Implicature Discernment in Natural Language Inference

Group 7
Jesse Gioannini
Charlie Guo
Thomas Phan
Leroy Wang

LING 575C: Analyzing Neural Network Models
2/25/2020

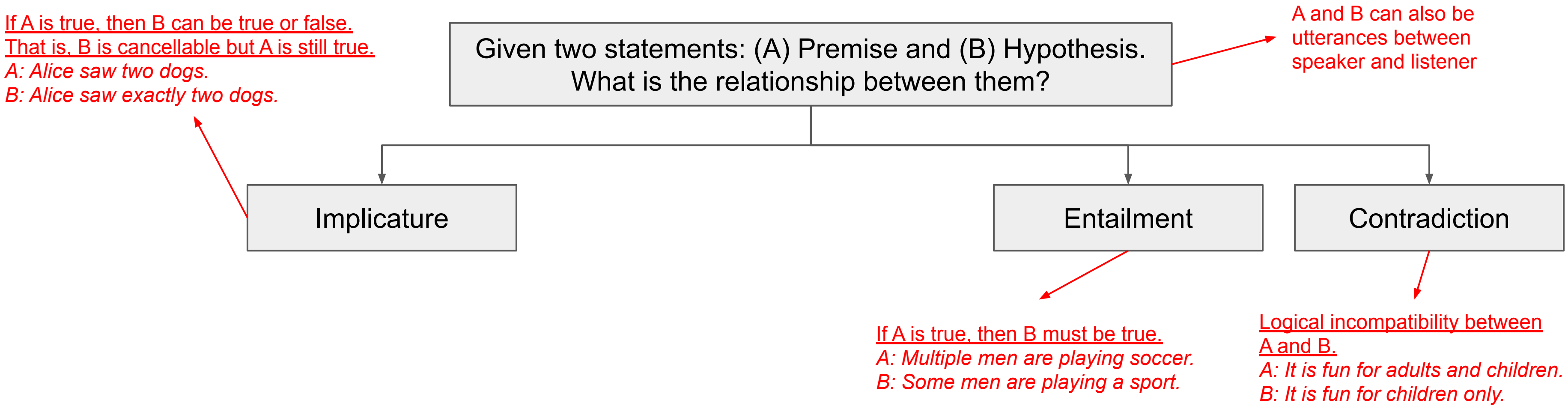
Overview

- Brief review of implicature, entailment, and contradiction
  - From the field of pragmatics
  - Studied by Grice in 1970s, not found in NN literature
- Two papers
  - “A Large Annotated Corpus for Learning Natural Language Inference”
  - “Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints”
- Our Project
  - Bringing implicatures to natural language inference

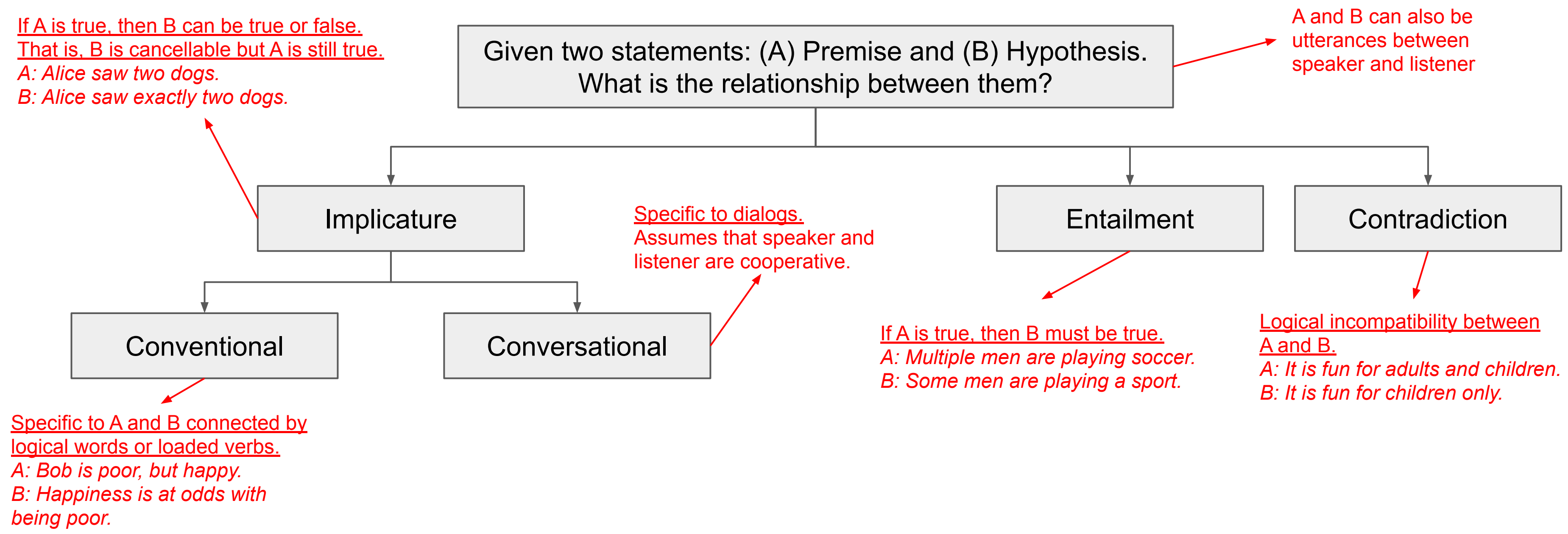
Brief review of implicature, entailment, and contradiction

Given two statements: (A) Premise and (B) Hypothesis.
What is the relationship between them?

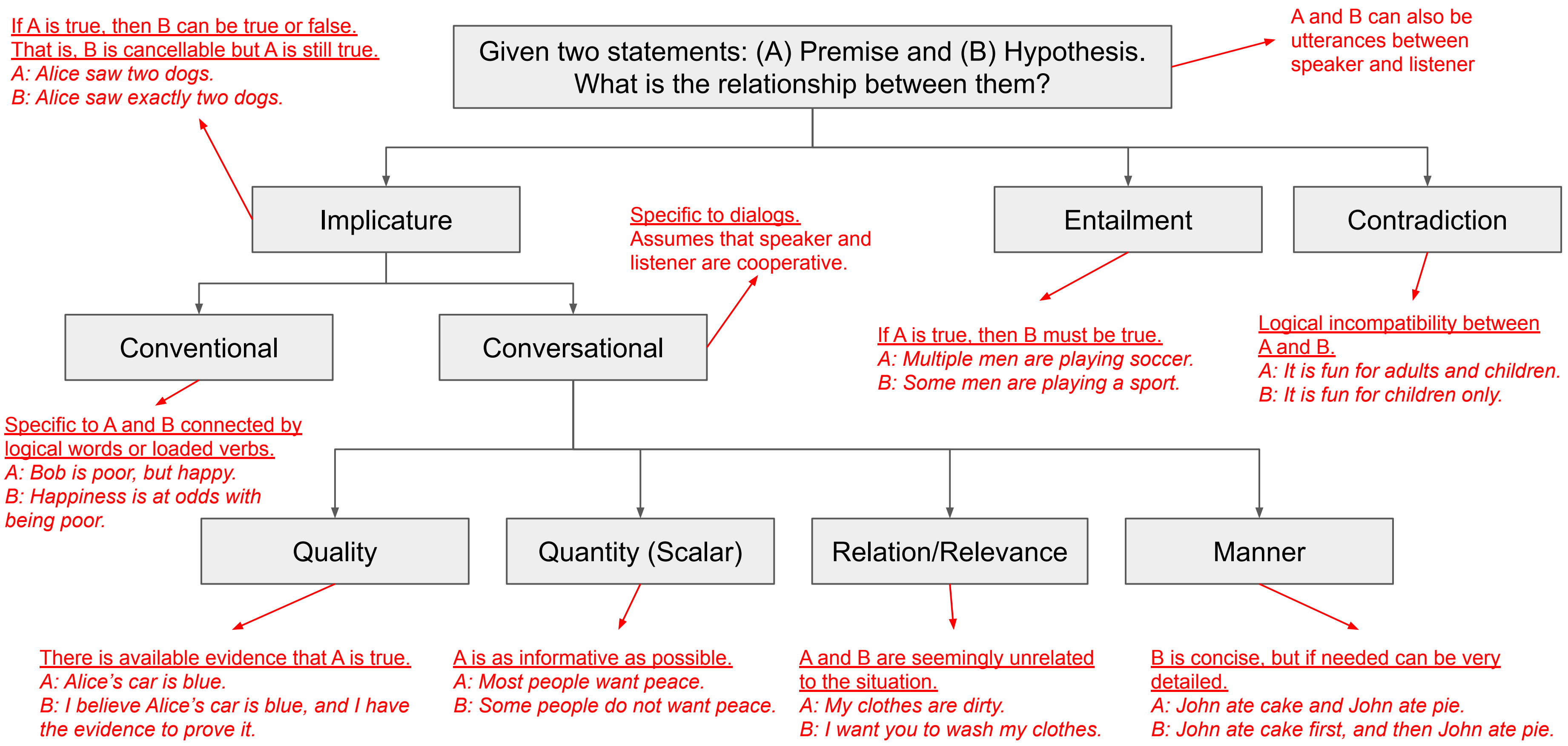
Brief review of implicature, entailment, and contradiction



Brief review of implicature, entailment, and contradiction

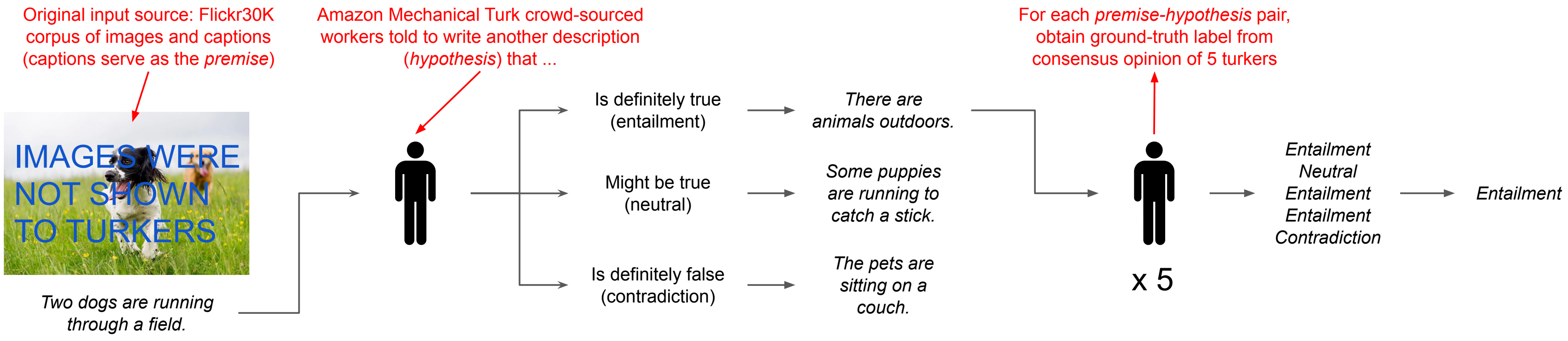


Brief review of implicature, entailment, and contradiction



Paper #1

- S. Bowman, G. Angeli, C. Potts, and C. Manning. "**A Large Annotated Corpus for Learning Natural Language Inference**," In Proceedings of EMNLP 2015.
- 1005 citations on Google Scholar
- Key ideas:
  - A novel dataset containing 570K labeled sentence pairs (previous sets were ~1k)
  - Hypothesis sentences were generated by humans (previous were partially synthetic)



Paper #1 (cont'd)

- Key results
  - **Availability** of Stanford Natural Language Inference (SNLI).

    https://nlp.stanford.edu/projects/snli/ (under Creative Commons Attribution-ShareAlike License)

  - **Validity** of SNLI

    Validated pairs: 56,951; Pairs w/ unanimous gold label: 58.3%; No gold label: 2%;

    Partitioned: train/test/dev; Parsed: via PCFG Parser 3.5.2; Large: two orders of magnitude larger than all other resources of its type.

  - **Utility** of SNLI

    Suitable for training parameter-rich models like neural networks.

Paper #1 (cont'd)

- Key results
  - **Utility** of SNLI (cont'd)

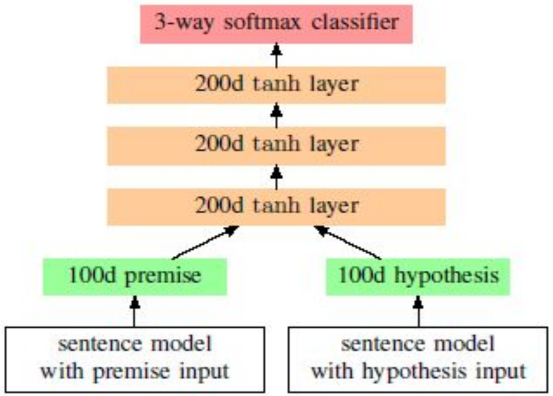


Figure 3: The neural network classification architecture: for each sentence embedding model evaluated in Tables 6 and 7, two identical copies of the model are run with the two sentences as input, and their outputs are used as the two 100d inputs shown here.

Sentence model	Train	Test
100d Sum of words	79.3	75.3
100d RNN	73.1	72.2
100d LSTM RNN	84.8	**77.6**

Table 6: Accuracy in 3-class classification on our training and test sets for each model.

Training sets	Train	Test
Our data only	42.0	46.7
SICK only	100.0	71.3
Our data and SICK (transfer)	99.9	**80.8**

Table 7: LSTM 3-class accuracy on the SICK train and test sets under three training regimes.

Paper #2

- L. Deng, J. Wiebe, Y. Choi. "**Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints**," In Proceedings of COLING 2014.
- 24 citations on Google Scholar
- Key ideas:
  - Infer implicit opinions over explicit sentiments and events that positively/negatively affecting entities. (GoodFor/BadFor event).

    "The reform would lower health care costs, which would be a tremendous positive change across the entire health-care system."

    Sentiment: *positive*; Event: "*reform lower costs*";

    Implicature: 1) *negative* to "cost"; 2) *positive* to "reform"

Paper #2 (cont'd)

- Key Ideas (cont'd)
  - Implicature rules: (s: sentiment; gf: good for; bf: bad for)

	s(gfbf)	gfbf	→	s(agent)	s(theme)		s(gfbf)	gfbf	→	s(agent)	s(theme)
1	positive	gf	→	positive	positive	3	positive	bf	→	positive	negative
2	negative	gf	→	negative	negative	4	negative	bf	→	negative	positive

Table 1: Rule Schema 1 & Rule Schema 3 (Deng and Wiebe, 2014)

e.g. "The reform would curb skyrocketing costs in the long run."

s(gfbf) = *positive*; Agent: "reform"; Theme: "costs"; gfbf: *bf* ("reform" *bf* "cost");
s("costs") = *negative*
Rule 3 applies: s("reform") = *positive*;

Paper #2 (cont’d)

- Key Ideas (cont’d)
  - Goal: Optimize a global function of all possible labels (pos/neg) on all agent/theme.
  - Method: Integer Linear Programming Framework.
  - Not a neural network model. (not really helpful to our project, but shows how accurately modelling implicatures’ behavior improves sentiment analysis; we think accurate detection of implicatures would improve the epistemic validity of automated reasoning on premises extracted from text).

:

Paper #2 (cont'd)

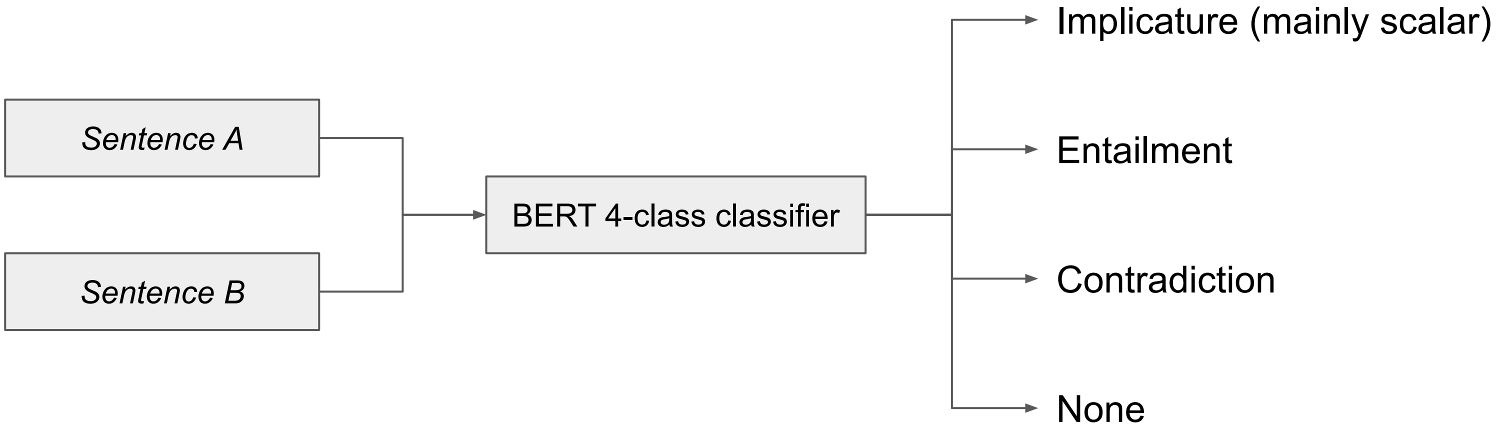
- Key results
  - Data "Affordable Care Act" corpus of DCW: 134 online editorials and blogs.
  - Results Comparison (on stats of Precision; Recall; F-measure)
  - Conclusion
    - The method improves over local sentiment recognition by almost 20 points in F-measure and over all sentiment baselines by over 10 points in F-measure.

		correct span subset			whole set, strict eval			whole set, relaxed eval		
		P	R	F	P	R	F	P	R	F
1	ILP	0.6421	0.6421	0.6421	0.4401	0.4401	0.4401	0.5939	0.5939	0.5939
2	Local	0.6409	0.3332	0.4384	0.4956	0.2891	0.3652	0.5983	0.3490	0.4408
3	ILP+coref	0.6945	0.6945	0.6945	0.4660	0.4660	0.4660	0.6471	0.6471	0.6471
4	Local+coref	0.6575	0.3631	0.4678	0.5025	0.3103	0.3836	0.6210	0.3834	0.4741
5	Majority	0.5792	0.5792	0.5792	0.3862	0.3862	0.3862	0.5462	0.5462	0.5462

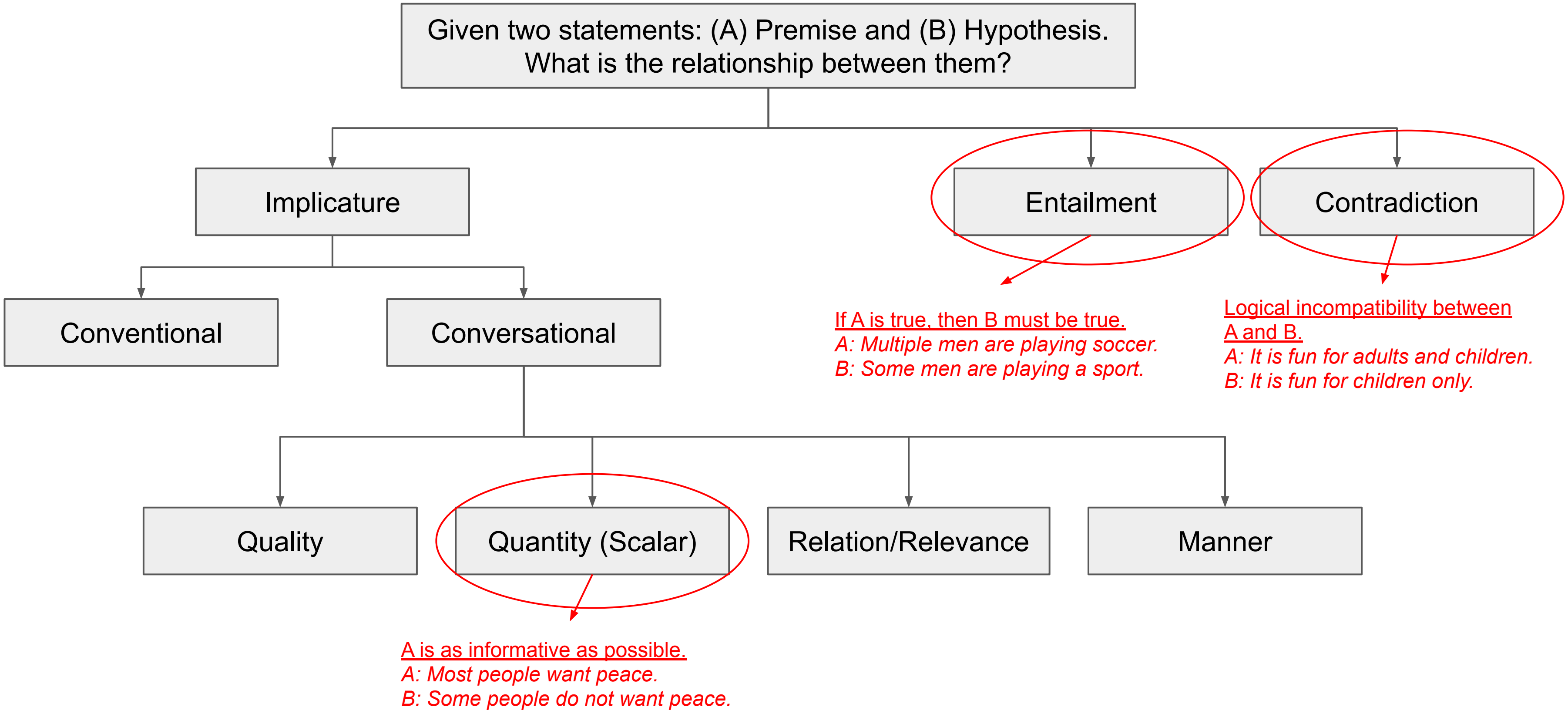
Table 3: Performances of sentiment detection

Our project

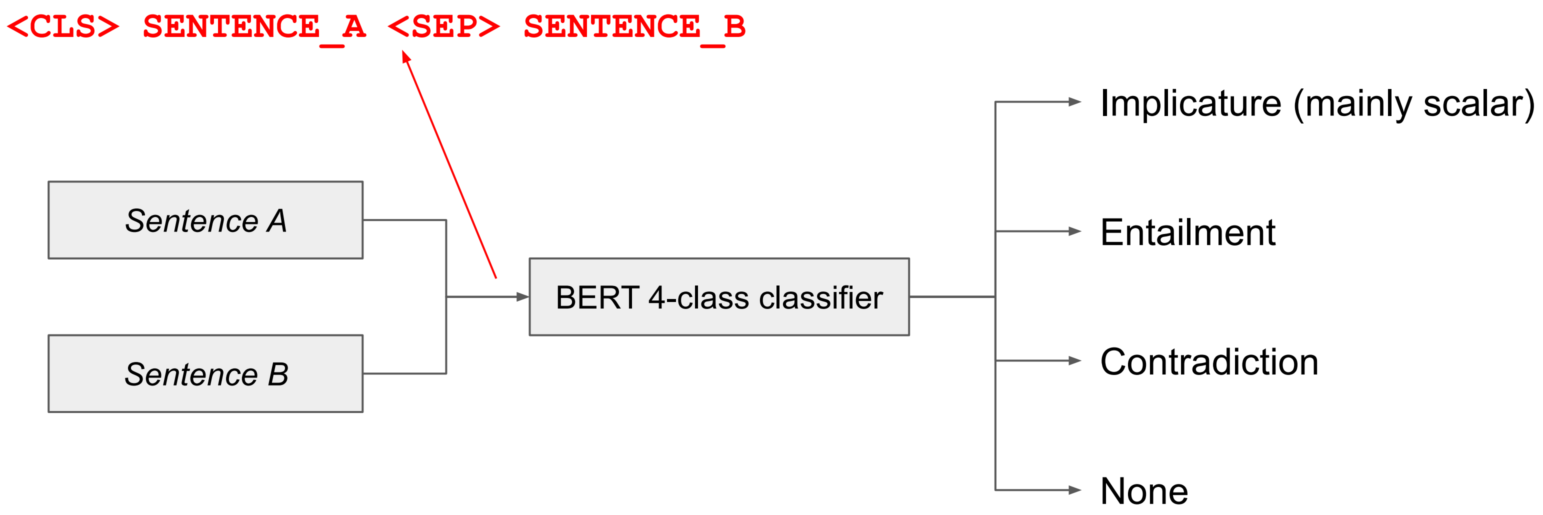
- Can the BERT contextual neural network language model distinguish between subtle inferential relationships (viz. implicature vs. entailment)?
- To the best of our knowledge, no other work has investigated this problem.



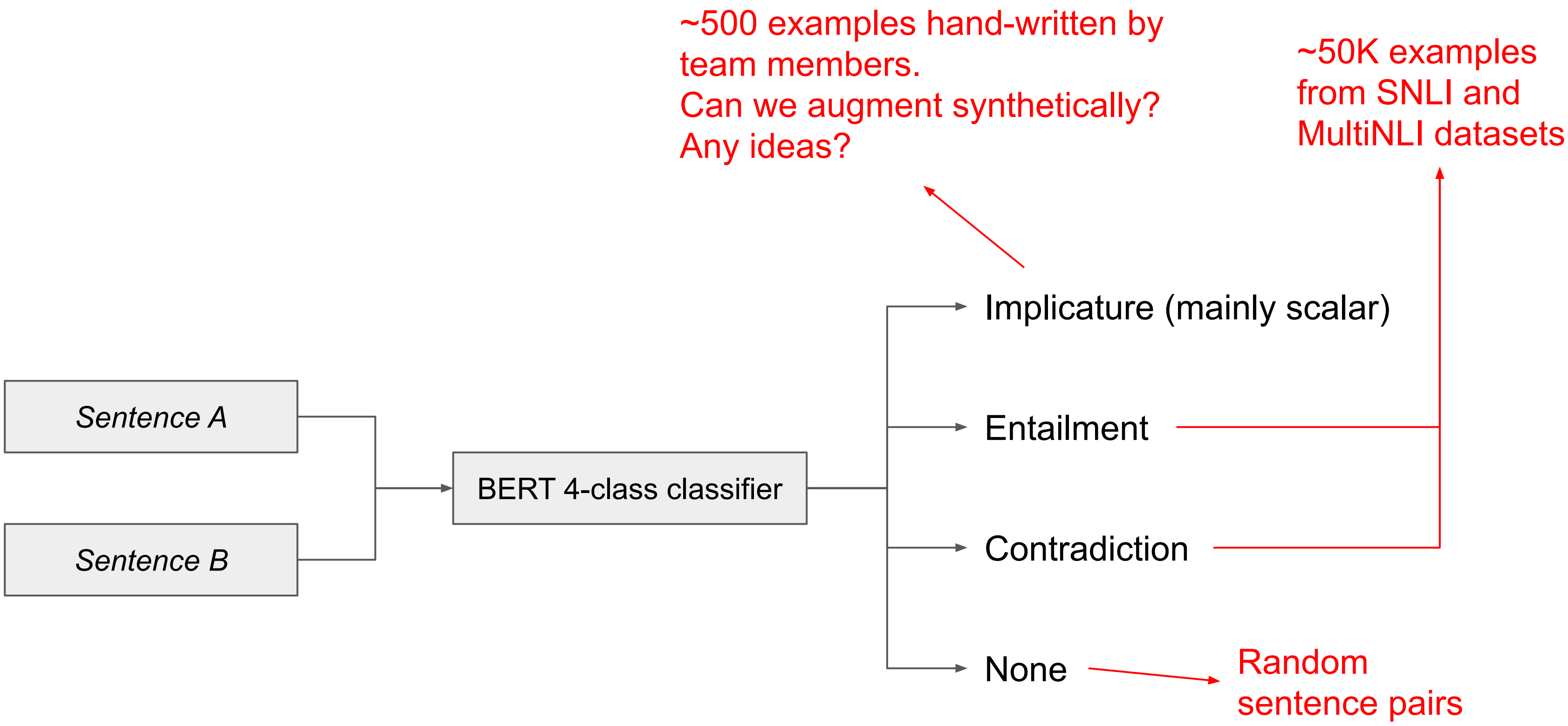
Brief review of implicature, entailment, and contradiction



Our project: Using BERT



Our project: Data availability



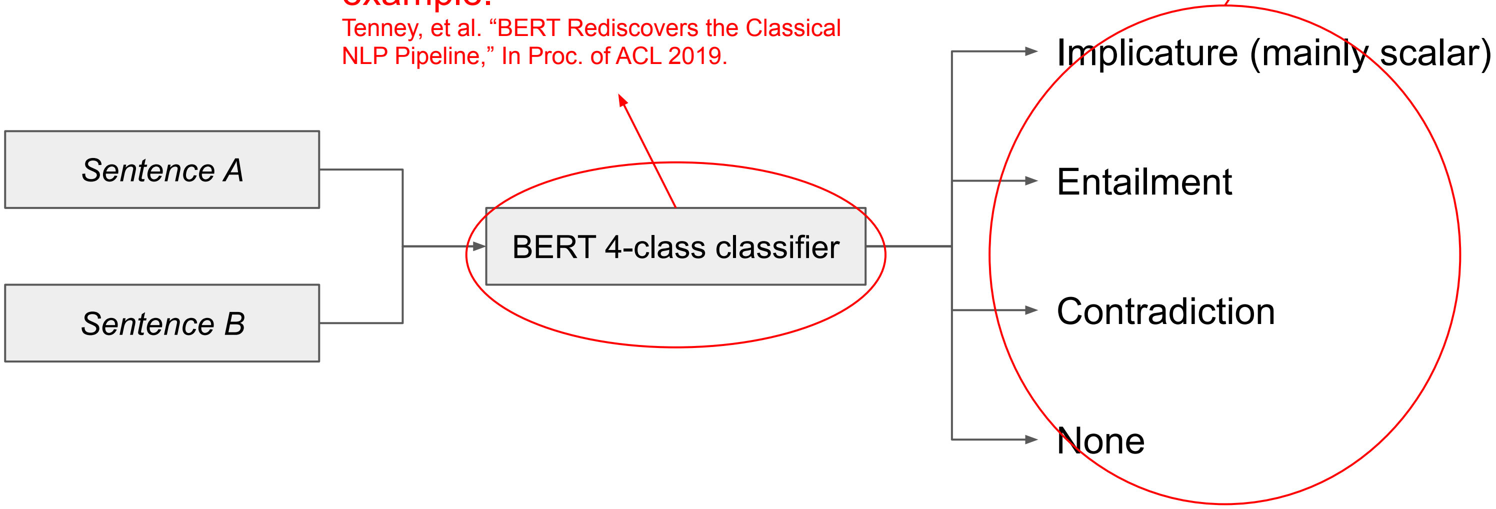
Our project: Experiments

<u>2. Stretch goal:</u>
At what layer does BERT gain the most knowledge?

Compute "expected layer" at which model correctly labels example.

Tenney, et al. "BERT Rediscovers the Classical NLP Pipeline," In Proc. of ACL 2019.

<u>1. Primary goal:</u>
What is the prediction F1 score or accuracy of untuned vs. tuned BERT?



Thank you