MedNLI Is Not Immune: Natural Language Inference Artifacts in the Clinical Domain

Christine Herlihy & Rachel Rudinger

University of Maryland, College Park

DEPARTMENT OF COMPUTER SCIENCE

Overview

- Research questions:
  - Does MedNLI contain annotation artifacts?
  - If so, how can we characterize them?
  - How can we improve the dataset construction process in knowledge-intensive domains?

Overview

- Research questions:
  - Does MedNLI contain annotation artifacts?
  - If so, how can we characterize them?
  - How can we improve the dataset construction process in knowledge-intensive domains?
- Motivation:
  - Improve the robustness of downstream clinical decision support (CDS) models trained on MedNLI.

Natural Language Inference (NLI)

Given a premise, p, and associated hypothesis, h, is h:

- Definitely true? (entailed)
- Possibly true? (neutral)
- Definitely false? (contradictory)

As a classification task:

- $f : (p, h) \in P \times H \mapsto \ell \in \{\text{entailed, neutral, contradictory}\}$

Domain-Agnostic NLI & Annotation Artifacts

- Neural NLI models require large-scale, labeled datasets.

Domain-Agnostic NLI & Annotation Artifacts

- Neural NLI models require large-scale, labeled datasets.
- Crowd-worker constructed datasets facilitate progress:
  - SNLI: 570k pairs; image caption corpus [Bowman et al., 2015]
  - MultiNLI: 433k pairs; multiple genres [Williams et al., 2017]

Domain-Agnostic NLI & Annotation Artifacts

- Neural NLI models require large-scale, labeled datasets.
- Crowd-worker constructed datasets facilitate progress:
  - SNLI: 570k pairs; image caption corpus [Bowman et al., 2015]
  - MultiNLI: 433k pairs; multiple genres [Williams et al., 2017]
- However, they are found to contain *annotation artifacts*
  - Gururangan et al. [2018]; McCoy et al. [2019]; Poliak et al. [2018]; Tsuchiya [2018]

Domain-Agnostic NLI & Annotation Artifacts

- Neural NLI models require large-scale, labeled datasets.
- Crowd-worker constructed datasets facilitate progress:
  - SNLI: 570k pairs; image caption corpus [Bowman et al., 2015]
  - MultiNLI: 433k pairs; multiple genres [Williams et al., 2017]
- However, they are found to contain *annotation artifacts*
  - Gururangan et al. [2018]; McCoy et al. [2019]; Poliak et al. [2018]; Tsuchiya [2018]
- Artifacts pose risks: model performance may be overestimated

MedNLI

- Domain-specific NLI dataset [Romanov and Shivade, 2018]

Table 1 from Romanov and Shivade [2018]

MedNLI

- Domain-specific NLI dataset [Romanov and Shivade, 2018]
- Premises are drawn from the `Past Medical History` sections of a random subset of clinical notes from MIMIC-III [Goldberger et al., 2000; Johnson et al., 2016].

Table 1 from Romanov and Shivade [2018]

MedNLI

- Domain-specific NLI dataset [Romanov and Shivade, 2018]
- Premises are drawn from the `Past Medical History` sections of a random subset of clinical notes from MIMIC-III [Goldberger et al., 2000; Johnson et al., 2016].
- Physicians asked to write one hypothesis per label per premise

Table 1 from Romanov and Shivade [2018]

MedNLI

- Domain-specific NLI dataset [Romanov and Shivade, 2018]
- Premises are drawn from the ***Past Medical History*** sections of a random subset of clinical notes from MIMIC-III [Goldberger et al., 2000; Johnson et al., 2016].
- Physicians asked to write one hypothesis per label per premise

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction

Table 1: Examples from the development set of MedNLI

Table 1 from Romanov and Shivade [2018]

MedNLI Contains Annotation Artifacts

- Compare: hypothesis-only `fastText` classifer vs. *majority class baseline*.
- Results suggest artifacts exist, confirming findings of Romanov and Shivade [2018].
- The `fastText` model is most likely to *misclassify* **entailment** as *neutral* and **neutral** and **contradiction** as *entailment*.

	dev	test
majority class	33.3	33.3
fastText	**64.8**	**62.6**

Performance (micro F1) of *fastText* classifier.

	entailment	neutral	contradiction
entailment	**255**	151	68
neutral	126	**290**	58
contradiction	69	60	**345**

Confusion matrix for *fastText* classifier.

Gururangan et al. [2018]; Joulin et al. [2016]; Poliak et al. [2018]

Lexical Artifacts by Class

- Top 15 tokens by $PMI(\text{token}, \text{class}) = log_2 \frac{p(\text{token, class})}{p(\text{token},\cdot)p(\cdot,\text{class})}$

entailment	%	neutral	%	contradiction	%
just	0.25%	cardiogenic_shock	0.33%	no_history_of_cancer	0.27%
high_risk	0.26%	pelvic_pain	0.30%	no_treatment	0.27%
pressors	0.25%	joint_pain	0.30%	normal_breathing	0.27%
possible	0.26%	brain_injury	0.32%	no_history_of_falls	0.27%
elevated_blood_pressure	0.26%	delerium	0.30%	normal_heart_rhythm	0.28%
responsive	0.25%	intracranial_pressure	0.30%	health	0.26%
comorbidities	0.26%	smoking	0.42%	normal_head_ct	0.26%
spectrum	0.27%	obesity	0.41%	normal_vision	0.26%
steroid_medication	0.25%	tia	0.32%	normal_aortic_valve	0.27%
longer	0.26%	acquired	0.31%	bradycardic	0.26%
history_of_cancer	0.26%	head_injury	0.31%	normal_blood_sugars	0.27%
broad	0.26%	twins	0.30%	normal_creatinine	0.28%
frequent	0.25%	fertility	0.30%	cancer_history	0.26%
failed	0.26%	statin	0.30%	cardiac	0.33%
medical	0.29%	acute_stroke	0.30%	normal_chest	0.28%

% of *class* training hypotheses containing token; [Gururangan et al., 2018]

Physician-Annotator Heuristics

- **Hypernym heuristic**: Let
  $\mathcal{X} := \{\textit{condition, medication, finding, procedure, event}\}$

$$\left(\bigvee \mathcal{X} \in p\right) \wedge (c = \textsc{entailment}) \wedge (\exists (t, t') \in p \times h \text{ s.t. } t <: t')$$

Physician-Annotator Heuristics

- **Hypernym heuristic**: Let $\mathcal{X} := \{\textit{condition, medication, finding, procedure, event}\}$

$$\left(\bigvee \mathcal{X} \in p\right) \wedge (c = \textsc{entailment}) \wedge (\exists (t, t') \in p \times h \text{ s.t. } t <: t')$$

- **Probable cause heuristic**:

$$(\textit{condition} \in p) \wedge (c = \textsc{neutral}) \wedge (h \text{ provides causal explanation for } p)$$

Physician-Annotator Heuristics

- **Hypernym heuristic**: Let $\mathcal{X} := \{\textit{condition, medication, finding, procedure, event}\}$

$$\left(\bigvee \mathcal{X} \in p\right) \wedge (c = \textsc{entailment}) \wedge (\exists (t, t') \in p \times h \text{ s.t. } t <: t')$$

- **Probable cause heuristic**:

$$(\textit{condition} \in p) \wedge (c = \textsc{neutral}) \wedge (h \text{ provides causal explanation for } p)$$

- **Everything's fine heuristic**:

$$(\textit{condition} \in p \vee \textit{finding} \in p) \wedge (c = \textsc{contradiction}) \wedge (h \implies \neg p)$$

Physician-Annotator Heuristics: χ^2 Results

heuristic	χ^2	p-value	top class
hypernym	59.15	1.4e-13‡	entail (45.2%)
probable cause	111.05	7.7e-25‡	neutral (57.8%)
everything fine	874.71	1.1e-190‡	contradict (83.8%)

Results of χ^2 test statistic by heuristic, computed using the combined MedNLI dataset (‡ $p < 0.001$, † $p < 0.01$, * $p < 0.5$). Top class presented with % of heuristic-satisfying pairs.

Adversarial Filtering

We employ `AFLite` to create *easy* and *difficult* partitions of MedNLI [Bras et al., 2020; Sakaguchi et al., 2020]:

	model	eval dataset	full	easy (Δ)	difficult (Δ)
no premise	majority class	dev	0.33	0.34 (+0.01)	0.35 (+0.02)
no premise	majority class	test	0.33	0.35 (+0.02)	0.37 (+0.04)
no premise	`fastText`	dev	0.65	0.67 (+0.02)	0.46 (-0.19)
no premise	`fastText`	test	0.63	0.65 (+0.02)	0.4 (-0.23)
with premise	majority class	dev	0.33	0.45 (+0.12)	0.36 (+0.03)
with premise	majority class	test	0.33	0.48 (+0.15)	0.37 (+0.04)
with premise	`fastText`	dev	0.53	0.6 (+0.07)	0.43 (-0.1)
with premise	`fastText`	test	0.51	0.55 (+0.04)	0.4 (-0.11)

Performance (micro F1-score) for the majority class baseline and `fastText` classifiers, with and without premise, by partition (e.g., *full, easy, difficult*).

See crherlihy/clinical_nli_artifacts for code and partition ids.

NLI Dataset Construction in Knowledge-Intensive Domains

- Zellers et al. [2019] advocate *adversarial dataset construction*; this may not scale in domains requiring expert validation.

NLI Dataset Construction in Knowledge-Intensive Domains

- Zellers et al. [2019] advocate *adversarial dataset construction*; this may not scale in domains requiring expert validation.
- In complex domains, information-rich inferences are more useful for downstream tasks than correct but trivial inferences.

NLI Dataset Construction in Knowledge-Intensive Domains

- Zellers et al. [2019] advocate *adversarial dataset construction*; this may not scale in domains requiring expert validation.
- In complex domains, information-rich inferences are more useful for downstream tasks than correct but trivial inferences.
- One option: adopt a *mechanism design perspective* to incentivize the production of hypotheses with high downstream utility [Ho et al., 2015; Liu and Chen, 2017]

NLI Dataset Construction in Knowledge-Intensive Domains

- Zellers et al. [2019] advocate *adversarial dataset construction*; this may not scale in domains requiring expert validation.
- In complex domains, information-rich inferences are more useful for downstream tasks than correct but trivial inferences.
- One option: adopt a *mechanism design perspective* to incentivize the production of hypotheses with high downstream utility [Ho et al., 2015; Liu and Chen, 2017]
- Another option: narrow the generative scope (and room for reliance on artifacts) by defining a set of inferences deemed to be useful for a specific task.

References

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*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL *https://www.aclweb.org/anthology/D15-1075*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases, 2020.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL *https://www.aclweb.org/anthology/N18-2017*.

Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 419–429, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741102. URL *https://doi.org/10.1145/2736277.2741102*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Yang Liu and Yiling Chen. Machine-learning aided peer prediction. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 63–80, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279. doi: 10.1145/3033274.3085126. URL *https://doi.org/10.1145/3033274.3085126*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL *https://www.aclweb.org/anthology/P19-1334*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL *https://www.aclweb.org/anthology/S18-2023*.

Alexey Romanov and Chaitanya Shivade. Lessons from Natural Language Inference in the Clinical Domain. *CoRR*, abs/1808.06752, 2018. URL *http://arxiv.org/abs/1808.06752*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, Apr. 2020. doi: 10.1609/aaai.v34i05.6399. URL *https://ojs.aaai.org/index.php/AAAI/article/view/6399*.

Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL *https://www.aclweb.org/anthology/L18-1239*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *CoRR*, abs/1704.05426, 2017. URL *http://arxiv.org/abs/1704.05426*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL *https://www.aclweb.org/anthology/P19-1472*.