding explanation in paragraph form.
- The model contributions show that jointly predicting the rating and explanation give better performance

NILE : Natural Language Inference with Faithful Natural Language Explanations(slight dangerous)

- Built on e-snli
- They have built a two stage model:
    - Candidate Explanation Generators: Given a claim and hypothesis, it generates label-specific explanations if the label was entail, contradict, neutral(for all three cases)
    - Explanation Processor: GIven the claim, hypothesis and the three generated explanations, it produces a label

Evaluating Factuality in Generation with Dependency-level Entailment(2010.05478.pdf (arxiv.org))

LoNLI: An Extensible Framework for Testing Diverse Logical Reasoning Capabilities for NLI(2112.02333.pdf (arxiv.org))

NLI datasets:

- MultiNLI

- SNLI
- MNLI
- ESNLI

**Formats**

| **Model** | **Input** | **Output** |
|---|---|---|
| **t5/bart** | <claim> Russia's capital is Berlin.<br><context> Russia 's capital Moscow is one of the largest cities in the world ; other major urban centers include Saint Petersburg , Novosibirsk , Yekaterinburg , Nizhny Novgorod and Kazan .<br><source> Russia's capital<br><relation> is<br><target> Berlin<br><mismatch> Moscow | <mislocation> Target-Head<br><mistype> T Sisters |
| **BERT/** | <SRC> Russia's capital </SRC> <REL> is </REL> <TAR> Berlin </TAR>. [SEP] Russia 's capital <MIS> Moscow </MIS> is one of the largest cities in the world ; other major urban centers include Saint Petersburg , Novosibirsk , Yekaterinburg , Nizhny Novgorod and Kazan . | Token classification in case of span annotation.<br><br>Sequence classification in case of mistype, mislocation and mismatching entity classification. |
| | | |