

UNIVERSITY OF CAMBRIDGE

A large annotated corpus for learning Natural Language Inference

Bowman, Angeli, Potts & Manning · EMNLP 2015

Jaehyuk (Jay) Choi

MPhil in Advanced Computer Science

L101 · Machine Learning for Language Processing
5 November 2026

NLI (Natural Language Inference)?

- NLI is the task of determining the semantic relationship between two sentences:

  1. **Premise** (P): A given statement assumed to be true.

  2. **Hypothesis** (H): A statement whose relationship to the premise is evaluated.

- NLI tests whether a model can understand semantic relationships beyond simple word matching

Examples of NLI

- Classify the relationship between two sentences

A man is playing soccer (P)

"A person is playing a sport."	ENTAILMENT
"The man is playing soccer with his friends."	NEUTRAL
"The man is sleeping."	CONTRADICTION

Examples of NLI

- Classify the relationship between two sentences

A man is playing soccer (P)

"A person is playing a sport."	ENTAILMENT
"The man is playing soccer with his friends."	NEUTRAL
"The man is sleeping."	CONTRADICTION

Motivation

01 | Lack of Large-Scale Data

Existing NLI datasets were too small to train data-intensive neural models.

02 | Artificial Data

Automatically generated sentences introduced spurious patterns.

03 | Label Ambiguity

Ambiguous event and entity coreference undermined annotation reliability.

A need emerged for a Large-scale, Naturalistic, and Reliably Annotated NLI benchmark

Related Works

DATASET / RESOURCE	SIZE	CONSTRUCTION	LIMITATION FOR NLI
RTE *(Dagan et al., 2006)*	**<1,000**	Human-annotated	Too small for data-intensive representation learning
SICK *(Marelli et al., 2014)*	**~4,500 train**	Partially automated construction	Artificial construction can create spurious patterns
Denotation Graph *(Young et al., 2014)*	**Millions**	Automatically generated and labeled	Noisy labels; better for supplementary training
Focused entailment graphs *(Levy et al., 2014)*	**Large**	Semi-automatic annotation of (subject, verb, object) triples	Not full-sentence inference
PPDB 2.0 *(Pavlick et al., 2015)*	**Large**	Automatically generated word & phrase-level entailment annotations	Not full-sentence inference
SNLI *(Bowman et al., 2015)*	**570,152**	Human-written hypotheses + separate validation	

How Was SNLI Constructed?

One premise, three human-written hypotheses, then independent label validation.

1. Premise from Flickr30k

Source: Image Caption

“Two dogs are running through a field.”

Workers saw the caption.
Not the Photo!

2. Write 3 hypotheses

One for each inference relation

ENTAILMENT

Animals are outdoors.

NEUTRAL

The dogs are chasing a stick.

CONTRADICTION

The dogs are sitting on a couch

3. Human validation

For a subset of pairs (~10%):

Original label + 4 extra votes

E E E E N

≥3 agreeing votes → gold label

No majority → “–” (excluded)

570,152 sentence pairs | **~2,500** crowdworkers | **98%** majority consensus*

** Among additionally validated pairs.*

What Models Did They Evaluate?

Three scalable model families; the EOP comparison uses a different label setup.

01 **EOP systems (Existing NLI Baselines)**
2-class NLI
Edit distance · classifier · + lexical resources

WHY EVALUATE?
Benchmark traditional NLI methods on SNLI

02 **Lexicalised* classifier**
3-class NLI
Hand-crafted overlap and cross-word features

WHY EVALUATE?
How far can hand-crafted lexical features go?

03 **Sentence encoders**
3-class NLI
Sum of words · RNN · LSTM + classifier

WHY EVALUATE?
Can neural models learn useful sentence representations?

Lexicalised: Uses specific words and word combinations as features for classification.

EOP Baselines

Initial 3-class	EOP's 2-class
Entailment	**Entailment**
Neutral	**Non-entailment**
Contradiction	

EOP baselines (2-class test accuracy)

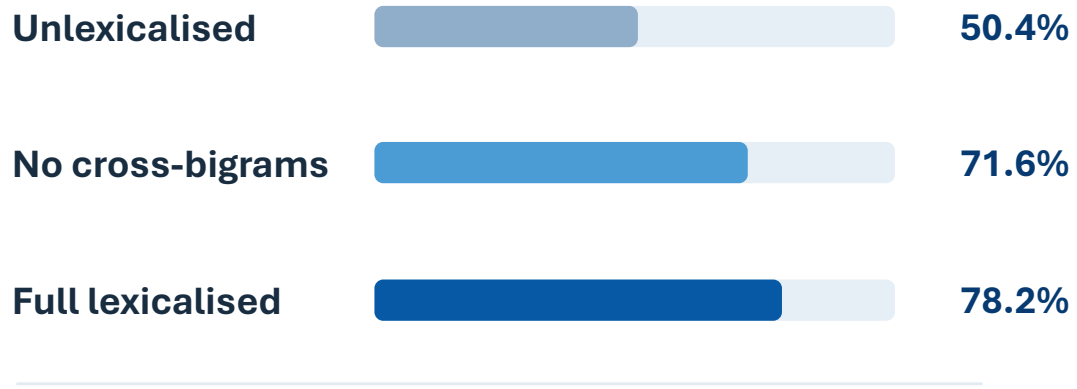
System	SNLI	SICK	RTE-3
Edit distance	71.9	65.4	61.9
Feature-based classifier	72.2	71.4	61.5
Classifier + Lexical resources	**75.0**	**78.8**	**63.6**

Most-frequent-class baselines: SNLI 66% · SICK 71%

Why Is a Simple Lexical Model So Strong?

Adding cross-sentence lexical features has a striking effect on SNLI accuracy.

SNLI test accuracy (3 classes)



+27.8 percentage points vs. unlexicalised

What information is added?

1. **Unlexicalised**
   BLEU, length difference, word overlap
2. **Lexicalised**
   Hypothesis unigrams and bigrams
3. **Cross-sentence**
   Cross-unigrams and cross-bigrams

Caveat: 99.7% train vs. 78.2% test indicates substantial overfitting.

Can Neural Models Match Feature-Based Classifiers?

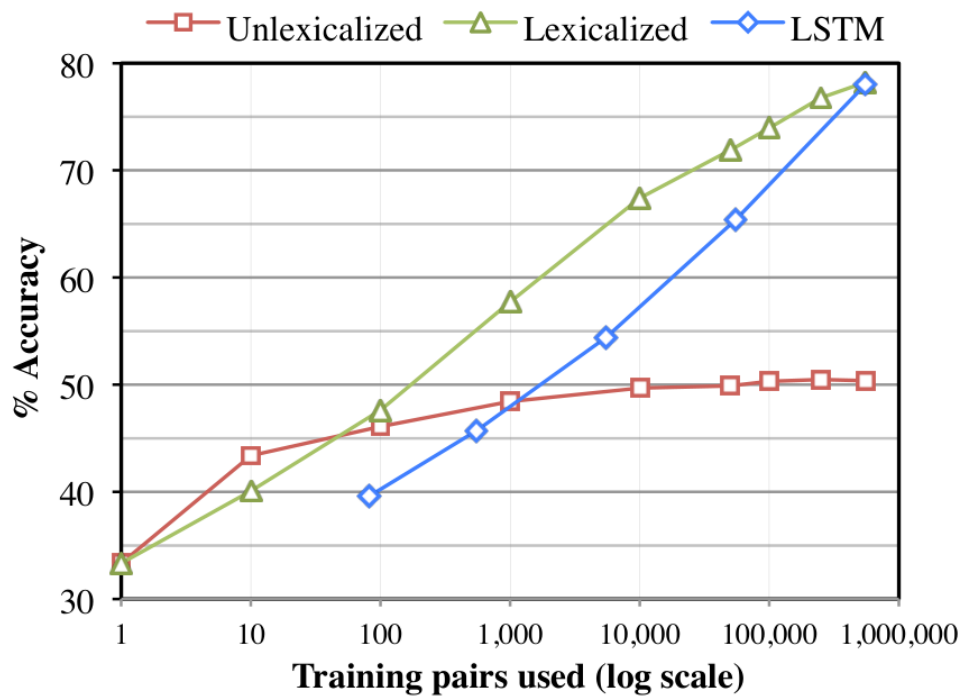
Accuracy (%)

Model	Train	Test	Interpretation
Unlexicalized (BLEU, length, overlap)	49.4	50.4	Overlap alone is weak
Lexicalized, no cross-bigrams	93.1	71.6	Word identity matters
Lexicalized + cross-bigrams	99.7	78.2	Best; memorizes training data
Sum of word vectors (GloVe)	79.3	75.3	No word order
Plain RNN	73.1	72.2	Below sum-of-words
LSTM	84.8	77.6	Nearly tied with lexicalized model; smaller train–test gap

Key result: **LSTM 77.6% ≈ feature-rich classifier 78.2% (same 3-class SNLI test).**

Does More Training Data Improve NLI?

Figure 4: the learning curves distinguish a capacity bottleneck from a data bottleneck.



What to notice

Unlexicalized

Plateaus around 50%

Lexicalized

Benefits strongly from scale

LSTM

Steeper late-stage increase

Original Figure 4 (Bowman et al., 2015) · x-axis is log-scaled; LSTM unavailable at very small sizes.

What Do the Models Still Get Wrong?

Three common failure modes reveal the limitations of high-performing NLI models.

Commonsense

PREMISE

A driver leaps from a burning car.

HYPOTHESIS

The driver is escaping danger.

Predicted: neutral

World knowledge and situational inference

Argument structure

PREMISE

A man is being bitten by a dog.

HYPOTHESIS

A man bit a dog.

Predicted: entailment

Who did what to whom is reversed

Lexical association

PREMISE

A girl throws sand toward the ocean.

HYPOTHESIS

A girl can’t stand the ocean.

Predicted: entailment

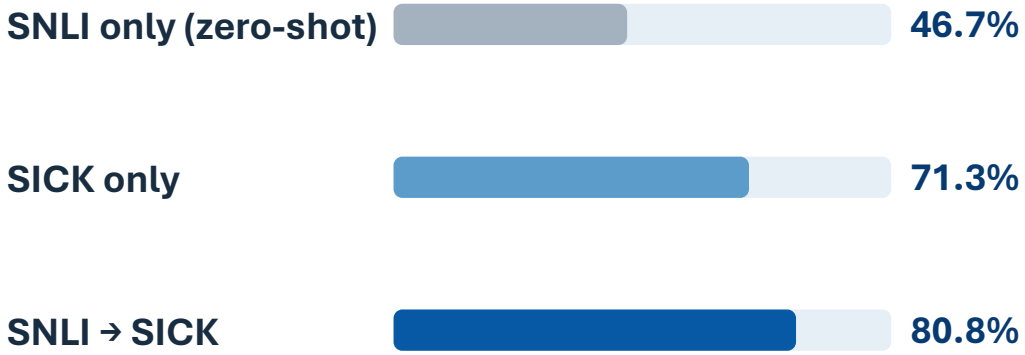
Lexical cue overrides idiomatic meaning

A strong NLI score can coexist with shallow lexical shortcuts and fragile compositional reasoning.

Does Knowledge Transfer Beyond SNLI?

Pretraining helps on SICK only after fine-tuning on the target task.

SICK test accuracy (3-class NLI)



Transfer protocol

Train LSTM on SNLI

↓ transfer learned parameters

Fine-tune on SICK

+9.5 pp

vs. training on SICK only

Limit: Zero-shot transfer is weak; RTE-3 adaptation was not competitive.

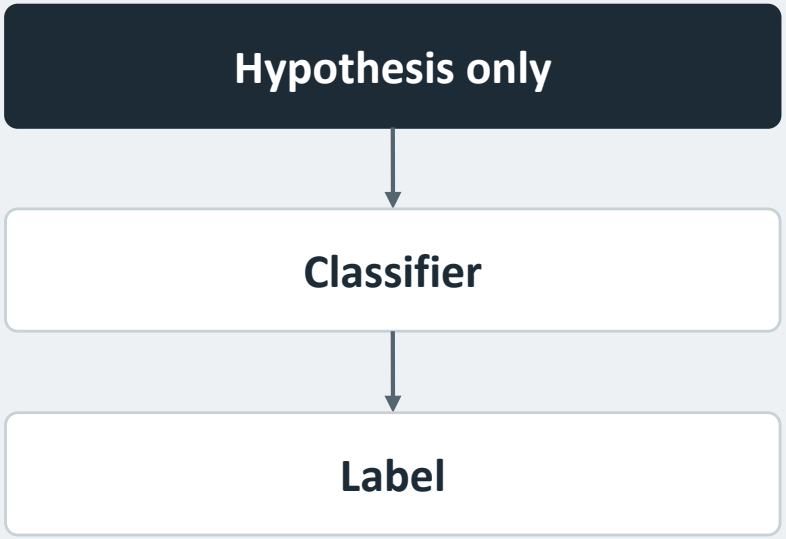
Contributions

Contribution	What it means
A large NLI dataset	570k human-written pairs: the first NLI data big enough to train neural models
A new way to collect it	Both sentences describe the **same** scene Hence, the “same event” ambiguity mostly disappears
A benchmark with baselines	Feature-based, neural and transfer results to beat Beaten within months (Rocktäschel et al., 2016), as a benchmark should be
A lasting format	Premise, hypothesis, three labels became standard Reused by MultiNLI and for NLI pre-training (InferSent, 2017)

SNLI's lasting contribution was establishing a large-scale NLI benchmark
Rapid improvements over its original baselines demonstrated its value in enabling and measuring progress

Predicting NLI Labels Without the Premise

Premise: not shown to the model



~67%

SNLI test accuracy from the hypothesis alone *(Gururangan et al., 2018)*
Chance is about 33%.

Cue in the hypothesis	Predicted label
Negation (nobody, not)	**Contradiction**
Vague or added detail	**Neutral**
More general words (animal, outdoors)	**Entailment**

If labels can be predicted without the premise, what is the benchmark measuring?

Does SNLI Generalise Beyond Image Captions?

	What SNLI covers well	Limitations
Sentences	Short descriptions of visible scenes	Longer text and discourse
Inference	Concrete entity and action relations	More complex reasoning about Time, Possibility, and Quantity
Transfer	Improves SICK after fine-tuning (71.3 → 80.8%)	Poor Zero-shot generalisation: 46.7% on SICK

Strong performance on SNLI does not guarantee general semantic understanding

Other limitations and future work

Limitation	Why it matters	Future work
Independent Sentence Encoding	No direct word-level interaction between premise and hypothesis	Cross-sentence attention mechanisms (83.5% - Tim Rocktäschel et al., 2016)
Single gold label	Human judgments can differ, especially for ambiguous examples (Neutral κ 0.60)	Evaluate against human label distributions (Nie et al., 2020)
Reproducibility	Some training settings (L2, dropout, classifier type, run variance, pay rates) not reported	Report hyperparameters, several runs, the full protocol

Future work should improve cross-sentence reasoning, account for human disagreement, and strengthen experimental reproducibility.

Overall Evaluation

What the paper claims	Holds?	Why
SNLI is a large, human-written NLI corpus	**Yes**	570k pairs, about 100× RTE and SICK
Reliable annotations	**Mostly**	98% majority agreement But only 10% independently validated
Neural models achieve competitive NLI performance	**Yes**	LSTM: 77.6% vs. lexical classifier: 78.2%
Learned representations generalise	**Partly**	Improves SICK after fine-tuning, but weak direct transfer

SNLI is a valuable benchmark
But strong performance does not necessarily imply genuine semantic understanding

References

1. Bowman et al. (2015). *A Large Annotated Corpus for Learning Natural Language Inference*. EMNLP 2015.
2. Dagan et al. (2006). *The PASCAL Recognising Textual Entailment Challenge*. Springer, LNCS 3944.
3. Marelli et al. (2014). *A SICK Cure for the Evaluation of Compositional Distributional Semantic Models.* LREC 2014.
4. Young et al. (2014). *From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions.* TACL 2014.
5. Magnini et al. (2014). *The Excitement Open Platform for Textual Inferences.* ACL 2014.
6. Rocktäschel et al. (2016). *Reasoning about Entailment with Neural Attention.* ICLR 2016.
7. Gururangan et al. (2018). *Annotation Artifacts in Natural Language Inference Data.* NAACL 2018.
8. Nie et al. (2020). *What Can We Learn from Collective Human Opinions on Natural Language Inference Data?* EMNLP 2020.

Thank you

Data Quality: What Was Actually Validated?

The evaluation sets were independently checked; most training labels were not.

Corpus / split (Table 2)

Train	**550,152**
Development	**10,000**
Test	**10,000**
Full corpus	**570,152**

Validation (Table 3)

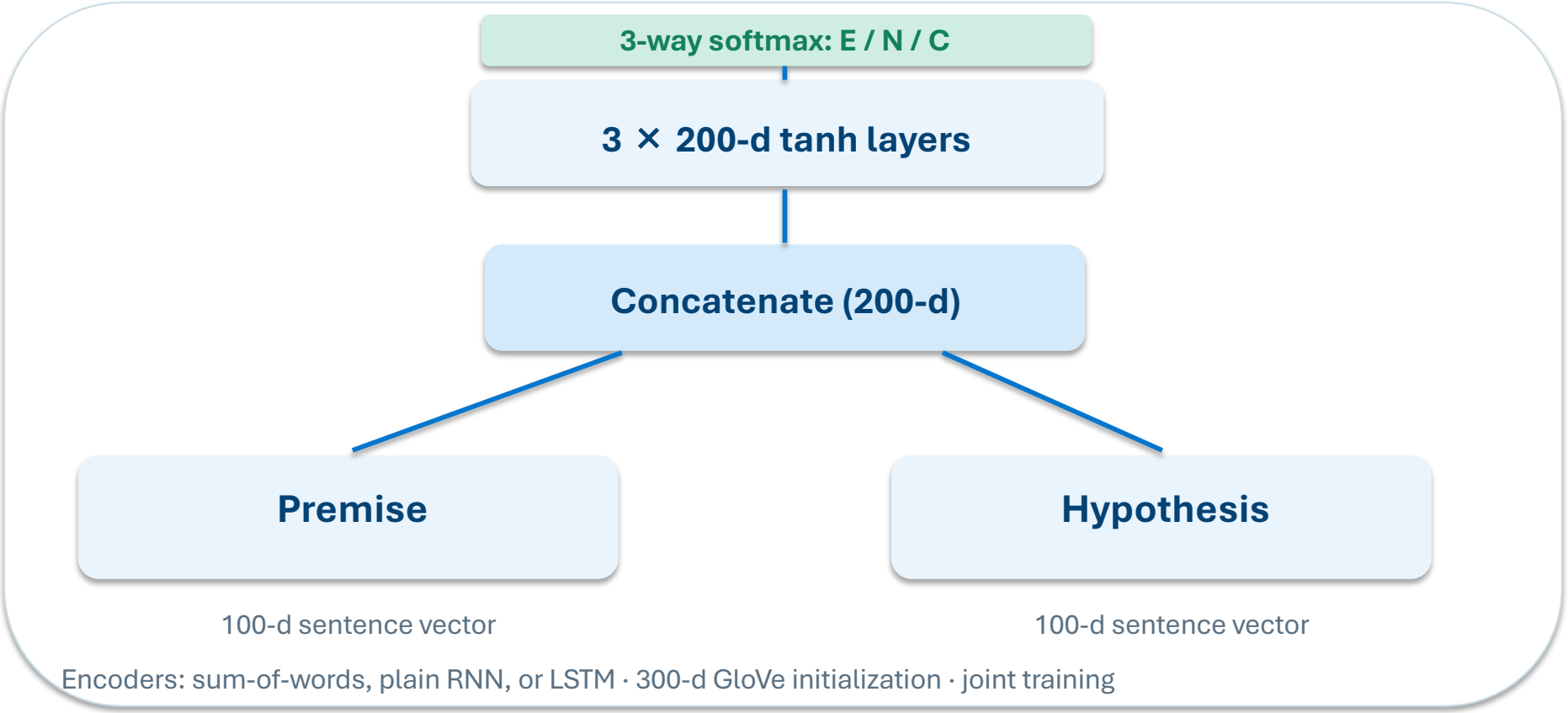
Validated pairs	**56,951**
Majority consensus	**98.0%**
Unanimous agreement	**58.3%**
Author = gold label	**91.2%**

Fleiss κ: 0.70 overall | contradiction .77 · entailment .72 · neutral .60

Table 2–3 · All dev/test examples validated; splits are separated by original Flickr caption.

What Does the Neural Model Actually Learn?

Figure 3: two independently encoded sentences feed a shared three-way classifier.



Adapted from Bowman et al. (2015), Figure 3. Encoder does not see the other sentence.