Stress Test Evaluation for Natural Language Inference

Aakanksha Naik*, **Abhilasha Ravichander***, Norman Sadeh, Carolyn Rose, Graham Neubig

Carnegie Mellon University
Language Technologies Institute

institute for SOFTWARE RESEARCH

Natural Language Inference

(a.k.a Recognizing Textual Entailment)

Premise: Stimpy was a little cat who believed he could fly

Hypothesis: Stimpy could fly

(Fyodorov, 2000, Condoravdi, 2003, Bos and Markert, 2005,
Dagan et.al, 2006, McCartney and Manning, 2009)

Natural Language Inference

(a.k.a Recognizing Textual Entailment)

Premise: Stimpy was a little cat who believed he could fly

Given a premise, determine whether a hypothesis is
True (**entailment**),
False (**contradiction**),
Undecided (**neutral**)

Hypothesis: Stimpy could fly

(Fyodorov, 2000, Condoravdi, 2003, Bos and Markert, 2005,
Dagan et.al, 2006, McCartney and Manning, 2009)

Natural Language Inference

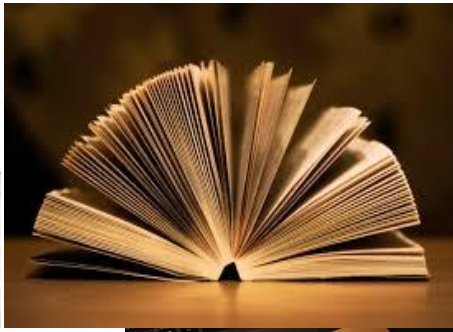
(a.k.a Recognizing Textual Entailment)

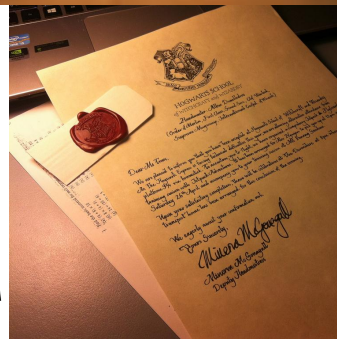
Benchmark task for Natural Language Understanding

- **Prevalent View:** To perform well at NLI, models must
  - learn good sentence representations: “handle nearly the full complexity of compositional semantics” (Williams et al, 2018)
  - reason over “difficult” phenomena like lexical entailment, quantification, coreference, tense, belief, modality, lexical and syntactic ambiguity (Dagan et al, 2009; McCartney and Manning, 2009; Marelli et al, 2014; Williams et al, 2018)

Natural Language Inference

(a.k.a Recognizing Textual Entailment)





MultiNLI

- Text from 10 genres!
- Covers written & spoken english
- Longer, more complex sentences
- Variety of linguistic phenomena
- Sentence-encoder SOTA: **74.5** % (Nie and Bansal, 2017)*

* All results shown for matched. Refer to paper for further details

Motivation

Neural networks can solve nearly ¾ examples from the challenging MultiNLI dataset!

Motivation

Neural networks can solve nearly ¾ examples from the challenging MultiNLI dataset!

But, more difficult cases occur rarely and are masked in traditional evaluation

↓

Optimistic estimate of model performance

Motivation

Neural networks can solve nearly ¾ examples from the challenging MultiNLI dataset!

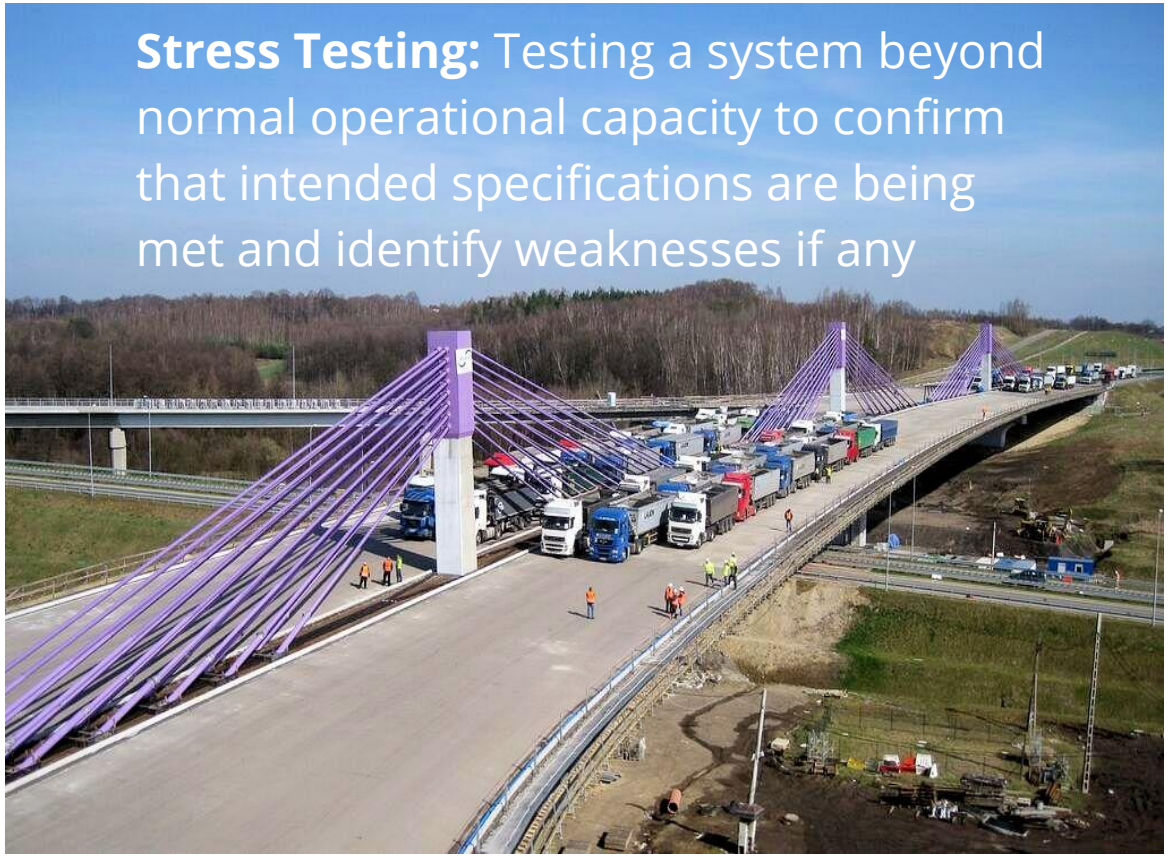
But, more difficult cases occur rarely and are masked in traditional evaluation

↓

Optimistic estimate of model performance

We want to figure out whether our systems have the ability to make real inferential decisions, and if so, to what extent.

What are Stress Tests?



For NLI

Building large-scale diagnostic datasets to exercise models on their weaknesses and better understand their capabilities.

Why Stress Tests?

Reward system **ability to reason** about task instead of encouraging reliance on misleading correlations in datasets.

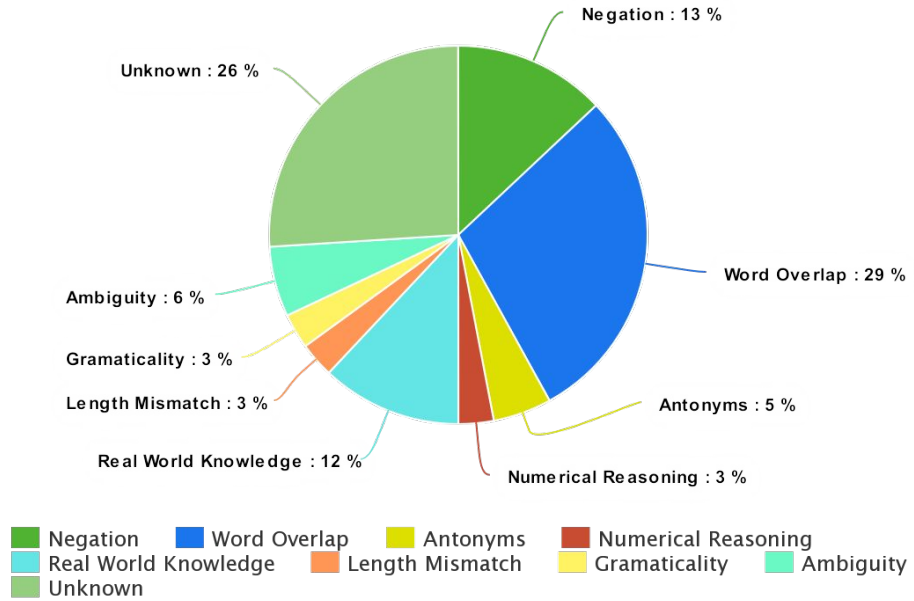
“Sanity checking” for NLP models

Analyze **strengths and weaknesses** of various models

Fine-grained **phenomenon-by-phenomenon** evaluation scheme

Weaknesses of SOTA NLI Models

- To construct stress tests, we must first identify **“bugs”** (potential weaknesses)
- Analyzed errors of Nie & Bansal (2017) (best-performing single model)



Word Overlap (29%)

Premise : And, could it not result in a decline in Postal Service volumes across–the–board?

Hypothesis :There may not be a decline in Postal Service volumes across–the–board.

Neutral → Entailment

Negation (13%)

Premise : Enthusiasm for Disney's Broadway production of The Lion King dwindles.

Hypothesis : The broadway production of The Lion King is no longer enthusiastically attended.

Entailment → Contradiction

Length Mismatch (3%)

Premise : So you know well a lot of the stuff you hear coming from South Africa now and from West Africa that's considered world music because it's not particularly using certain types of folk styles.

Hypothesis : They rely too heavily on the types of folk styles.

Contradiction → Neutral

Numerical Reasoning (3%)

Premise : Deborah Pryce said Ohio Legal Services in Columbus will receive a $200,000 federal grant toward an online legal self-help center.

Hypothesis : A $900,000 federal grant will be received by Missouri Legal Services, said Deborah Pryce.

Contradiction → Entailment

Antonymy (5%)

Premise : “Have her show it,” said Thorn

Hypothesis : Thorn told her to hide it.

Contradiction → Entailment

Grammaticality (3%)

Premise : So if there are something interesting or something worried, please give me a call at any time.

Hypothesis : The person is open to take a call anytime.

Contradiction → Neutral

Real World Knowledge (12%)

Premise : It was still night.

Hypothesis : The sun hadn’t risen yet, for the moon was shining daringly in the sky.

Entailment → Neutral

Ambiguity (6%)

Premise: Outside the cathedral you will find a statue of John Knox with Bible in hand.

Hypothesis: John Knox was someone who read the Bible.

Entailment → Neutral

Unknown (26%)

Premise : We’re going to try something different this morning, said Jon.

Hypothesis : Jon decided to try a new approach.

Entailment → Contradiction

Constructing Stress Tests

Competence Tests

Evaluate model **ability** to reason about quantities and understand antonyms

Target error categories: antonymy, numerical reasoning

Construction framework: Heuristic rules, external knowledge sources

Distraction Tests

Evaluate model **robustness** to shallow distractions

Target error categories: word overlap, negation, length mismatch

Construction framework: label-preserving perturbations using propositional logic

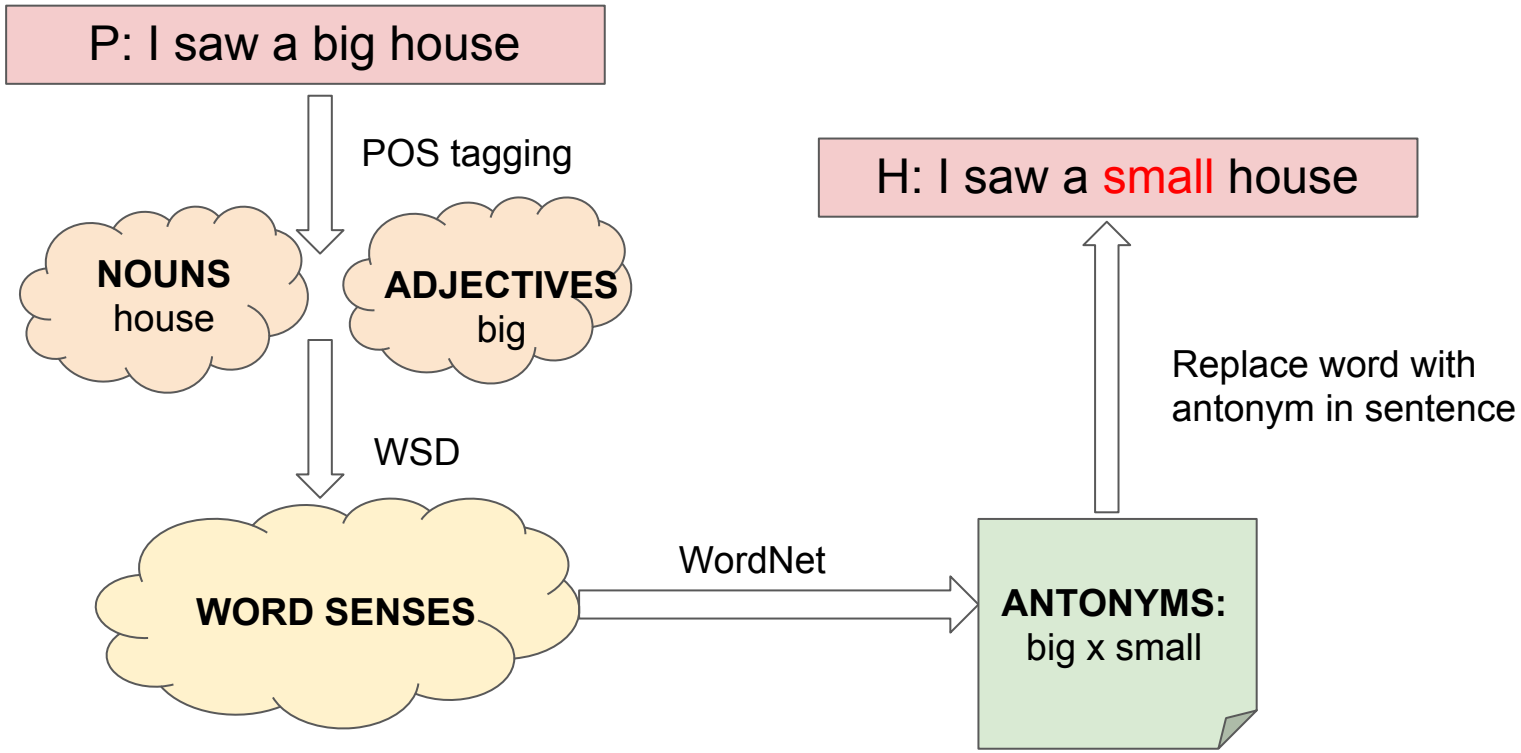
Noise Tests

Evaluate model **robustness** to noise in data

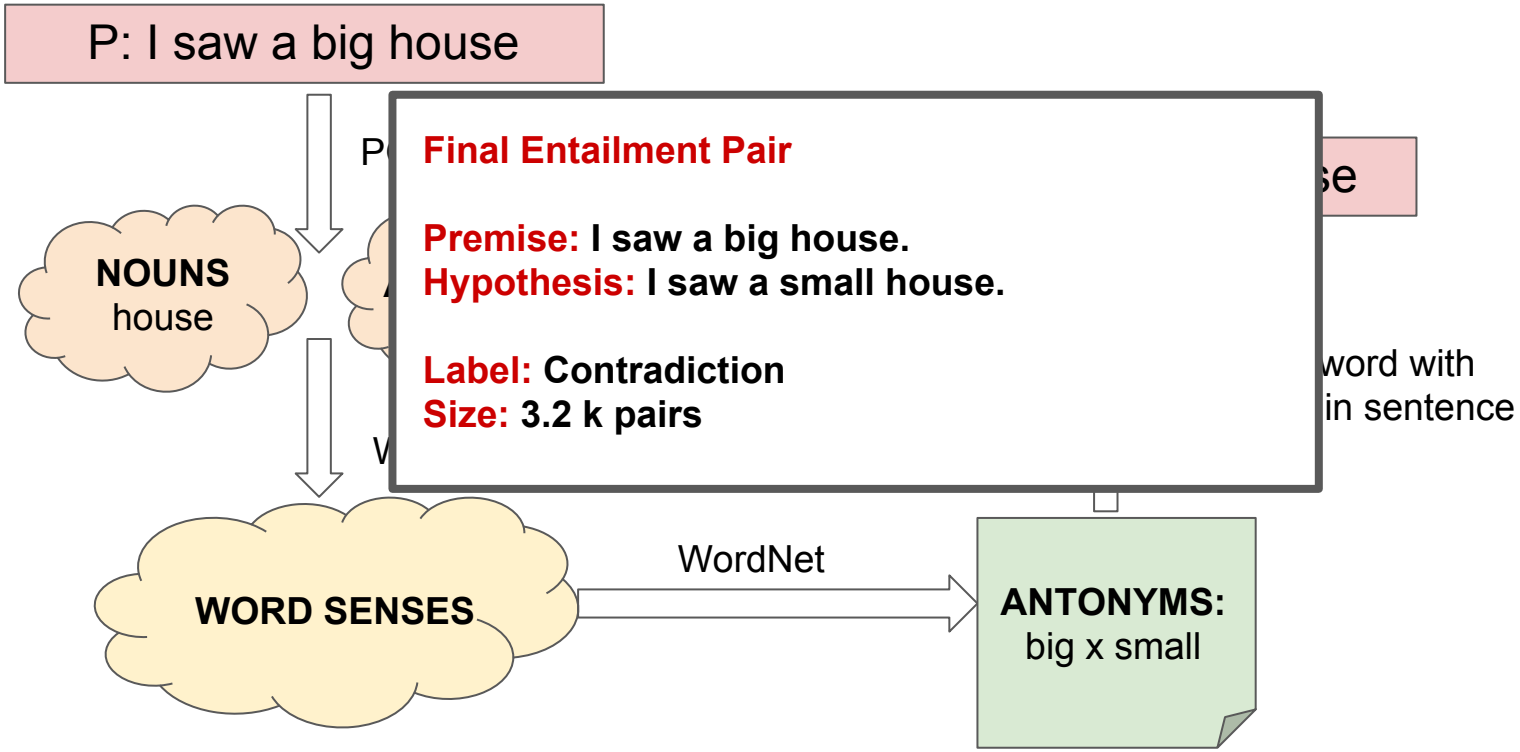
Target error categories: grammaticality

Construction framework: Random perturbation

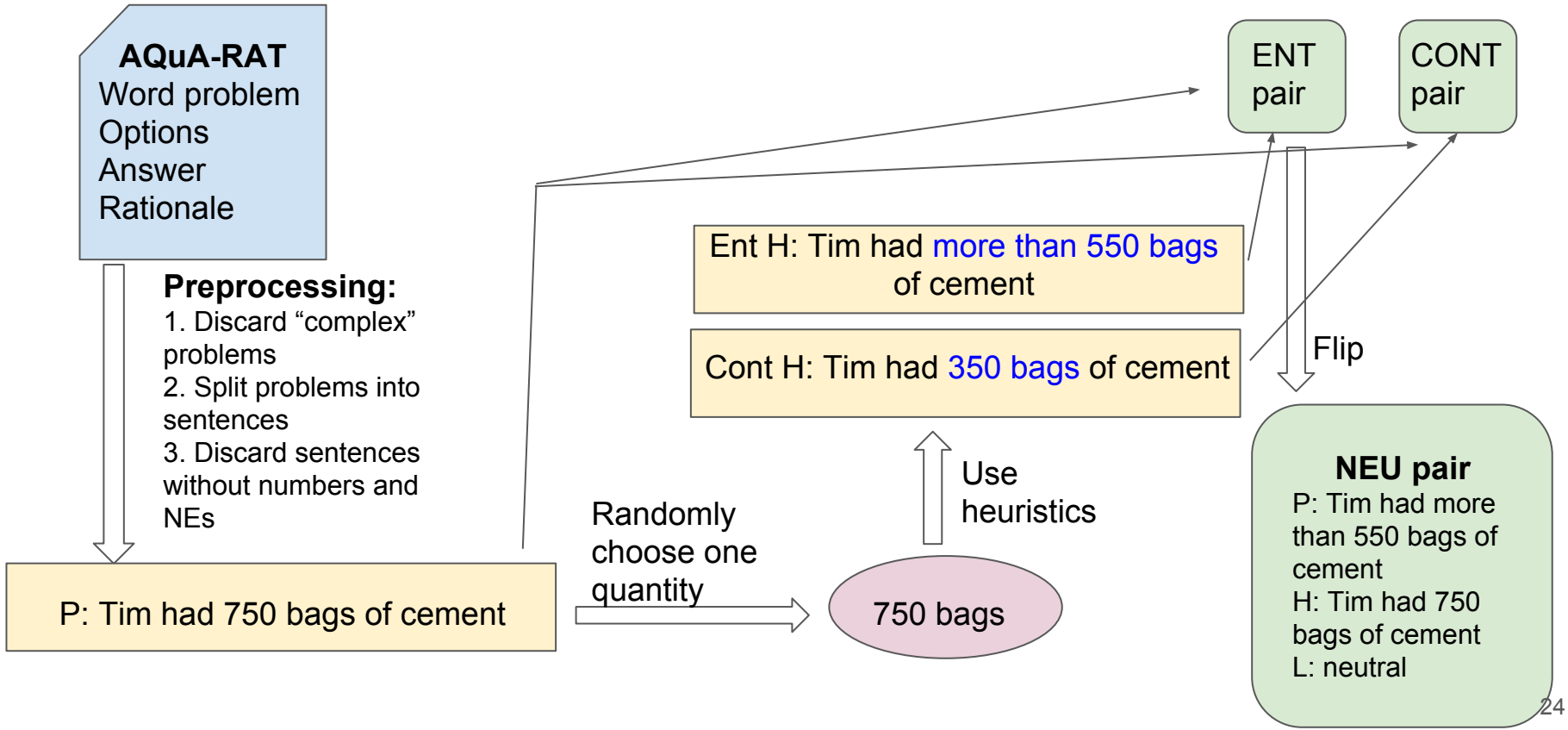
Constructing Competence Tests: Antonymy



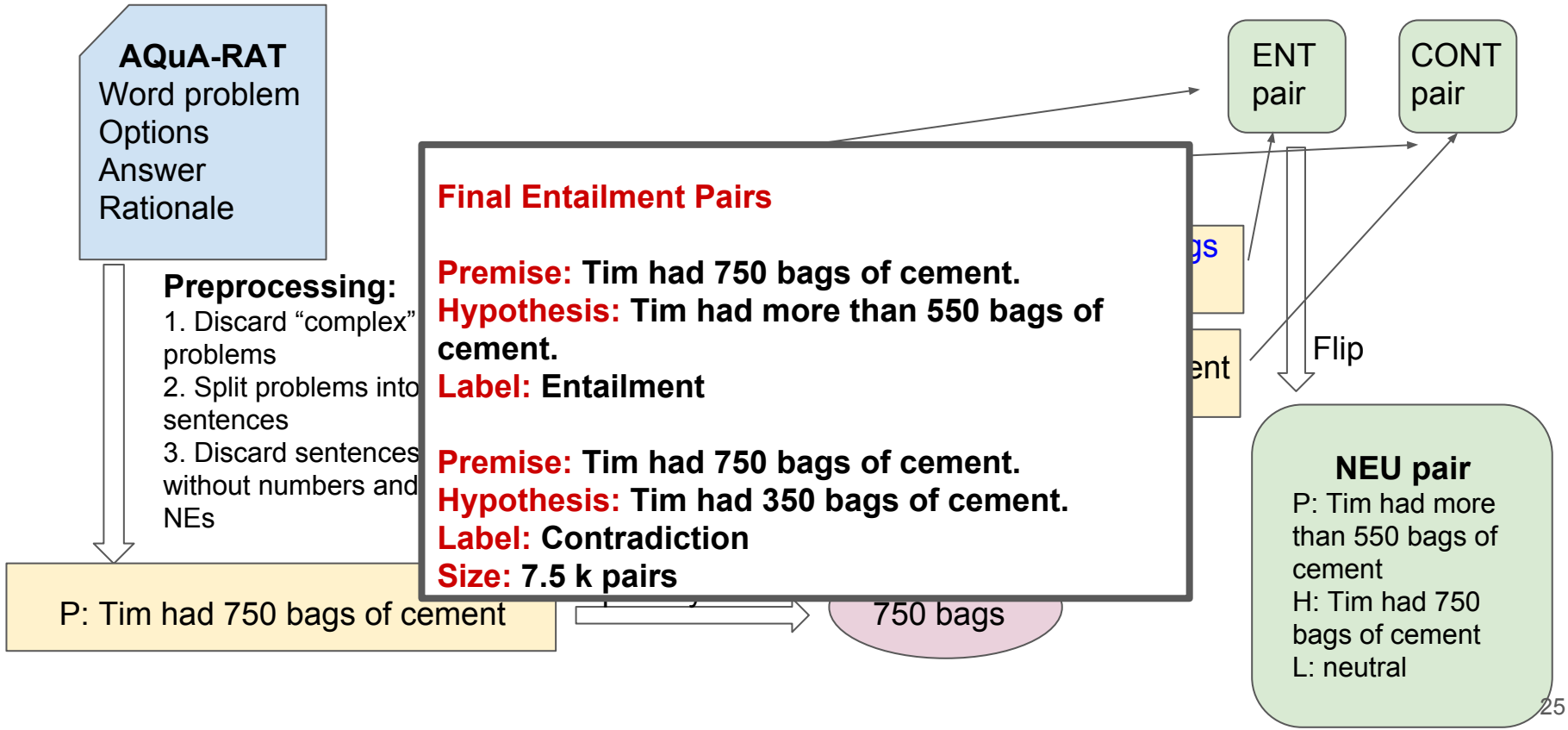
Constructing Competence Tests: Antonymy



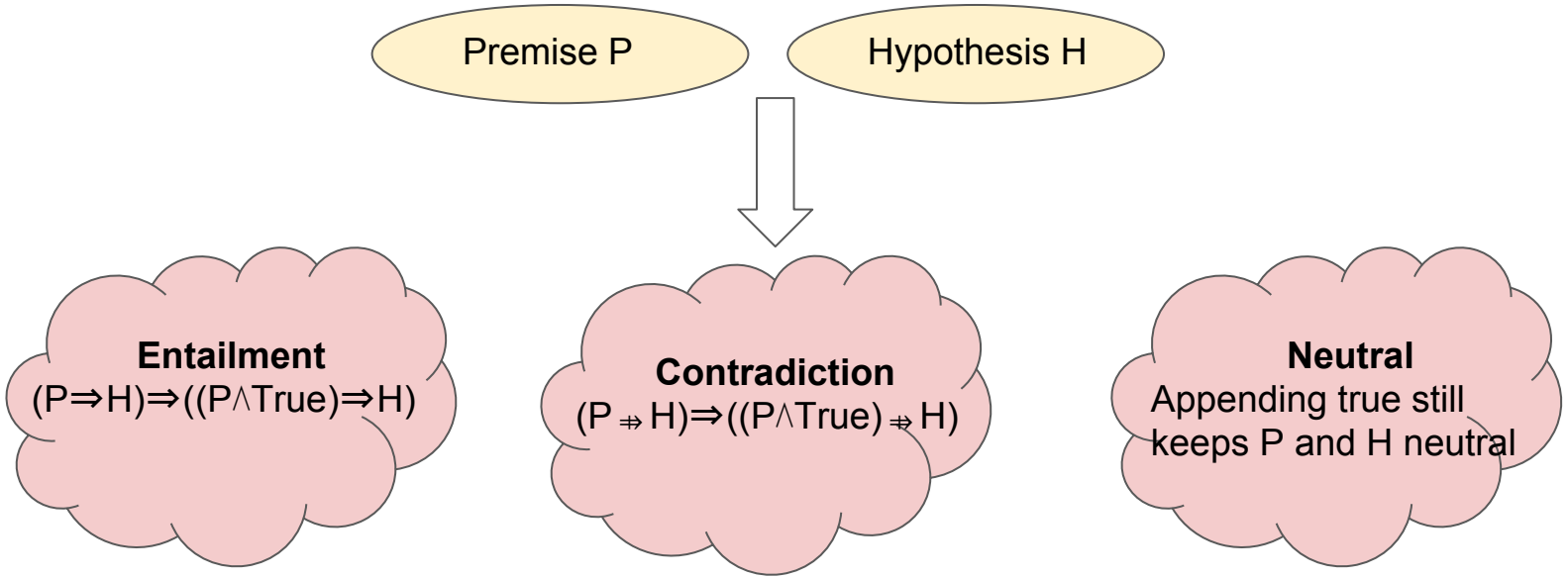
Constructing Competence Tests: Numerical Reasoning



Constructing Competence Tests: Numerical Reasoning



Constructing Distraction Tests



Logic Framework for Distraction Test Construction: Appending tautology to either premise or hypothesis

Constructing Distraction Tests (Cont.)

Word Overlap

- Tautology: true is true
- Append to hypothesis
- Reducing word overlap

Negation

- Tautology: false is not true
- Append to hypothesis
- Introducing strong negation

Length Mismatch

- Tautology: (true is true)*5
- Append to premise
- Add irrelevant information to premise

Constructing Distraction Tests (Cont.)

Word Overlap

Final Entailment Pair

Premise: Possibly no other country has had such a turbulent history and true is true.
Hypothesis: The country's history has been turbulent.

Label: Entailment

Negation

Final Entailment Pair

Premise: Possibly no other country has had such a turbulent history and false is not true.
Hypothesis: The country's history has been turbulent.

Label: Entailment

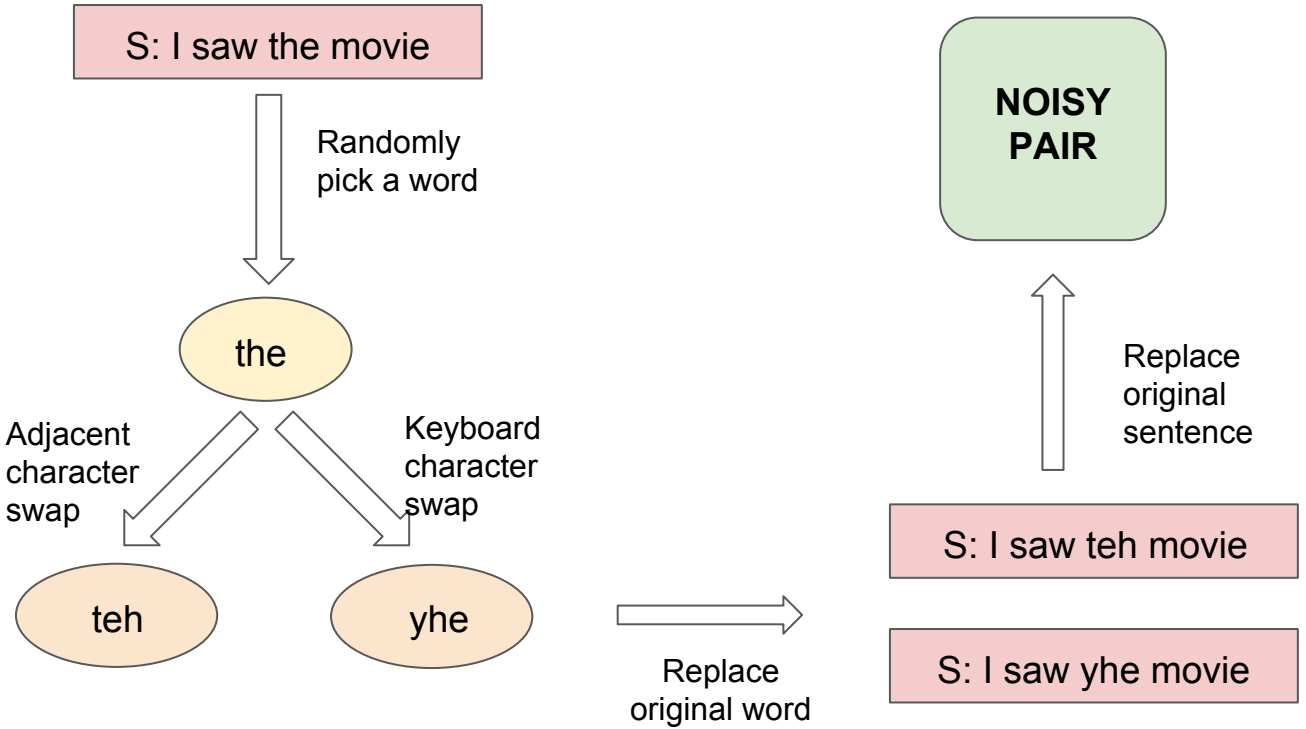
Length Mismatch

Final Entailment Pair

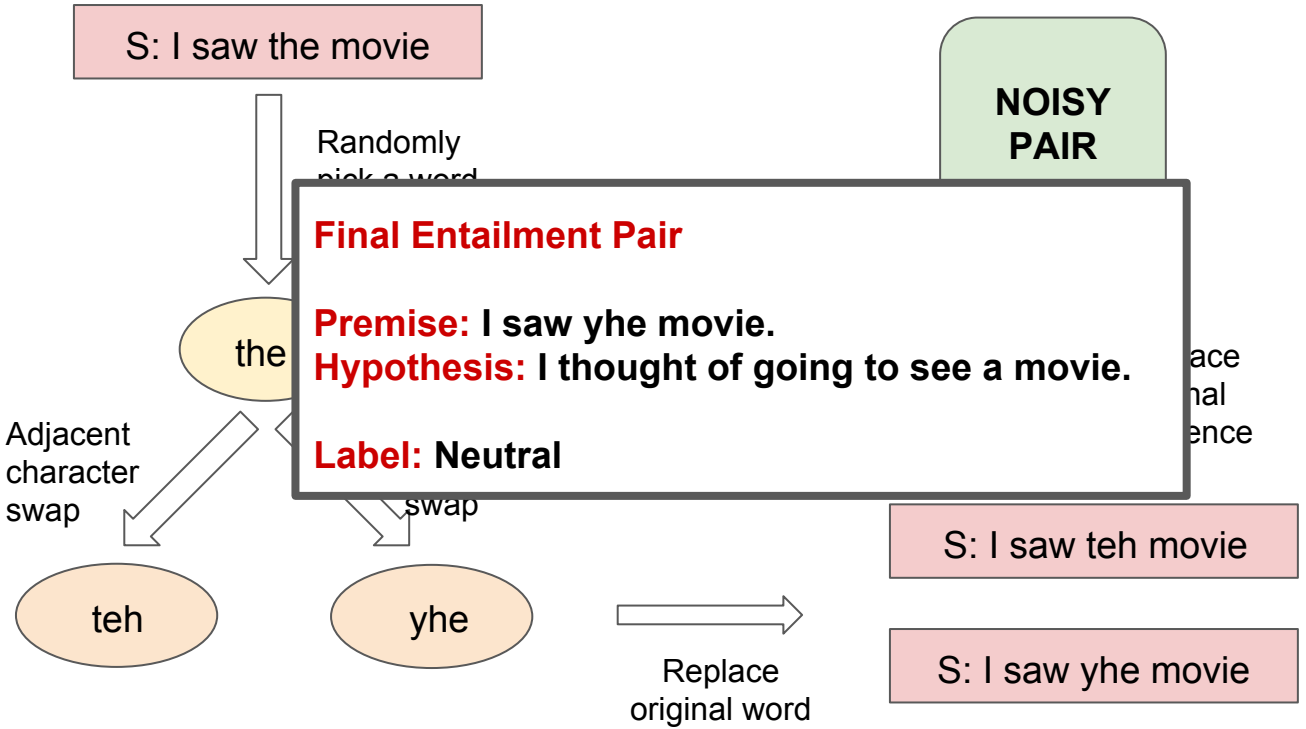
Premise: Possibly no other country has had such a turbulent history and true is true and true is true and true is true and true is true and true is true.
Hypothesis: The country's history has been turbulent.

Label: Entailment

Constructing Noise Tests



Constructing Noise Tests

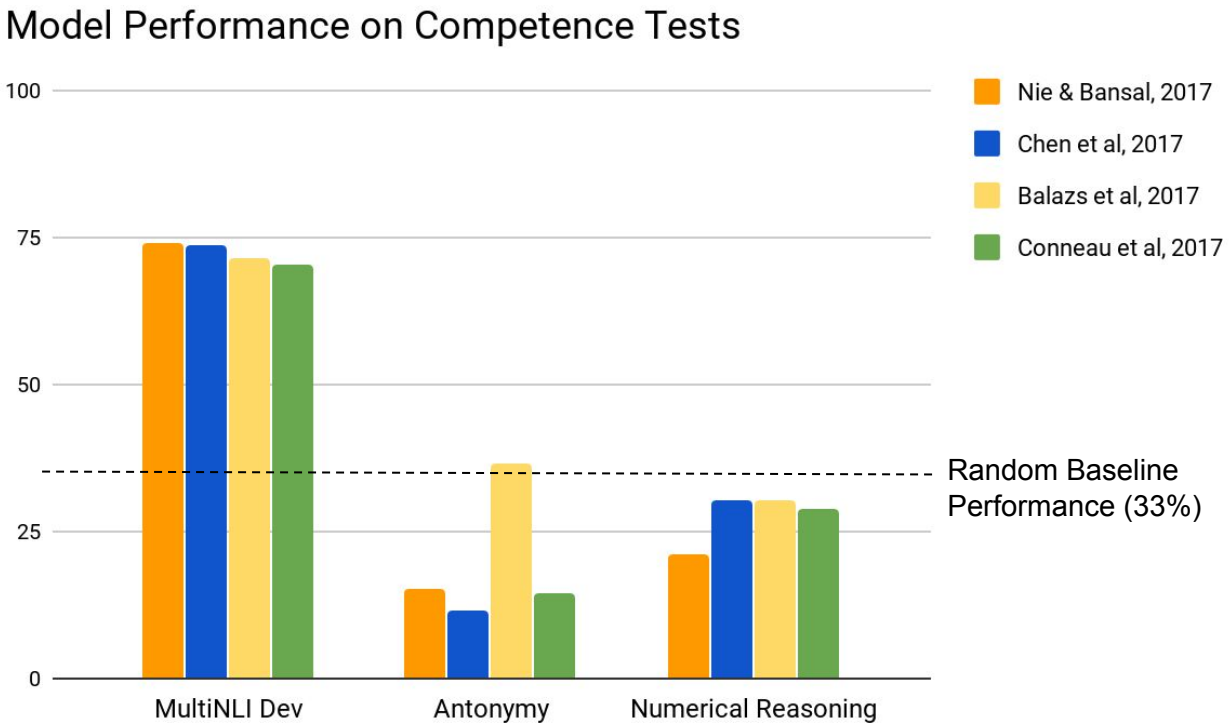


Experimental Setup: Models

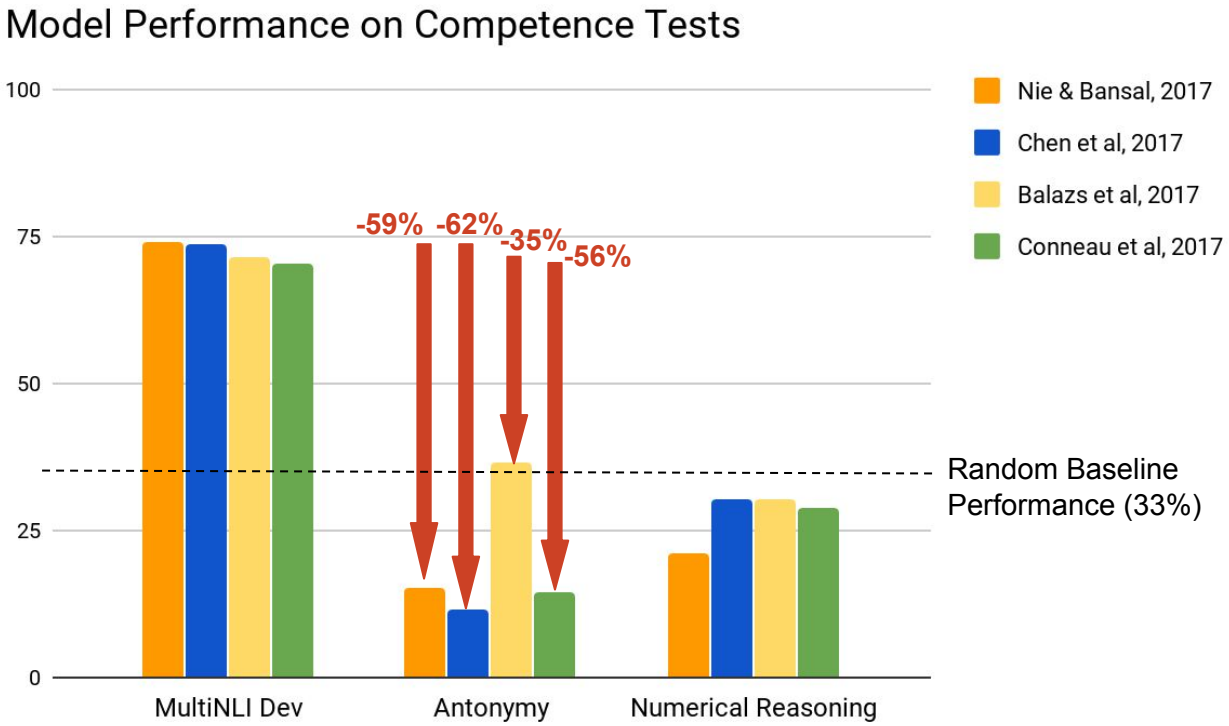
SOTA

- **Nie & Bansal, 2017:** Shortcut-Stacked BiLSTMs
- **Chen et al, 2017:** Shortcut-Stacked BiLSTMs + Char CNN
- **Balazs et al, 2017:** BiLSTMs with Inner Attention
- **Conneau et al, 2017:** BiLSTMs with Max Pooling

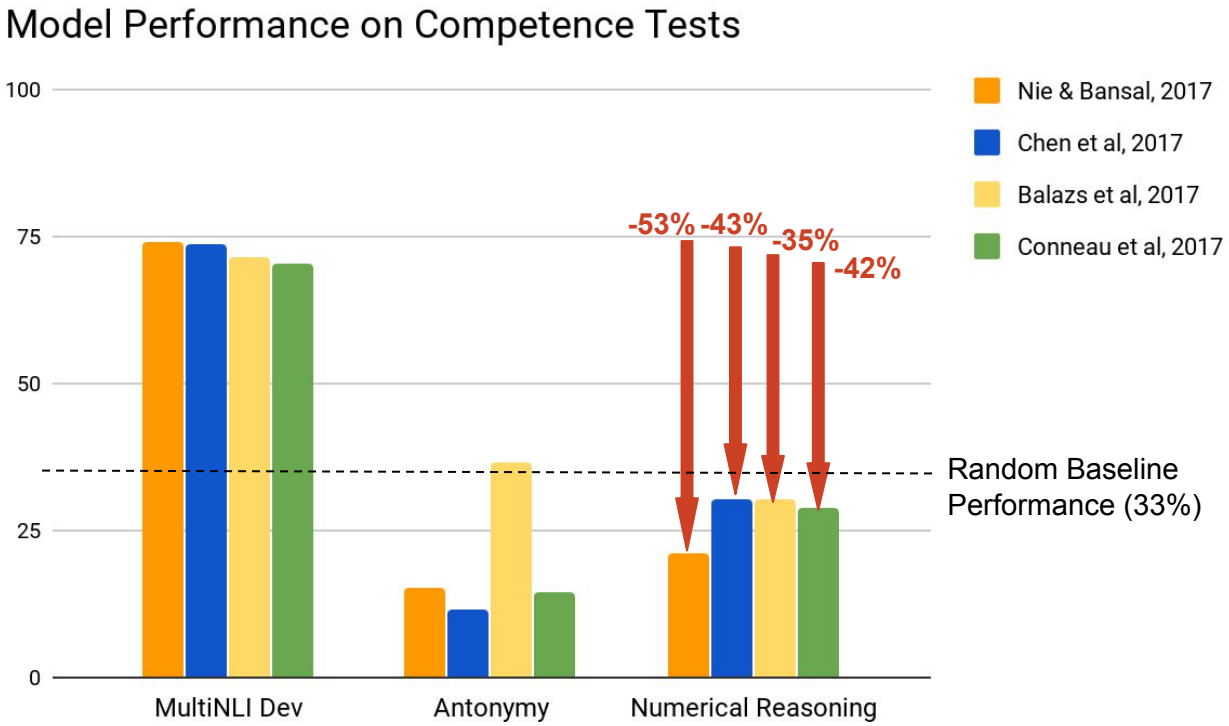
Results on Competence Tests



Results on Competence Tests



Results on Competence Tests



Performance Analysis on Competence Tests

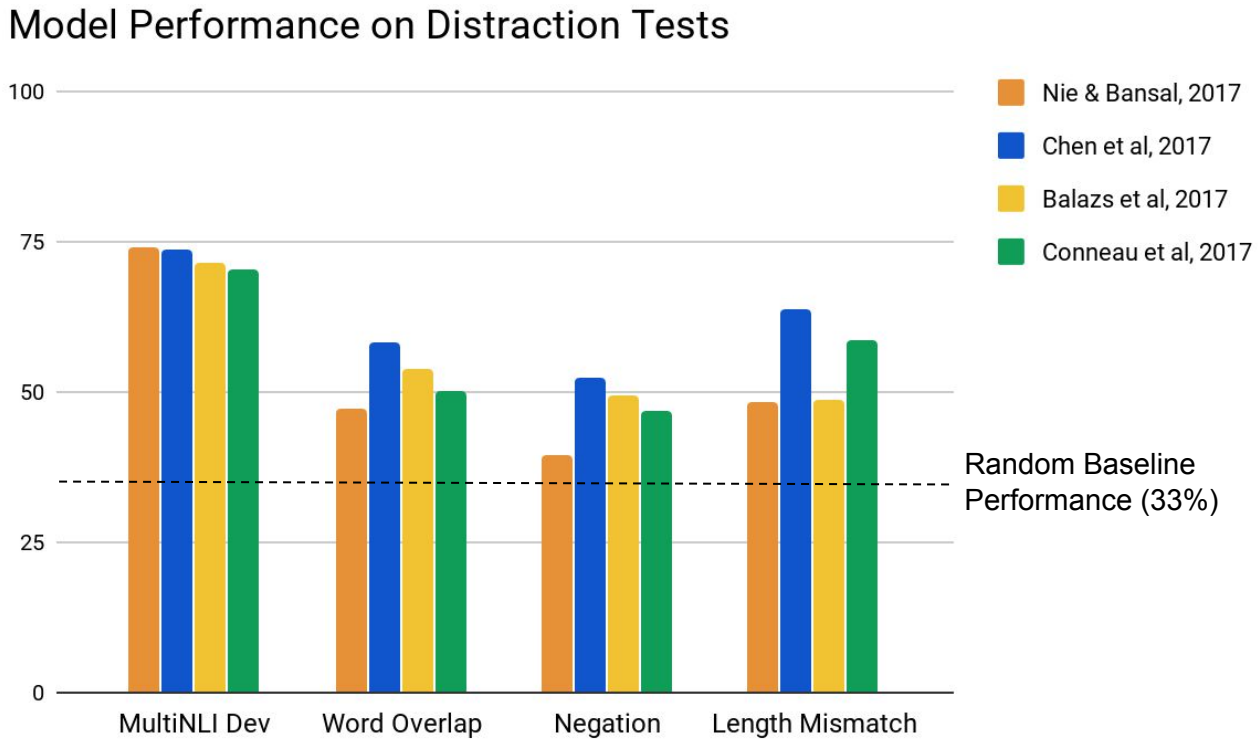
ANTONYMY

- All models overpredict entailment relations: **86.4**% of all errors are **Contradiction → Entailment** predictions!
- Some success on “easy” antonym-pairs (seen before in training data)
- Antonym pairs recognized by weakest model occur nearly twice as often in training data as pairs recognized by stronger model

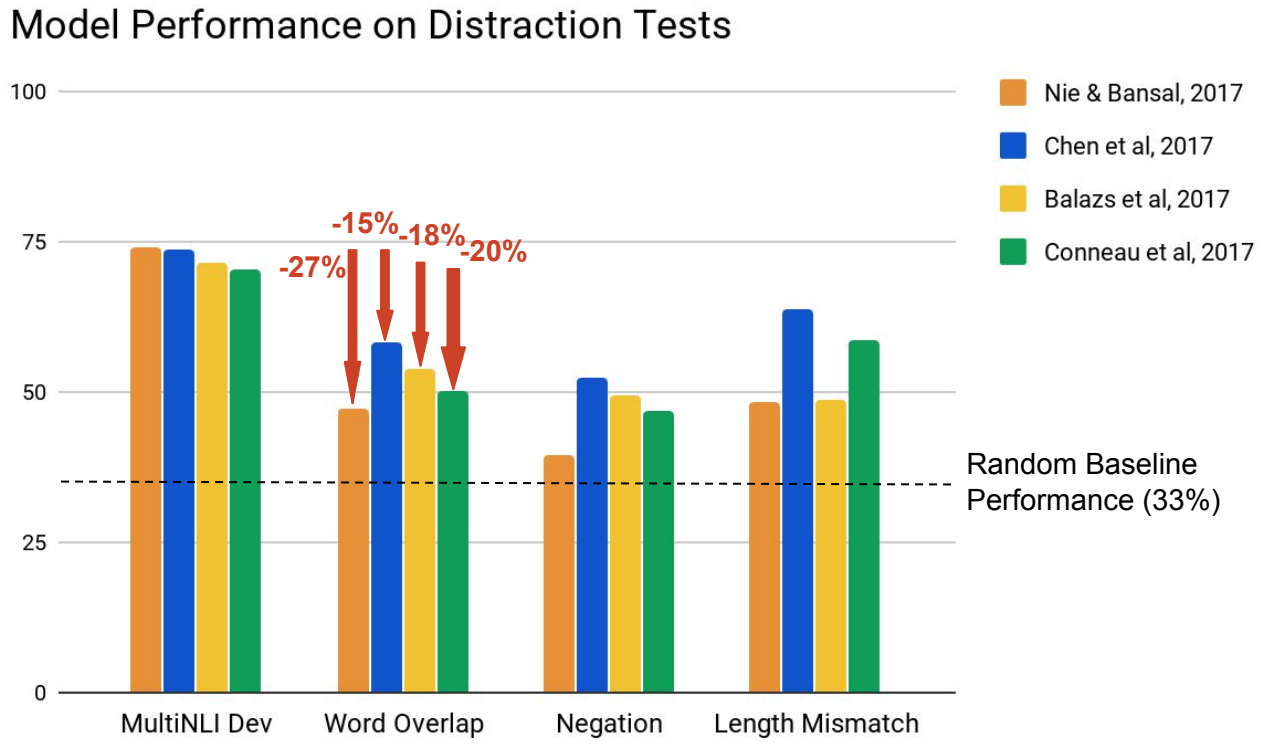
NUMERICAL REASONING

- All models overpredict entailment relations. **78.3%** of all errors are **Neutral → Entailment** or **Contradiction → Entailment**
- Models rely on lexical cues due to artifacts in training data
- No quantitative reasoning performed

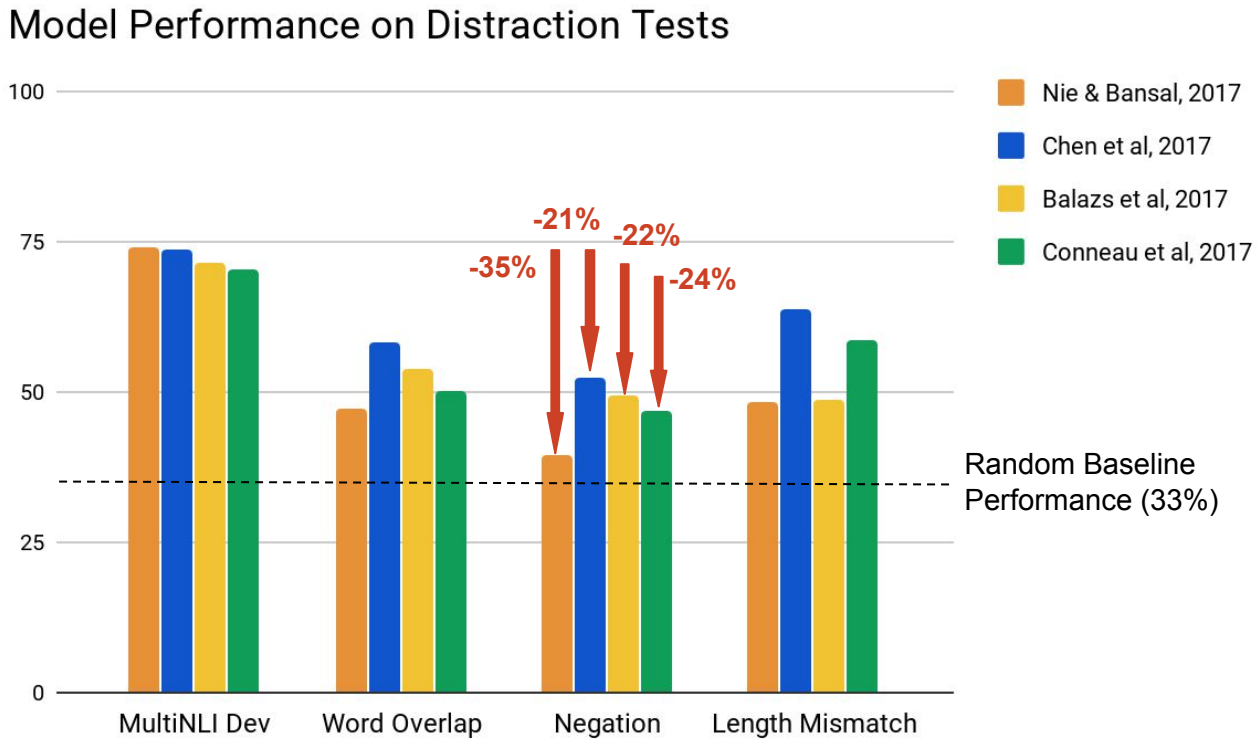
Results on Distraction Tests



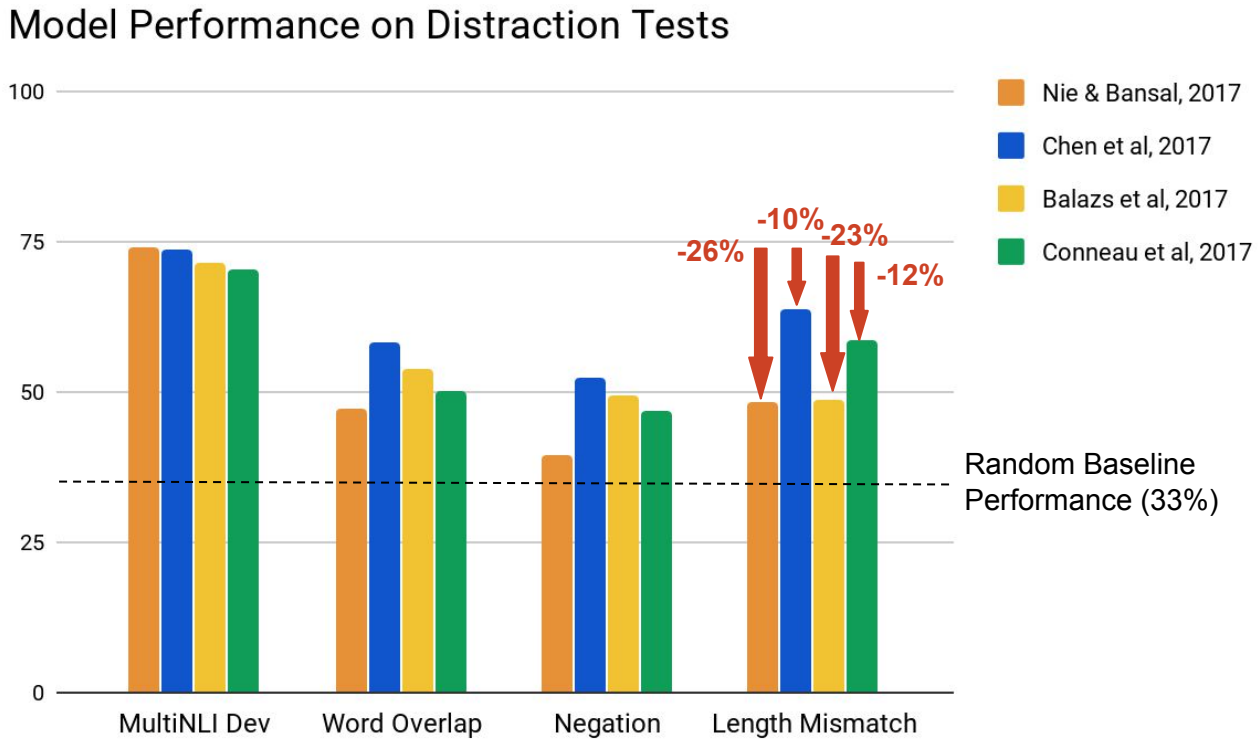
Results on Distraction Tests



Results on Distraction Tests



Results on Distraction Tests



Performance Analysis on Distraction Tests

High proportion of **false neutral** errors:

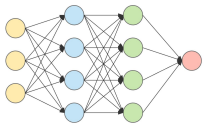
Account for **67.1%** of errors on word overlap and length mismatch tests (tests with ↓ lexical similarity), compared to **35.05%** of errors on MultiNLI Dev

Crowdworkers shown sentence:
Asked to construct entailed, neutral and contradictory hypotheses

P: Stimpy the cat believed he could fly
H: Stimpy thought he could fly

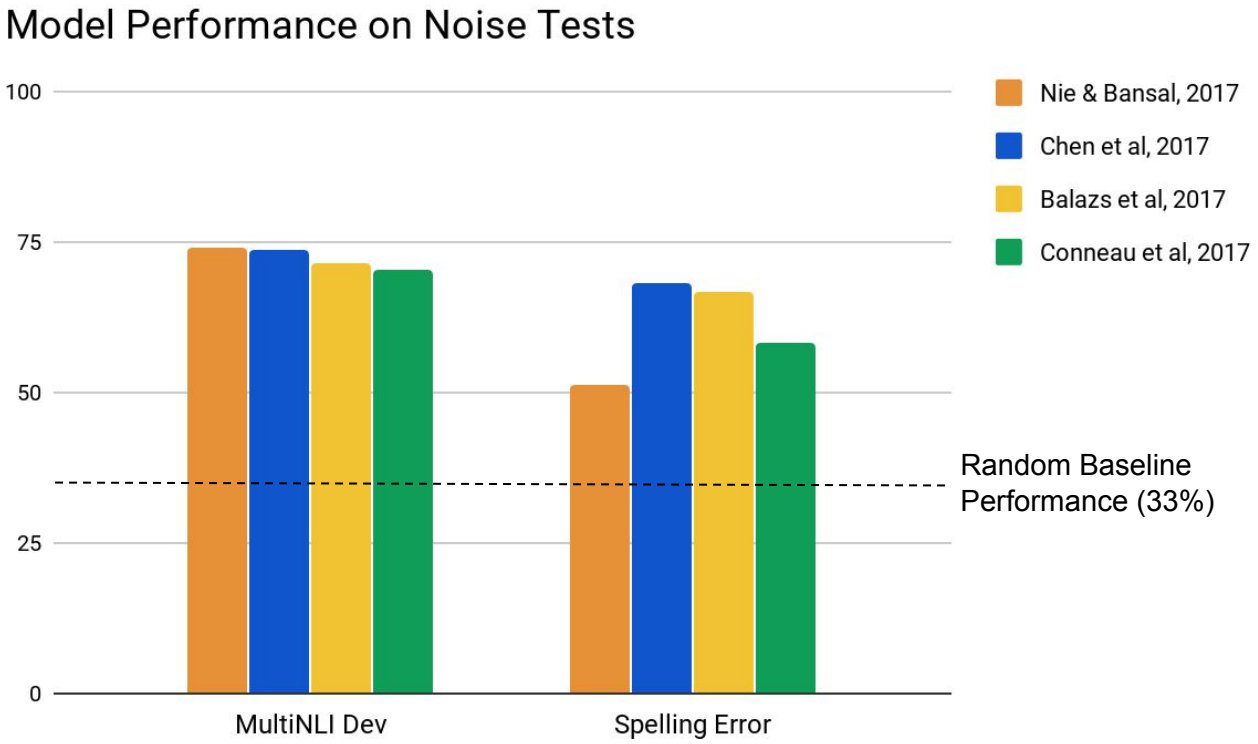
Construct hypotheses with inherent biases → biased datasets

(Bar-Haim et al, 2006; Parent et. al, 2010; Wang et. al, 2012; Yih et al, 2013; Gururangan et. al, 2018; Poliak et.al, 2018)

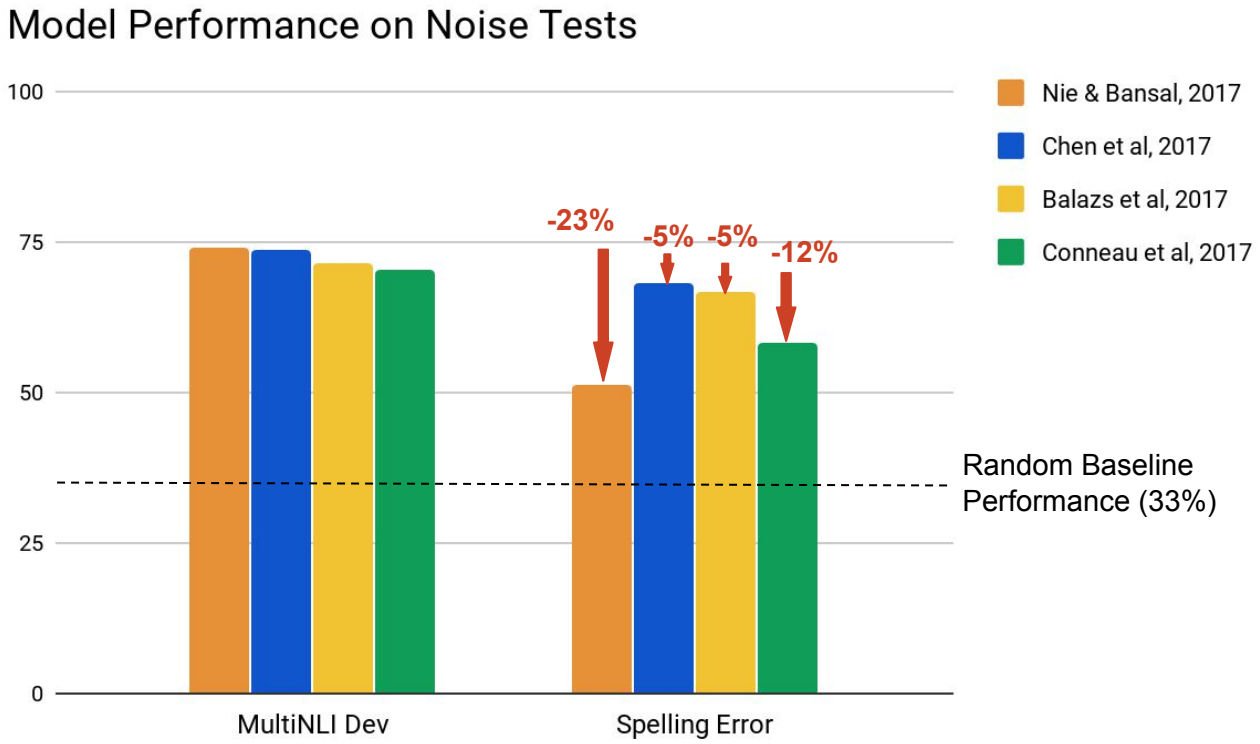


Neural models predict **entailment for high word overlap**, regardless of semantic meaning, and **neutral for low word overlap**, regardless of semantic meaning

Results on Noise Tests

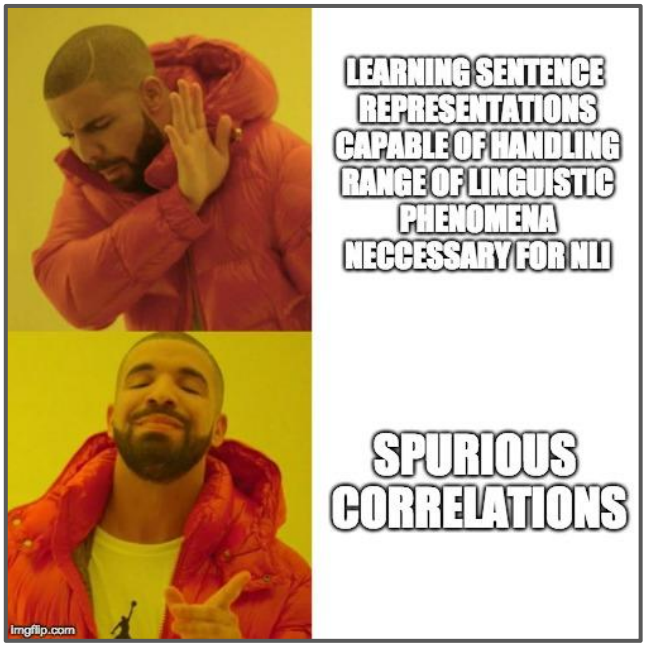


Results on Noise Tests



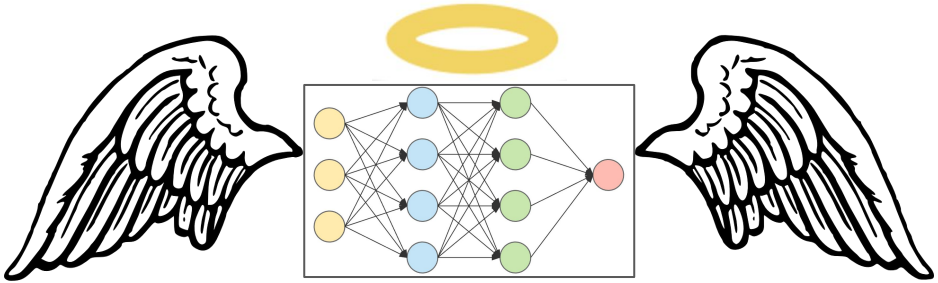
Conclusion

Neural networks will **take shortcuts**! Latch onto misleading correlations in datasets.



Conclusion

- Must ensure that our models are learning reasonable things. Good performance on dev set != good performance on task.



- **Avoid attributing good model performance to doing the kinds of reasoning we imagine the task needs.**
- Calls for diagnostic datasets - **Stress Tests**!

Conclusion

All stress tests/code/leaderboard now available: https://abhilasharavichander.github.io/NLI_StressTest/ .

Use the data to show models are actually learning!

THANK YOU!

Visit our website: https://abhilasharavichander.github.io/NLI_StressTest/

For questions, contact: anaik@cs.cmu.edu, aravicha@cs.cmu.edu