Universität Konstanz


# NLI Data and Annotations: a Look behind the Scenes

AIKATERINI-LIDA (KATERINA) KALOULI

16. JULY 2020

NALOMA @ WESSLLI 2020

# Natural Language Inference (NLI)

*NLI is the task of determining whether a natural language sentence, also called hypothesis (H), can be inferred from another natural language sentence, also called the premise (P).*

*(McCartney, 2009)*

➔ 3-way classification: ENTAILMENT, CONTRADICTION or NEUTRAL

Given two sentences P and H, what is the relation between P and H?

- ENTAILMENT: if P is true, then H is also true
- CONTRADICTION: if P is true, H is highly unlikely to be true (de Marneffe et al, 2014)
- NEUTRAL: no relation can be established between P and H

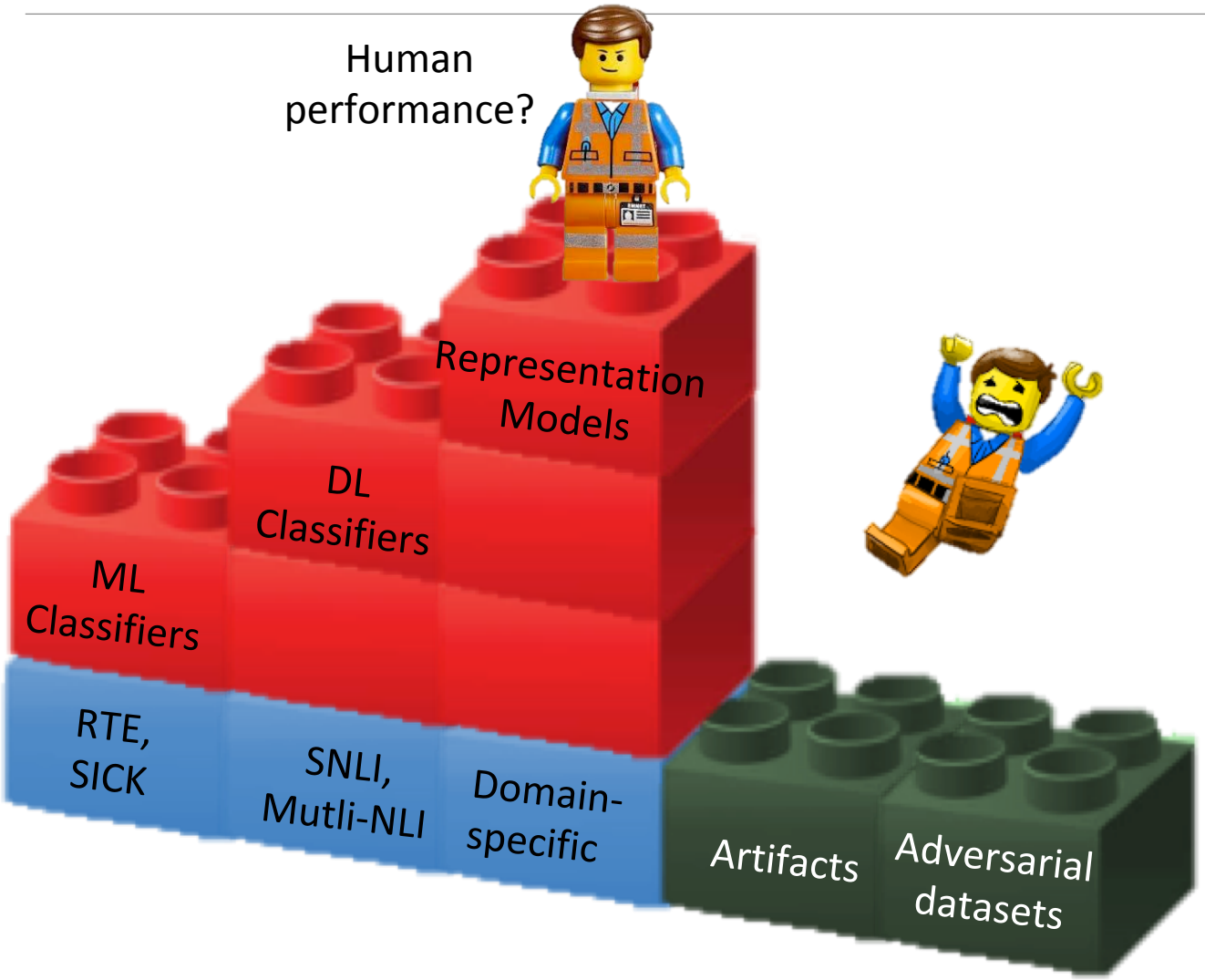
# NLI: where are we?

*NLI is a necessary metric for evaluating an NLU system since it forces a model to perform many distinct types of reasoning.*
*(Condoravdi et al, 2013)*

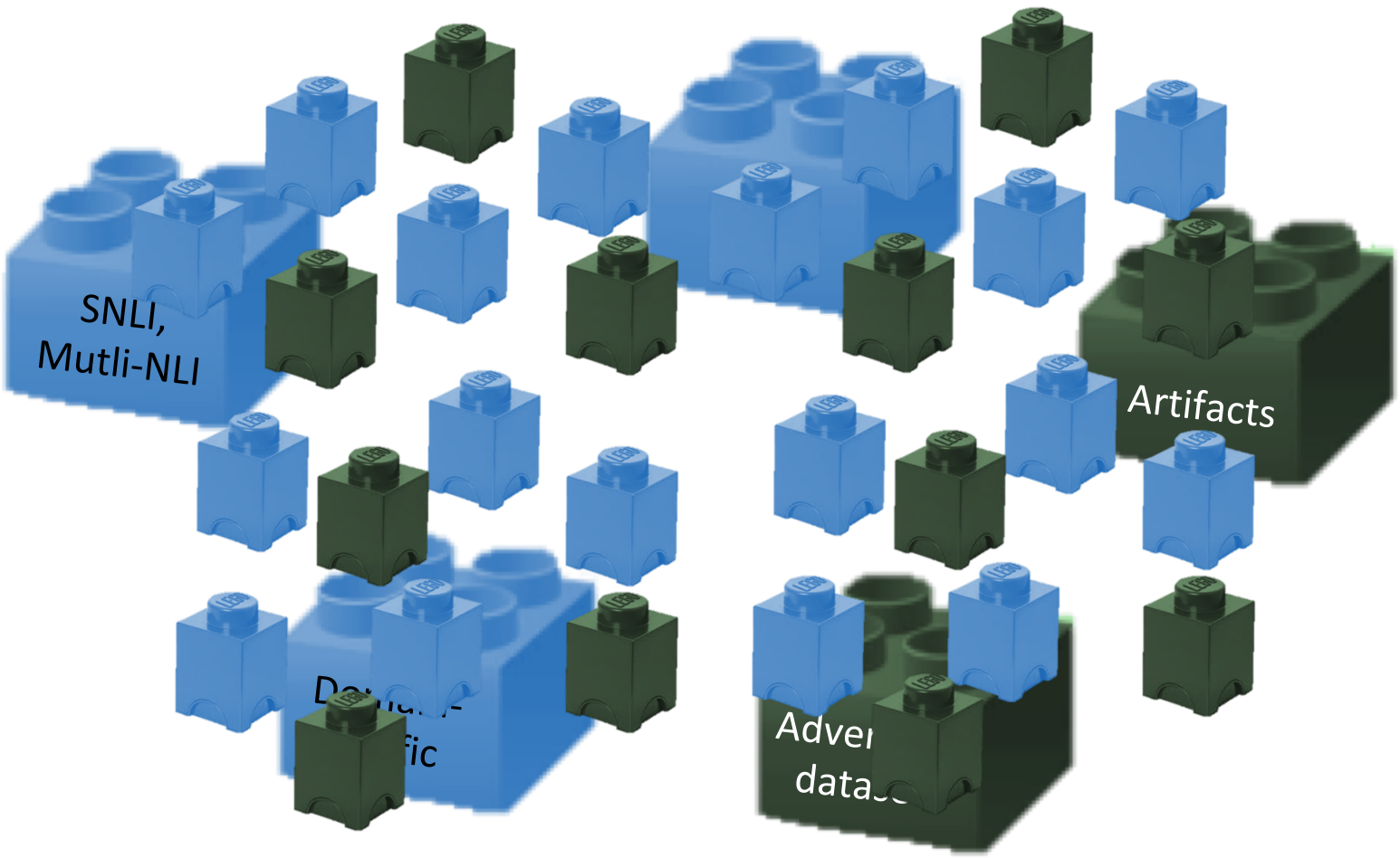
*… solving [NLI] perfectly entails human level understanding of language…*
*(Goldberg and Hirst, 2017)*

*… in order for a system to perform well at natural language inference, it needs to handle nearly the full complexity of natural language understanding…*
*(Nangia et al., 2017)*

# NLI: where are we?



# Take a closer look



# SICK (Marelli et al, 2014)

- 9840 English pairs annotated for degree of similarity and inference relation
- based on captions of pictures and a 3-step generation process
- aimed at every-day, common-sense sentences with no complex linguistic phenomena (e.g., no NEs, no MWEs, progressive tense, etc.)
- guidelines: no strict definitions, one example per relation

# SICK is sick (Kalouli et al, 2017a, 2017b, 2018)

- asymmetrical/illogical contradictions: one pair direction is contradiction, the other one isn't:

  P: A black and white dog is carrying a small stick on the green grass.
  H: A black and white dog is carrying a huge stick on the green grass.

- non-binding referents (coreference issues):

  P: An Asian woman in a crowd is not carrying a black bag.
  H: An Asian woman in a crowd is carrying a black bag.

- ungrammatical cases:

  The black and white dog isn't running and there is no person standing behind.

- nonsensical cases:

  A motorcycle is riding standing up on the seat of the vehicle.

# SICK is sick (Kalouli et al, 2017a, 2017b, 2018)

- non clear-cut definitions:

  P: There is no man on a bicycle riding on the beach.

  H: A person is riding a bicycle in the sand beside the ocean.

- alternative concepts:

  P: The lady is cracking an egg into a bowl.

  H: The lady is cracking an egg into a dish.

- plain errors:

  P: The blond girl is dancing behind the sound equipment.

  H: The blond girl is dancing in front of the sound equipment.

→ most of these issues also appear in SNLI and MultiNLI, especially the issue with coreference

# Goals & Methodology

- measure annotation quality when guidelines are improved and annotators provide explanations
- discover linguistic phenomena that are hard to annotate
- present aspects of NLI to be considered in future corpora

*Explaining Simple Natural Language Inference*
(Kalouli A-L., A. Buis, L. Real, M. Palmer, V. de Paiva, 2019)

- annotation experiment at the University of Colorado (CU)
- quantitative meta-experiment on the previous experiment
- corpus of investigation: SICK

# CU Experiment

- 12 graduate students of the University of Colorado Boulder
- improved, (supposedly) uncontroversial guidelines, building on previous literature, e.g., targeting coreference issues
- inference annotation, decision justification and intuitive "computational feasibility" (CF) score
- data: 224 randomly chosen SICK pairs, annotated in both directions
- Inter-Annotator Agreement (IAA): 73.2%, Cohen's k: 0.68 (substantial)
- 17% disagreement between our annotators and the original annotators

# Observations – Justifications I

- less-informative justifications, e.g., “sentences mean same thing”
- justifications describing the relation, e.g., “someone != no one”
- confusion about contradictory/neutral pairs:

  a) pairs marked as contradictory despite no coreference between them:

  P: Two sumo ringers are fighting.

  H: A man is riding a water toy in the water.

  → “subjects and activities are completely different”

  b) pairs marked as neutral despite obvious contradiction:

  P: A girl is getting a tattoo removed from her hand.

  H: A girl is getting a tattoo on her hand.

  → “could be getting both at the same time”

# Observations – Justifications II

- different agreement rates depending on the pair direction:

  P: A light brown dog is sprinting in the water.

  H: A light brown dog is running in the water.

  → A -> B: unanimously entailment but B -> A: 25% entailment

- "loose definitions" more prone to errors:

  P: A white dog is standing on a hill covered by grass.

  H: A dog is standing on the side of a mountain.

  → hill covered by grass = mountain?

- high CF scores in highly unambiguous pairs, e.g., one word difference

  → annotators give high CFs when they are themselves sure of the inference

# What if?

What if these observations are not merely random but can indeed be classified in phenomena and observed in other NLI data?

What if there is measurable correlation among the phenomena and the low IAA, so that these phenomena lead to statistically worse agreements?

Meta-Experiment

# Meta-Experiment

- 5 co-authors annotated the pairs for:

  - directionality: mark whether the current pair direction is easier, harder or equally hard to annotate → easier, harder or equal
  - coreference: does the pair contain events or entities hard to be assumed coreferent? → True or False
  - loose definitions: does the pair contain loose/subjective concepts? → True or False
  - atomicity: is each sentence atomic, i.e., does it contain one predicate-argument structure? → True or False for P and H
  - negation: does each sentence contain negation? → True or False for P and H
  - quantification phenomena: does each sentence contain a quantifier? → True or False for P and H

# Analysis

## Goal

check whether IAA and CF scores are statistically worse in pairs with such phenomena

## Method

- ✦ calculate IAA and CF score for each pair and each of the six meta-annotations
- ✦ IAA: Generalized Additive Mixed Models (GAMMs, Wood 2011, 2017) → check for main effects and interactions
- ✦ CF: logistic mixed-effects regression model → check for main effects and interactions

# Results IAA

- main effects of coreference, directionality, loose definitions and negation
- coreference, directionality and loose definitions confirm initial observations
- negation: counter-intuitive but only clear-cut textbook negations
- quantifiers: too few or easier for humans than for machines
- atomicity: no clear picture, more testing required

| Phenomenon | IAA | |
|---|---|---|
| | True | False |
| A_is_atomic | 72.06 | 79.4 |
| B_is_atomic | 72.6 | 76.81 |
| A_is_negated | **88.88** | **71.46** |
| B_is_negated | **90.47** | **71.27** |
| A_has_quant | 79.67 | 72.6 |
| B_has_quant | 80.48 | 72.5 |
| hard_coref | **62.45** | **77.27** |
| loose_def | **59.6** | **77.19** |

| Directionality | | | |
|---|---|---|---|
| **Measure** | **Easier** | **Harder** | **Equals** |
| IAA | **81.18** | **58.33** | **74.9** |

# Results CF score

- main effects of coreference and negation
- coreference: as hypothesized, an intuitively detectable factor that annotators "catch" by giving such pairs lower CF scores
- negation: due to clear-cut textbook negations pairs are more unambiguous and thus higher scores

| Phenomenon | CF score | |
|---|---|---|
| | True | False |
| A_is_atomic | 6.81 | 6.68 |
| B_is_atomic | 6.83 | 6.59 |
| A_is_negated | **7.66** | **6.68** |
| B_is_negated | **7.51** | **6.7** |
| A_has_quant | 7.03 | 6.76 |
| B_has_quant | 7.05 | 6.75 |
| hard_coref | **6.22** | **6.99** |
| loose_def | 6.2 | 6.95 |

| Directionality | | | |
|---|---|---|---|
| **Measure** | **Easier** | **Harder** | **Equals** |
| CF score | 6.57 | 6.58 | 6.88 |

# Discussion I

1. Improvement of the NLI process:
   - ✦ better guidelines → coreference, loose definitions
   - ✦ corpora and training and testing methods based on notion of *human* inference which has inherent variability, e.g.,
     - corpora without these phenomena ➔ systems should achieve perfect performance (like humans would)
     - corpora with all phenomena ➔ lower performance because not even humans agree
     - suitable training: easy pairs are more reliable ➔ higher training weights than harder pairs
     - suitable evaluation: performance measured separately on easier vs. harder pairs

# Discussion II

2. Enhancement of annotation tasks with justifications
   - ✦ reveal quality and weaknesses of guidelines ➔ IAA and Kappa not enough
   - ✦ reveal aspects of the task that should be taken into account
   - ✦ exploit justification for training: justification as extra rules, patterns, weights for more explainable models (cf. Camburu et al, 2018; Thorne et al, 2019; Kumar & Talukdar, 2020)

# Discussion III

3. Enhancement of the annotation task with a Difficulty Score

- CF score: estimated difficulty for the machine ➔ limited conclusions allowed
- Difficulty Score: real difficulty for the annotator
  - for training: easier pairs more reliable ➔ higher attention/weight
  - for testing: performance for easy vs. hard pairs: SOTA models might struggle with *annotation-easy* pairs (for humans) , which contain hard linguistic phenomena ➔ explainability, models' power
  - capture artifacts, e.g., pairs with the word *sleep* in H are always judged contradictory and easy, no matter what P is (due to the artifact that sleeping is used to contradict any other action)

# Conclusion & Future Work

- ✦ research focus not only on better models and larger or more complex datasets but also on quality of datasets
- ✦ reliable datasets ➔ reliable models
- ✦ plain annotation labels are not enough ➔ need for justifications for the decision
- ✦ IAA and Kappa not enough to measure quality
- ✦ certain aspects of NLI are measurable qualities found in other corpora as well ➔ adjust training and testing processes

→ apply experiment on other datasets

→ design corpus of *human inference*

# Thank you

https://ling.sprachwiss.uni-konstanz.de/pages/home/kalouli/
https://github.com/kkalouli/

Aikaterini-Lida Kalouli
Department of Linguistics, Univesity of Konstanz

Aikaterini-lida.kalouli@uni-konstanz.de