Experimental Designs: Adversarial Examples

Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

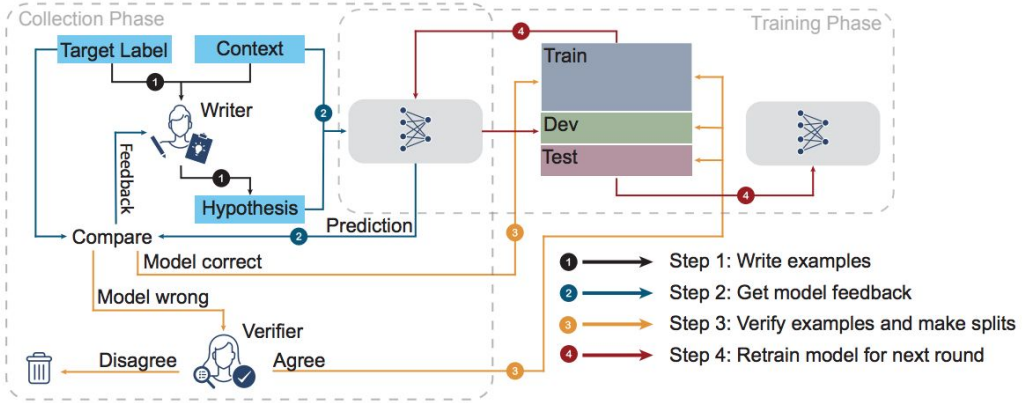
Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversarial NLI: Nie et al. (2019)



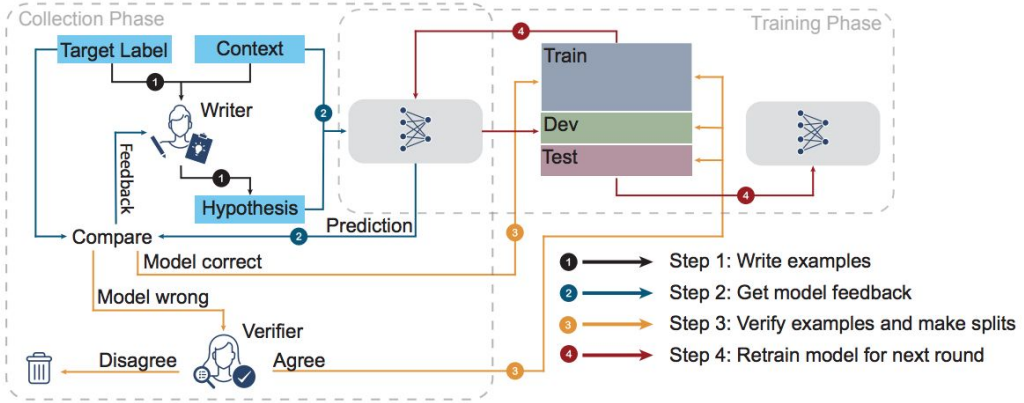
Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses
- Pros: Practical analysis of failures; can be used as training to improve model
- Cons: Sets age quickly; are model/data specific; “whack-a-mole” approach

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversarial NLI: Nie et al. (2019)



Construction Methods

- Sources of Data

- Example/Label Generation

Construction Methods

- Sources of Data
  - Sentences drawn from existing corpora
  - Sentences drawn from existing benchmark sets/test suites
  - Templates
  - Manual Generation
- Example/Label Generation

Construction Methods

- Sources of Data
  - Sentences drawn from existing corpora
  - Sentences drawn from existing benchmark sets/test suites
  - Templates
  - Manual Generation
- Example/Label Generation
  - Labels are given by-definition (e.g. if using templates or manual generation)
  - Automatically manipulate sentences and assume heuristic labels (+/- human filtering)
  - Purely automatic (e.g. adversarial)
  - Purely manual labeling (e.g. human generated examples)

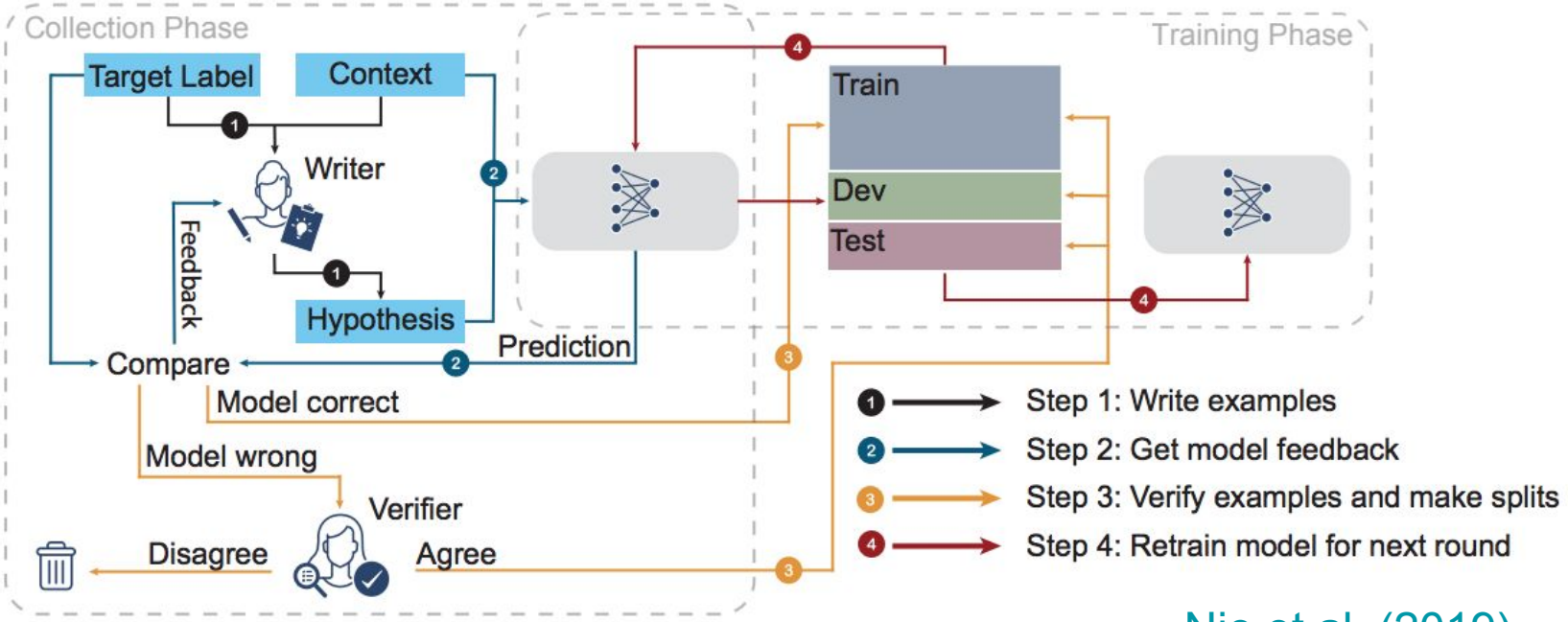
Construction Methods: Entirely Manual

Construction Methods: Entirely Manual

- Examples: Build-It-Break-It, Adversarial NLI

Construction Methods: Entirely Manual

- Examples: Build-It-Break-It, Adversarial NLI



Nie et al. (2019)

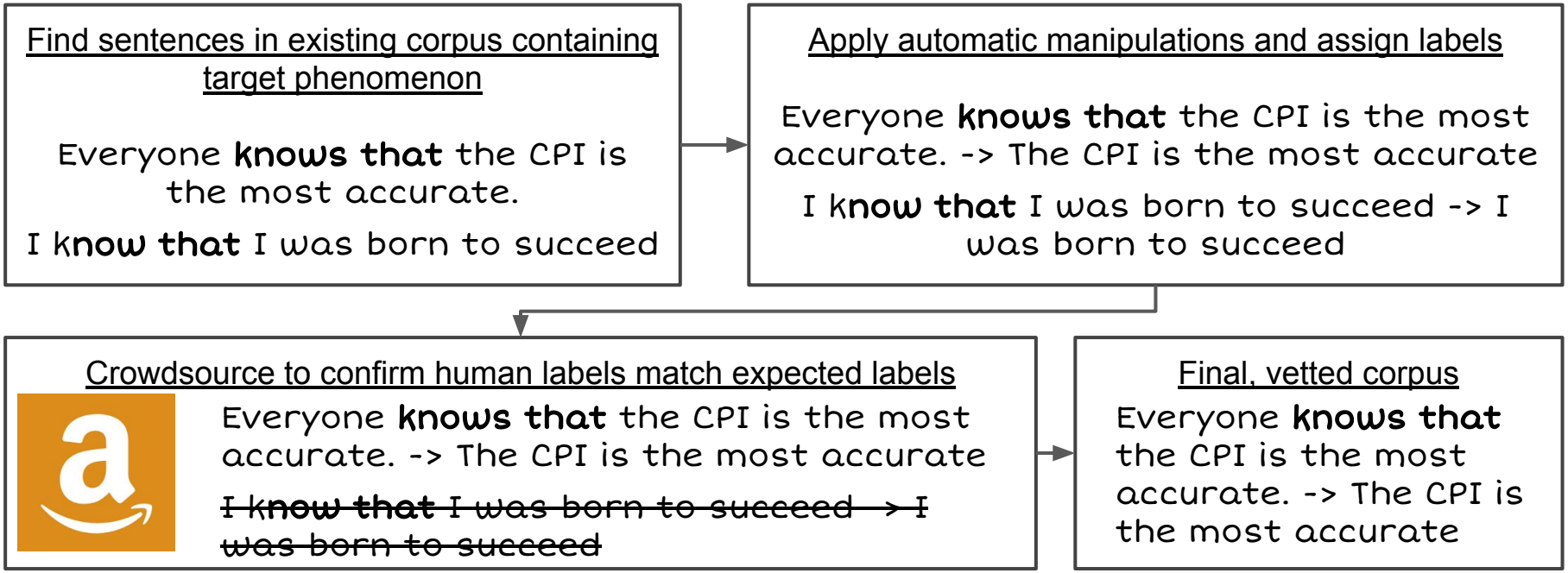
Construction Methods: Semi-Automatic

Construction Methods: Semi-Automatic

- Manipulate Existing Corpora, Filter with Crowdsourcing
  - Examples: Ross and Pavlick (2018), Kim et al. (2018), Poliak et al. (2018)

Construction Methods: Semi-Automatic

- Manipulate Existing Corpora, Filter with Crowdsourcing
  - Examples: Ross and Pavlick (2018), Kim et al. (2018), Poliak et al. (2018)



Construction Methods: Semi-Automatic

- Hand-crafted templates that produce known labels
  - Examples: Ettinger et al. (2018), McCoy et al. (2019)

Construction Methods: Semi-Automatic

- Hand-crafted templates that produce known labels
  - Examples: Ettinger et al. (2018), McCoy et al. (2019)

Subcase	Template	Example
Entailment: Conjunctions	The N_1 and the N_2 V the N_3 → The N_2 V the N_3	The actor and the professor mentioned the lawyer. → The professor mentioned the lawyer.
Non-entailment: NP/S	The N_1 V_1 the N_2 V_2 the N_3 ↛ The N_1 V_1 the N_2	The managers heard the secretary encouraged the author. ↛ The managers heard the secretary.

McCoy et al. (2019)

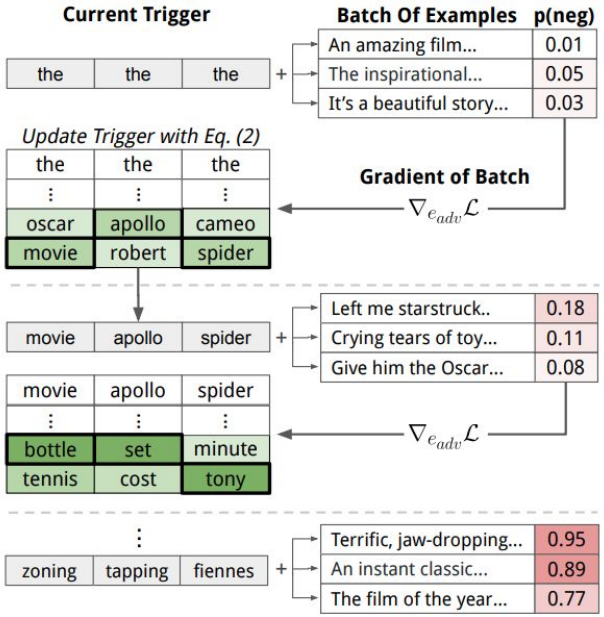
Construction Methods: Fully Automatic

Construction Methods: Fully Automatic

- Examples: Ebrahimi et al. (2018), Wallace et al. (2019)

Construction Methods: Fully Automatic

- Examples: Ebrahimi et al. (2018), Wallace et al. (2019)



Wallace et al. (2019)

Challenge Sets: Limitations

Challenge Sets: Limitations

- Availability
  - Limited coverage of tasks and languages
  - Need to expand beyond English and to more NLP tasks

Challenge Sets: Limitations

- Availability
  - Limited coverage of tasks and languages
  - Need to expand beyond English and to more NLP tasks
- Methodology
  - What does failure on a challenge set tell us?
  - Who is to blame, the model or its training data?
  - Lie et al. (2019) fine-tune a model on a few challenge set examples and re-evaluate
  - Rozen et al. (2019) diversify both the training and test data
  - Geiger et al. (2019) propose method for determining whether a generalization task is “fair”