EXPLICITLY STATING ASSUMPTIONS REDUCES HALLUCINATIONS IN NATURAL LANGUAGE INFERENCE

Wenchuan Mu, Kwan Hui Lim
Singapore University of Technology and Design
{wenchuan_mu,kwanhui_lim@sutd.edu.sg}

ABSTRACT

A natural language inference (NLI) model might hold a hallucination that a 'premise' infers a 'hypothesis'. This hallucination is possibly not due to insufficient training on the data but rather is inherent in the data themselves. A training label might suggest that 'a premise infers a hypothesis', but this inferential link could be overstated. This overstating might arise from the mismatch in intensity or context. We propose that reinforcing the premise could mitigate this problem. To address this, we introduce a method that employs predicate logic and natural deduction to explicitly articulate assumptions that validate the inferential link. This method allows for a transparent inference process, drawing logically plausible conclusions based on explicit premises. The NLI model may thus achieve higher reliability in understanding and processing natural language.

1 INTRODUCTION: NATURAL LANGUAGE INFERENCE AND HALLUCINATION

NLI focuses on discerning inferential links between two text statements, but hallucination arises when a model relies on false perceptions to discern these links. An overstated inferential link may not stem from inadequate training but rather from the way training samples are presented. For example, a premise is "About two weeks before the trial started, I was in Shapiro's office in Century City" and a hypothesis is "Shapiro works in Century City" (Haim et al., 2006). While a well-trained NLI model may be confident to give an affirmative answer, this premise may fail to infer this hypothesis in real applications. Suppose Shapiro gets a job offer at the end of 2020 when people can work from home, that "Shapiro works in Shapiro's office" cannot be guaranteed. This kind of assumption, when not explicitly stated, reflects the hallucination tendency of an NLI model. Even though experimental accuracy might suggest the efficacy of such inference without pinging its underlying assumption, it can be unstable or become quickly outdated when applied in the real world.

To equip NLI models for the dynamic nature of real-world scenarios, the training dataset needs to capture conditions under which inferences are valid. We propose that adding unstated assumptions to the premises of training samples can be beneficial, as it clarifies the circumstances in which a premise logically leads to a hypothesis. First, we introduce a method using predicate logic and natural deduction to extract sufficient assumptions from existing premise/hypothesis pairs. Next, we fine-tune NLI models with these enhanced pairs. We then assess the performance of these models in real-world scenarios (GRE Argument). We find that this method enhances the model's capability to make precise inferences, effectively reducing the influence of incorrect assumptions in its outputs.

2 METHOD: INFERENCE WITH SUFFICIENT ASSUMPTION(S)

Here, we outline our method for adding assumptions to the premises of training samples. Typically, a standard NLI sample looks like this: **Premise** A turtle danced. **Hypothesis** A turtle moved. **label** Inferred. Our method involves explicitly connecting actions, such as 'dance' to 'move', by adding an assumption that 'dancing' involves 'moving'. This results in a sample like **Premise** A turtle danced. Dancing is a kind of movement. **Hypothesis** A turtle moved. **label** Inferred. The question is how to systematically derive such assumptions.

First, we parse the existing premise P and hypothesis H into predicate logic, defining the undetermined conclusion X as $(P \rightarrow H)$. In the turtle example, **P** $\exists\, e\, x.\ \mathrm{Act}(e, x) \wedge v_{\text{dance}}(e) \wedge n_{\text{turtle}}(x)$;

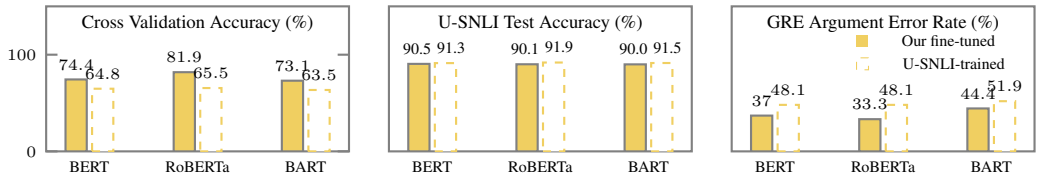


Figure 1: Result of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2020) when fine-tuned on enhanced samples. U-SNLI-trained models are only tested for reference.

H $\exists\, e\, x.\ \mathrm{Act}(e, x) \wedge v_{\text{move}}(e) \wedge n_{\text{turtle}}(x)$. Next, we find an assumption S that, if plausible ($\top(S)$), makes X also plausible. When P is known, S is a sufficient assumption. Formally, $P \wedge S \vdash H$; $P \wedge S \nvdash \bot$; $S \nvdash H$. Then, we apply the following natural deduction to obtain S.

Theorem. *Any S is a sufficient assumption if $P \wedge \neg H \vdash \neg S$. Any S is a sufficient assumption, if S' is a sufficient assumption and $\neg S' \vdash \neg S$.*

Proof. The natural deduction goes as `1` $P \wedge S \vdash H$, `2` $P, S \vdash H$ (Simplification), `3` $P \vdash S \rightarrow H$ (Modus Ponens), `4` $P, \neg H \vdash \neg S$ (Modus Tollens), `5` $P \wedge \neg H \vdash \neg S$ (Conjunction). □

An sufficient assumption of the turtle example can thus be $\neg\exists\, e\, x.\ \mathrm{Act}(e, x) \wedge v_{\text{dance}}(e) \wedge n_{\text{turtle}}(x) \vee \exists\, e\, x.\ \mathrm{Act}(e, x) \wedge v_{\text{move}}(e) \wedge n_{\text{turtle}}(x)$. After sorting and further simplification, we arrive at S : $\forall\, x. v_{\text{dance}}(x) \rightarrow v_{\text{move}}(x)$. Essentially, Dancing is (a kind of) movement.

After adding explicit assumptions to the premises, we revalue the labels of both enhanced and original premise-hypothesis pairs. In the turtle example, adding the assumption does not alter the sample label. However, other samples may need re-annotation, *e.g.*, that "there are two young ladies jogging, by the *ocean side*" is thought to infer that "two women are jogging by the *beach*" (Bowman et al., 2015). This inference could be faulty, even if merely a strong, not valid, argument is required. For instance, even in areas where beaches are common (Harvey & Caton, 2010), they make up less than 70% of coastal types (cliffs, wetlands, etc.) If 70% were a sufficient basis for inference, we could wrongly conclude "Peter cannot have three daughters" from "Peter has three children" (Simonoff, 2010), which is correct 85% of the time but dubious as an inference. Therefore, without assuming "ocean sides are (mostly) beaches", the sample should be labelled Not Inferred (`False`), but with the assumption, it could be Inferred (`True`). The revalued samples are then used to fine-tune NLI models.

3 Experiment: Does Our Method Reduce Model Hallucination?

We take 1,057 original premise/hypothesis pairs from U-SNLI (Chen et al., 2020) and use DRS Boxer (Bos, 2008) to parse texts. Note that NLI typically involves ternary classification with three classes: `True`, `False`, and `True negation`. The third class means contradiction, *i.e.*, $P \rightarrow \neg H$. Adding assumptions creates additional enhanced samples, leading to a balanced distribution of 500 samples per class, totalling 1,500 samples. We fine-tune three NLI models on these samples, using a five-fold cross-validation scheme, and also test them on the U-SNLI's original test set. The fine-tuned models show improved accuracy on our validation splits compared to the original U-SNLI training, though they underperform on the U-SNLI test set due to different label distributions. To test real-world applicability, we applied these models to 27 GRE test passages, all expected to be Not Inferred (`False`). Our fine-tuned models outperformed the U-SNLI-trained models by 11.1% (absolutely) lower in error rate, demonstrating their effectiveness in practical scenarios.

Conclusion In this work, we propose a novel approach to reduce hallucinations in natural language inference models by explicitly stating assumptions within premise-hypothesis pairs. Integrating predicate logic with natural deduction, our approach significantly enhanced the accuracy of NLI models. This improvement is not merely a step forward in model performance; it also represents a pivotal shift in data structure design, with far-reaching implications for the NLI community.

REFERENCES

Johan Bos. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pp. 277–286. College Publications, 2008. URL https://aclanthology.org/W08-2222.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8772–8779, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.774. URL https://aclanthology.org/2020.acl-main.774.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, 2006.

Nick Harvey and Brian Caton. *Coastal management in Australia*. University of Adelaide Press, 2010.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Jeffrey Simonoff. Statistics and data analysis (cor1-gb.1305), 2010.

URM STATEMENT

We acknowledge that all authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

A ASSUMPTION SIMPLIFICATION DETAILS

It is crucial to ensure that the assumption after simplification is as valid as $P \wedge \neg H$ is. To this end, we use NLTK Resolution Prover[1] for the turtle expression we derive in the method section.

```
[ 1] {-v_dance(z4), v_move(z4)}                          A
[ 2] {v_dance(z9)}                                       A
[ 3] {n_turtle(z10)}                                     A
[ 4] {Actor(z12,z11)}                                    A
[ 5] {-v_move(z13), -n_turtle(z14), -Actor(z13,z14)}     A
[ 6] {v_move(z9)}                                        (1, 2)
[ 7] {-n_turtle(z14), -Actor(z13,z14), -v_dance(z13)}    (1, 5)
[ 8] {-n_turtle(z14), -Actor(z13,z14)}                   (2, 7)
[ 9] {-Actor(z13,z14), -v_move(z13)}                     (3, 5)
[10] {-Actor(z13,z14), -v_dance(z13)}                    (1, 9)
[11] {-Actor(z13,z14)}                                   (2, 10)
[12] {-Actor(z13,z14), -v_dance(z13)}                    (3, 7)
[13] {-Actor(z13,z14)}                                   (2, 12)
[14] {-Actor(z13,z14)}                                   (3, 8)
[15] {-v_move(z13), -n_turtle(z14)}                      (4, 5)
[16] {-n_turtle(z14), -v_dance(z13)}                     (1, 15)
[17] {-n_turtle(z14)}                                    (2, 16)
[18] {-v_move(z13)}                                      (3, 15)
[19] {-v_dance(z13)}                                     (1, 18)
[20] {}                                                  (2, 19)
```

B WHAT DOES AN EXISTING NLI MODEL POSSIBLY HALLUCINATE?

Besides the few aforementioned examples, we list more details on possible hallucinations of an existing NLI model in Table 1 and 2.

B.1 HYPER-PARAMETER OF EXPERIMENTS

In the conducted experiment, the team employed a fine-tuning approach on a pre-existing Roberta architecture. The batch size was set at 16, and the training process was iterated over 5 epochs. The AdamW optimizer was utilized for this purpose.

[1]Tableau Prover presents similar results, though its proof is longer. https://www.nltk.org/howto/inference.html

Table 1: We first identify the assumptions for the provided label to be valid. Next, we evaluate the plausibility of these assumptions. On the right-most column, we present reasons or explanations for why these assumptions, which may have been implicitly suggested to the model during training, could lead to undesirable hallucinations. The images are generated from Dalle-mini `https://huggingface.co/spaces/dalle-mini/dalle-mini`.

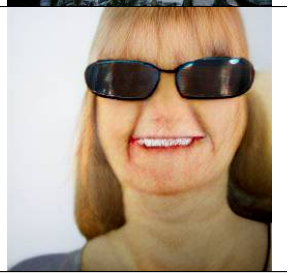
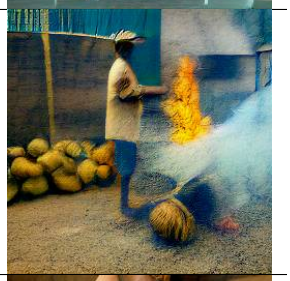
#	Expression	Why hallucination
1	**Premise** A car is loaded with items on the top. **Hypothesis** The car is a convertible. **U-SNLI label** Contradiction. **Assumption** Can convertible cars ever get loaded on their top? Yes, they can.	
2	**Premise** A family walking with a soldier. **Hypothesis** A group of people strolling. **U-SNLI label** Inferred. **Assumption** Do family members have to stroll, if they walk together with a soldier? No, they do not have to.	
3	**Premise** A man is wearing a blue and yellow racing uniform while holding a bottle. **Hypothesis** This guy is jumping rope. **U-SNLI label** Contradiction. **Assumption** Does one have to give up his bottle or stop jumping rope when he wears a blue and yellow racing uniform? No, he does not.	
4	**Premise** Young Asian girl is sitting on the ground in rubble. **Hypothesis** The young Asian girl is outside in the rubble. **U-SNLI label** Inferred. **Assumption** Must rubble be outside? No.	
5	**Premise** A woman wearing sunglasses is frowning. **Hypothesis** A woman wearing sunglasses is not smiling. **U-SNLI label** Contradiction. **Assumption** Have you ever seen someone smile with frowning? Yes, try a forced smile.	

Table 2: (Continue Table 1) We first identify the assumptions for the provided label to be valid. Next, we evaluate the plausibility of these assumptions. On the right-most column, we present reasons or explanations for why these assumptions, which may have been implicitly suggested to the model during training, could lead to undesirable hallucinations. The images are generated from Dalle-mini `https://huggingface.co/spaces/dalle-mini/dalle-mini`.

#	Expression	Why hallucination
6	**Premise** A statue at a museum that no seems to be looking at. **Hypothesis** Tons of people are gathered around the statue. **U-SNLI label** Contradiction. **Assumption** Could tons of people around a statue do something else? Yes, they could.	
7	**Premise** A blond-haired doctor and her African-American assistant looking through new medical manuals. **Hypothesis** A man is eating PB and J. **U-SNLI label** Contradiction. **Assumption** Can a doctor eat PB and J while doing something else? Yes, they can.	
8	**Premise** A young family enjoys feeling ocean waves lap at their feet. **Hypothesis** A family is out at a restaurant. **U-SNLI label** Contradiction. **Assumption** Are there any beach restaurants where ocean waves are just around? Yes, there could be.	
9	**Premise** A person wearing a straw hat, standing outside working a steel apparatus with a pile of coconuts on the ground. **Hypothesis** A person is burning a straw hat. **U-SNLI label** Contradiction. **Assumption** Can one wear a hat and burn another? Yes.	
10	**Premise** Man chopping wood with an axe. **Hypothesis** The man is outside. **U-SNLI label** Inferred. **Assumption** Must one chop wood outdoors/outside? Not really.	