Beyond the Tip of the Iceberg:

Assessing Coherence of Text Classifiers

Shane Storks & Joyce Chai

(he/him)

Situated Language and Embodied Dialogue (SLED)

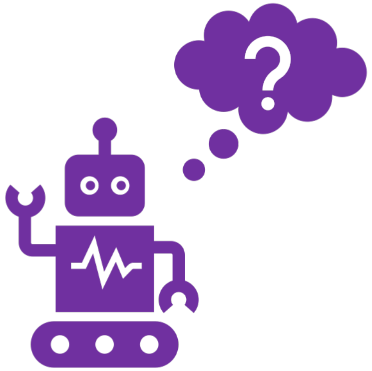
University of Michigan, Computer Science and Engineering Division

sstorks@umich.edu

Findings of EMNLP 2021 Short Paper

Introduction

- Today, language understanding is often boiled down to **high-level classification tasks**



Textual Entailment

Dialog:

A_1: Yeah, yeah. Is that why you like aerobics classes, because you're not, sort of, someone else is doing the counting for you, so,

B_1: Yeah.

…

B_2: And, someone else is telling me, okay, you know, let's move this way, let's move that way,

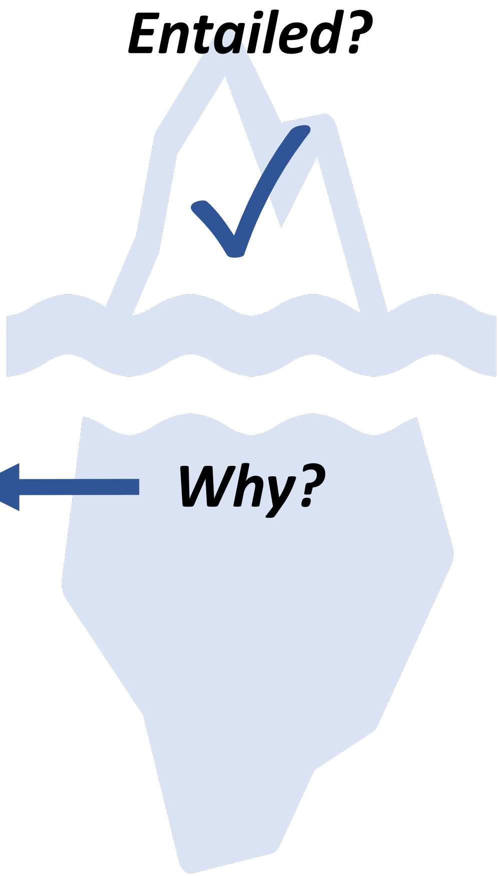
A_2: Uh-huh, uh-huh.

B_3: instead of me having to think about it so much.

…

Hypothesis:

Speaker **B** likes the aspect of Aerobics that someone else is leading.



Zhang, C., & Chai, J.Y. (2010). Towards Conversational Entailment: An Empirical Investigation. In EMNLP 2010.

Coherence

Dialog:

A_1: Well, ironically enough I'm sitting here with a cast on my leg because I resumed an aerobics class the night before last.
B_1: Oh, no.
A_2: I ripped the ligaments in my right ankle.

✗ ✗ ✓

Hypothesis:

Speaker **A** ripped the ligaments in her ankle at aerobics class.

Accuracy:
full-text correct

Strict Coherence:
all spans correct

Lenient Coherence:
average accuracy on spans

Zhang, C., & Chai, J.Y. (2010). Towards Conversational Entailment: An Empirical Investigation. In EMNLP 2010.

Empirical Results

- Despite high accuracy from SOTA text classifiers, we see significant drops from accuracy to coherence across the board!

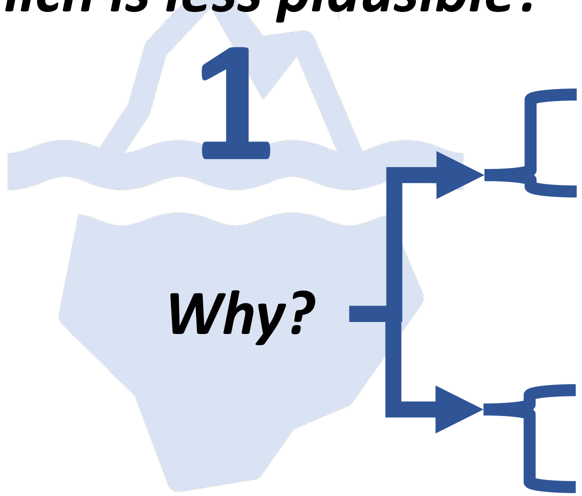
CE, *test*:

Model	**Accuracy** (%)	**Strict Coherence** (Δ; %)		**Lenient Coherence** (Δ; %)	
majority	57.8	–		–	
BERT	55.8	28.5	(-27.3)	35.7	(-20.1)
RoBERTa	70.9	39.0	(-31.9)	47.5	(-23.4)
↪ + MNLI	78.5	50.6	(-27.9)	58.2	(-20.3)
DeBERTa	67.4	37.2	(-30.2)	45.2	(-22.2)

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692
Williams, A., Nangia, N., & Bowman, S.R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. NAACL HLT 2017.
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv: 2006.03654.

Abductive Reasoning in narrative Texts (ART)

Which is less plausible?



Story 1:

Kelly wanted to try out for soccer this year.
Kelly tried out for the soccer team but was cut.
Kelly celebrated by getting pizza.

Story 2:

Kelly wanted to try out for soccer this year.
Kelly made it onto the team.
Kelly celebrated by getting pizza.

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W., & Choi, Y. (2020). Abductive commonsense reasoning. In ICLR 2020.

Empirical Results

- Despite high accuracy from SOTA text classifiers, we see significant drops from accuracy to coherence across the board!

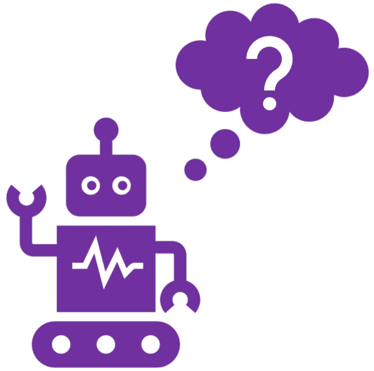
ART, *validation*:

Model	**Accuracy** (%)		**Strict Coherence** (Δ; %)		**Lenient Coherence** (Δ; %)	
majority	55.0	(50.1)	–		–	
BERT	66.7	(66.7)	42.3	(-24.4)	43.7	(-23.0)
RoBERTa	87.8	(84.2)	55.0	(-32.8)	59.3	(-28.5)
DeBERTa	88.4	(85.7)	59.8	(-28.6)	61.8	(-26.6)

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W., & Choi, Y. (2020). Abductive commonsense reasoning. In ICLR 2020.
Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv: 2006.03654.

Conclusion

- We proposed a quick, effective, and versatile paradigm for measuring the coherence of a text classifier's predictions
  - Unlock strong insights from small amount of annotation!



Thank you!