Analyzing Compositionality-Sensitivity of NLI Models

Yixin Nie*, Yicheng Wang*, Mohit Bansal

UNC NLP

Natural Language Inference

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

(Premise, Hypothesis) → Label { Entailment, Contradiction, Neutral }

Importance of NLI

The concepts of **entailment** and **contradiction** are central to all aspects of natural language meaning.

Importance of NLI

The concepts of **entailment** and **contradiction** are central to all aspects of natural language meaning.

Building computation systems that can recognize these relationships is essential to many NLP tasks such as **question answering** and **summarization**.

Difficulty of NLI

At a high level, NLI is a complicated task with many components.

Difficulty of NLI

At a high level, NLI is a complicated task with many components.

Intuitively, success in natural language inference needs a high degree of **sentence-level understanding**.

Difficulty of NLI

At a high level, NLI is a complicated task with many components.

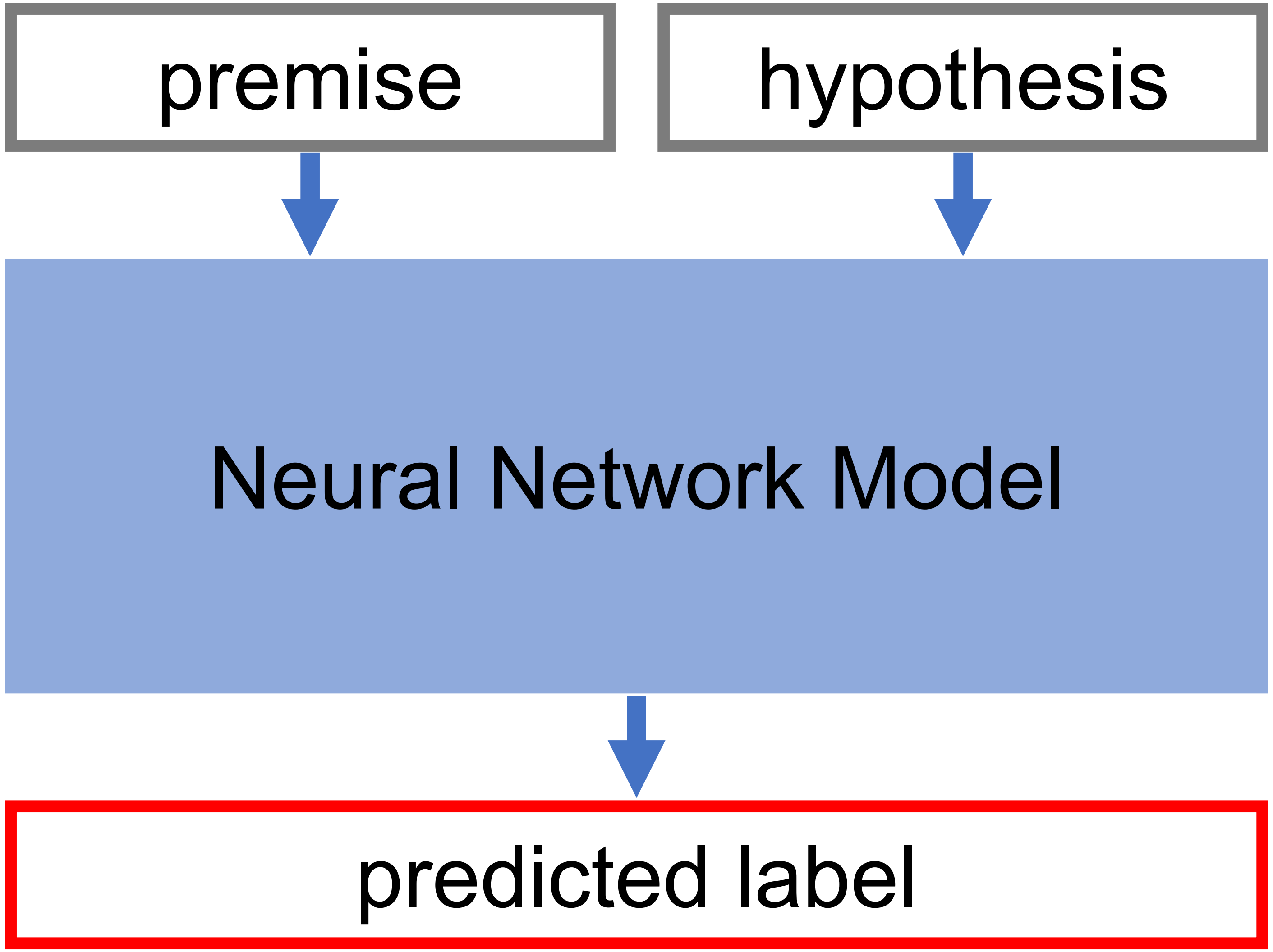
Intuitively, success in natural language inference needs a high degree of **sentence-level understanding**.

Sentence-level understanding requires a model to capture both lexical and **compositional** semantics.

Datasets

- Stanford Natural Language Inference (SNLI)
  570k pairs (image caption genre)
- Multi-Genre Natural Language Inference (MNLI)
  433k pairs (multiple genres e.g. news, fiction)

Models



Trained on provided training set.

Current Model and Motivation

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

Despite their **high** performance, it is unclear if models employ semantic understanding or are simply performing **shallow** pattern matching.

Current Model and Motivation

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

Despite their **high** performance, it is unclear if models employ semantic understanding or are simply performing **shallow** pattern matching.

Counterintuitive model designs indicate an **over-focus** on **lexical** information, which is **different** from human reasoning.

Current Model and Motivation

UNC NLP

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

Despite their **high** performance, it is unclear if models employ semantic understanding or are simply performing **shallow** pattern matching.

Counterintuitive model designs indicate an **over-focus** on **lexical** information, which is **different** from human reasoning.

This motivates our analytic study of models' **compositionality-sensitivity**.

Current Model and Motivation

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

Model	**SNLI**	**Type**	**Representation**
RSE	86.47	Enc	Sequential
G-TLSTM	85.04	Enc	Recursive (latent)
DAM	85.88	CoAtt	Bag-of-Words
ESIM	88.17	CoAtt	Sequential
S-TLSTM	88.10	CoAtt	Recursive (syntax)
DIIN	88.10	CoAtt	Sequential
DR-BiLSTM	88.28	CoAtt	Sequential

Analysis experiments

- Adversarial Evaluation
  - Expose models' compositional-unawareness and over reliance on lexical feature.
- Compositional-removal analysis
  - Reveal the limitation of current evaluation.
- Compositional-sensitivity testing
  - Provide a tool to explicitly analysis models' compositionality-sensitivity.

Semantic-based Adversaries

Goal:

To show that models are **over-reliant** on word-level information and have limited ability to process compositional structures.

Semantic-based Adversaries

Goal:

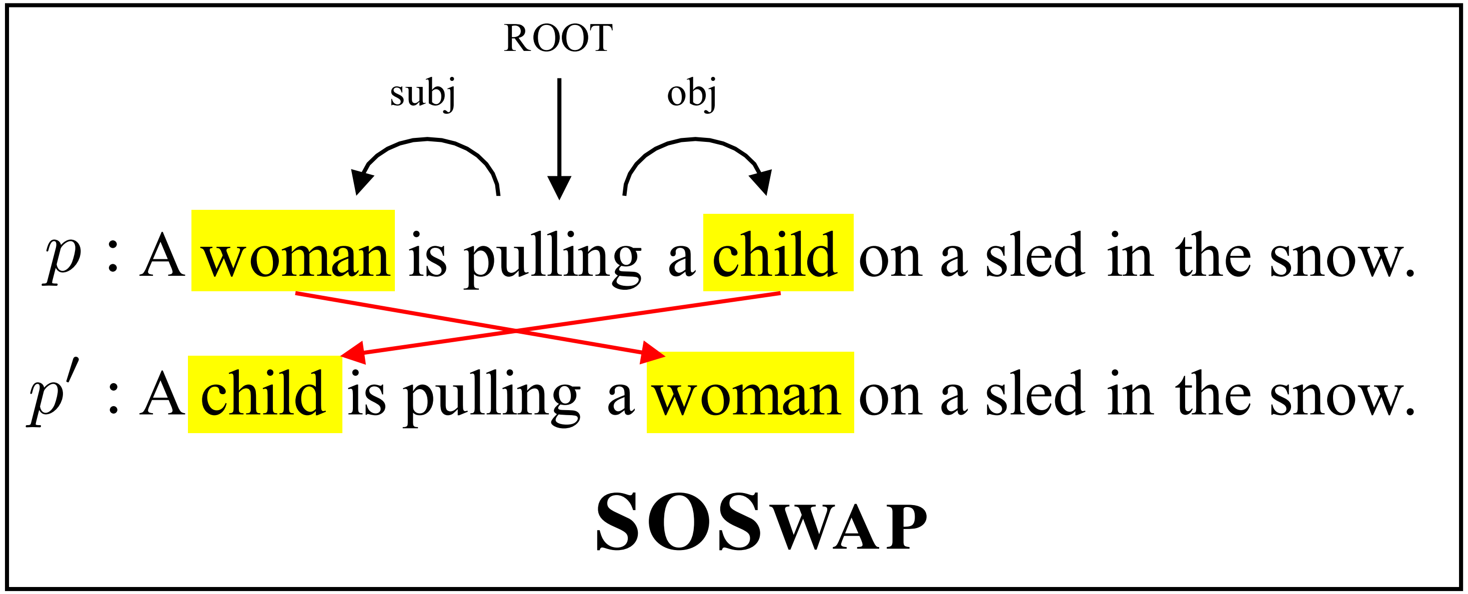
To show that models are **over-reliant** on word-level information and have limited ability to process compositional structures.

Method:

Created adversaries whose logical relations cannot be extracted from lexical information **alone**.

SubObjSwap:

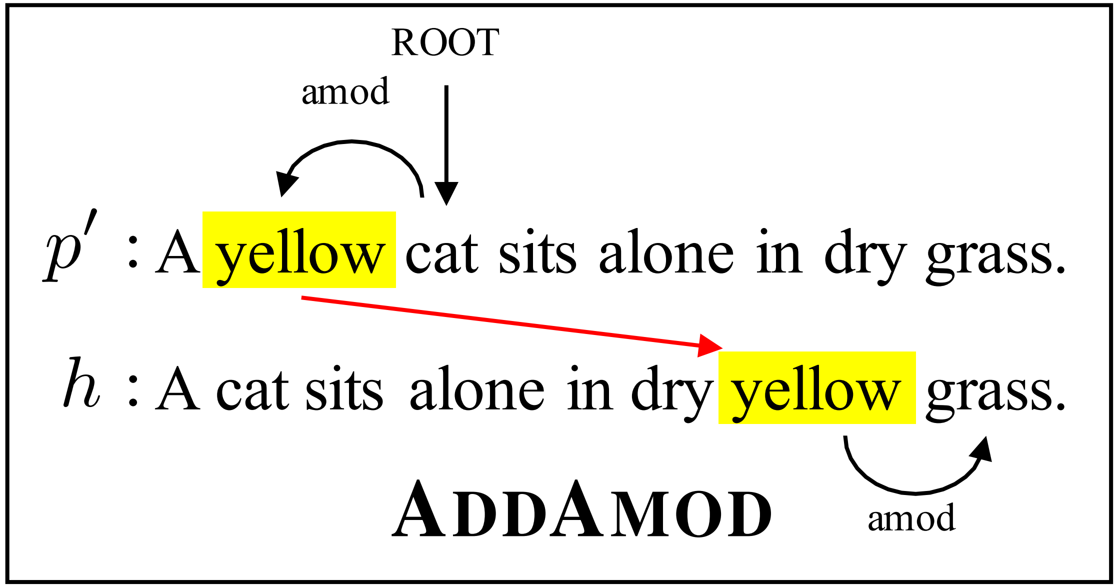
- Take a premise with a subject-verb-object structure;
- Create the hypothesis by swapping the subject and object.



Semantic-based Adversaries

AddAmod:

- Take a premise that has at least two different noun entities;
- Pick an adjective modifier;
- Create the premise by adding the modifier to one of the nouns, and the hypothesis by adding it to the other.



Adversarial Evaluation Results

Model	SNLI dev	SOSWAP			ADDAMOD		
		E	C	N	E	C	N
RSE	86.5	**92.5**	2.1	5.5	**95.2**	0.2	4.6
G-TLSTM	85.9	**97.2**	1.2	1.5	**95.9**	1.2	2.9
DAM	85.0	**99.7**	0.3	0.0	**99.9**	0.0	0.1
ESIM	88.2	**96.4**	2.1	1.5	**85.6**	9.6	4.8
S-TLSTM	88.1	**92.1**	4.4	3.5	**90.4**	1.1	8.5
DIIN	88.1	**84.9**	4.5	10.6	**55.0**	0.4	44.6
DR-BiLSTM	88.3	**89.7**	5.5	4.8	**82.1**	8.9	9.0
Human	-	2	**84**	14	10	2	**88**

Limitations of Regular Evaluation

Goal:

To show that regular evaluation **fails** to assess models deeper compositional understanding.

Limitations of Regular Evaluation

Goal:

To show that regular evaluation **fails** to assess models deeper compositional understanding.

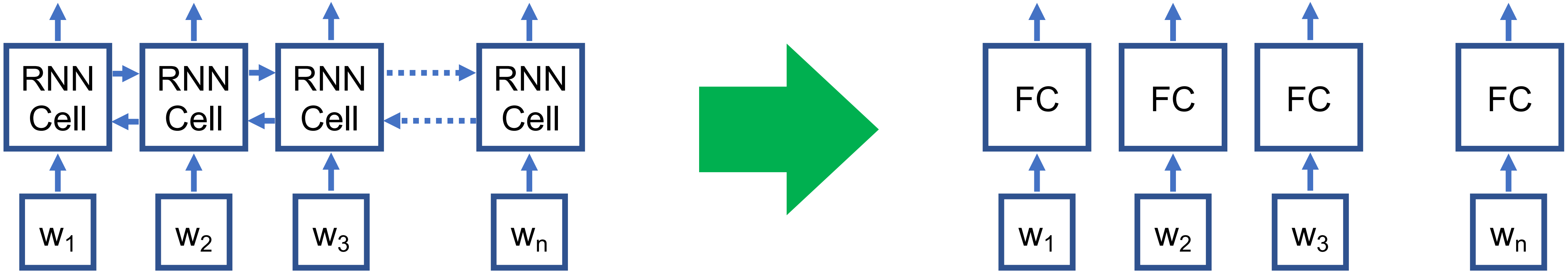
Method:

Train models with **compositional structures explicitly removed** and compare their results with those before on regular evaluation.

Limitations of Regular Evaluation

RNN replacement:

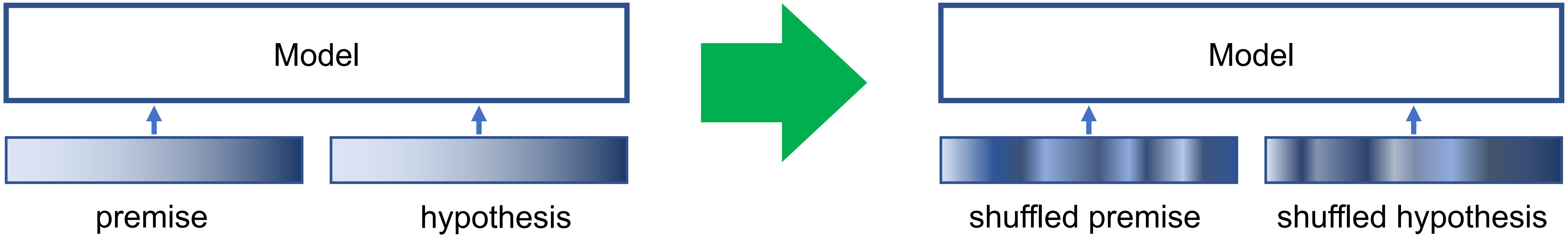
Create strong bag-of-words-like models by replacing RNN layers with fully-connected layers, and train them on the standard training set.



Limitations of Regular Evaluation

Word-Shuffled Training:

We train the NLI models with the words of the two input sentences shuffled, such that the compositional information is diluted and hard to learn.



Results

Model	SNLI			MNLI Matched			MNLI MisMatched		
	Original	BoW	WS	Original	BoW	WS	Original	BoW	WS
RSE	86.47	85.02	–	72.80	70.02	–	74.00	71.10	–
ESIM	88.17	82.37	86.79	76.16	68.98	73.70	76.22	69.77	74.20
DR-BiLSTM	88.28	82.81	86.90	76.90	70.11	73.27	77.49	70.70	73.25

Table 3: The ”Original” columns show results for vanilla RSE, ESIM and DR-BiLSTM on SNLI, MNLI matched, and MNLI mismatched dev set. The ”BoW” column show results for BoW-like variant of RSE, ESIM, and DR-BiLSTM by replacing their RNNs with fully-connected layers. The ”WS” columns show results for ESIM and DR-BiLSTM with words of input sentences shuffled during training.

Removing compositional structures doesn’t induce as much performance drop as expected.

Compositionality-Sensitivity Testing

We know that:

- Models are overly relying on lexical features via adversarial evaluation.
- Standard evaluation fails to reveal this issue.

How can we analyze models' compositionality sensitivity directly from **existing** natural datasets?

Compositionality-Sensitivity Testing

UNC NLP

Formalization:

Perfect Model: $p(y \mid x) = f_\theta(S_p, S_h, \Pi_p, \Pi_h)$

Bag-of-Words Model: $p(y \mid x) = g_\theta(S_p, S_h)$

Current Model: $p(y \mid x) = \hat{f}_\theta(\tilde{S}_p, \tilde{S}_h, \tilde{\Pi}_p, \tilde{\Pi}_h)$

$\tilde{S}_p \subseteq S_p$ and $\tilde{S}_h \subseteq S_h$ Sets of lexical features model captured

$\tilde{\Pi}_p \subseteq \Pi_p$ and $\tilde{\Pi}_h \subseteq \Pi_h$ Sets of compositional features model captured

Our hypothesis: $\tilde{\Pi}_p \ll \Pi_p$ and $\tilde{\Pi}_h \ll \Pi_h$

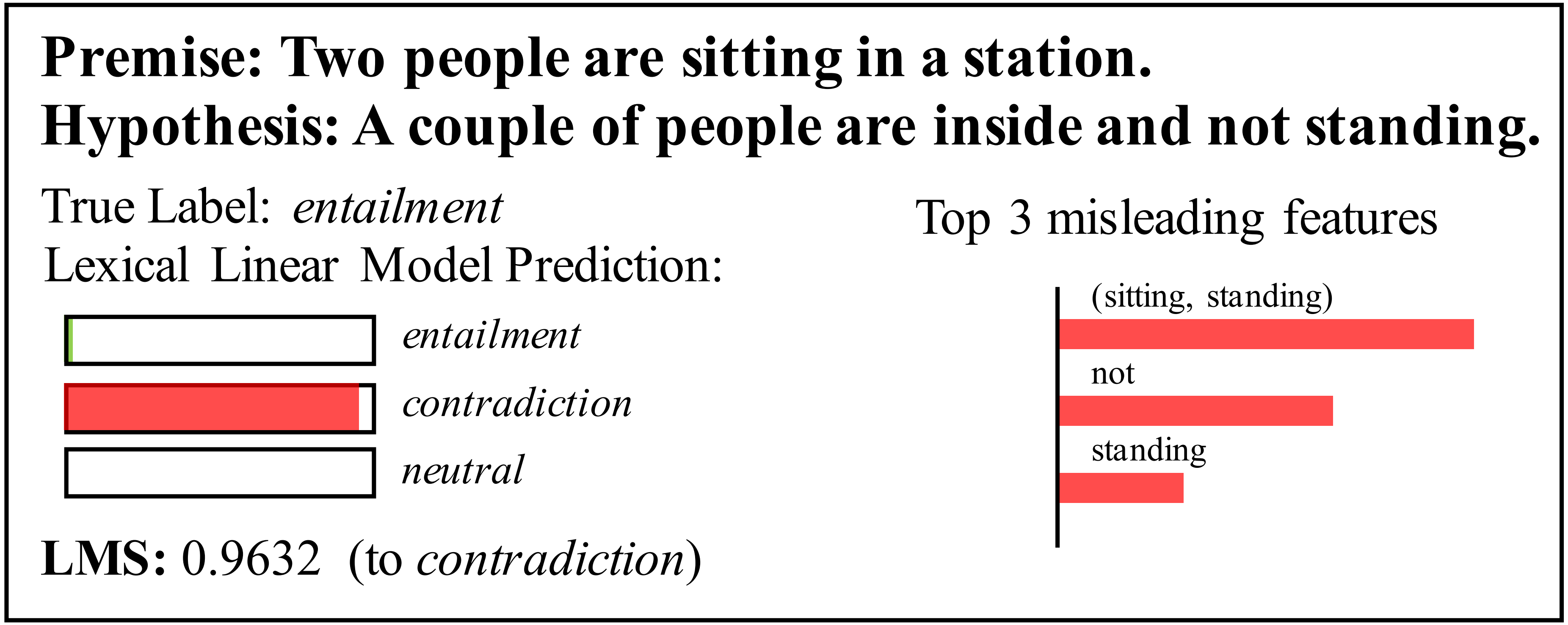
Lexically-Misleading Score

Formally, we define the **Lexically-Misleading Score** (LMS) of an NLI datapoint (x, c^*) as:

$$f_{LMS}(x, c^*) = \max_{c \in L \setminus \{c^*\}} p(c \mid x)$$

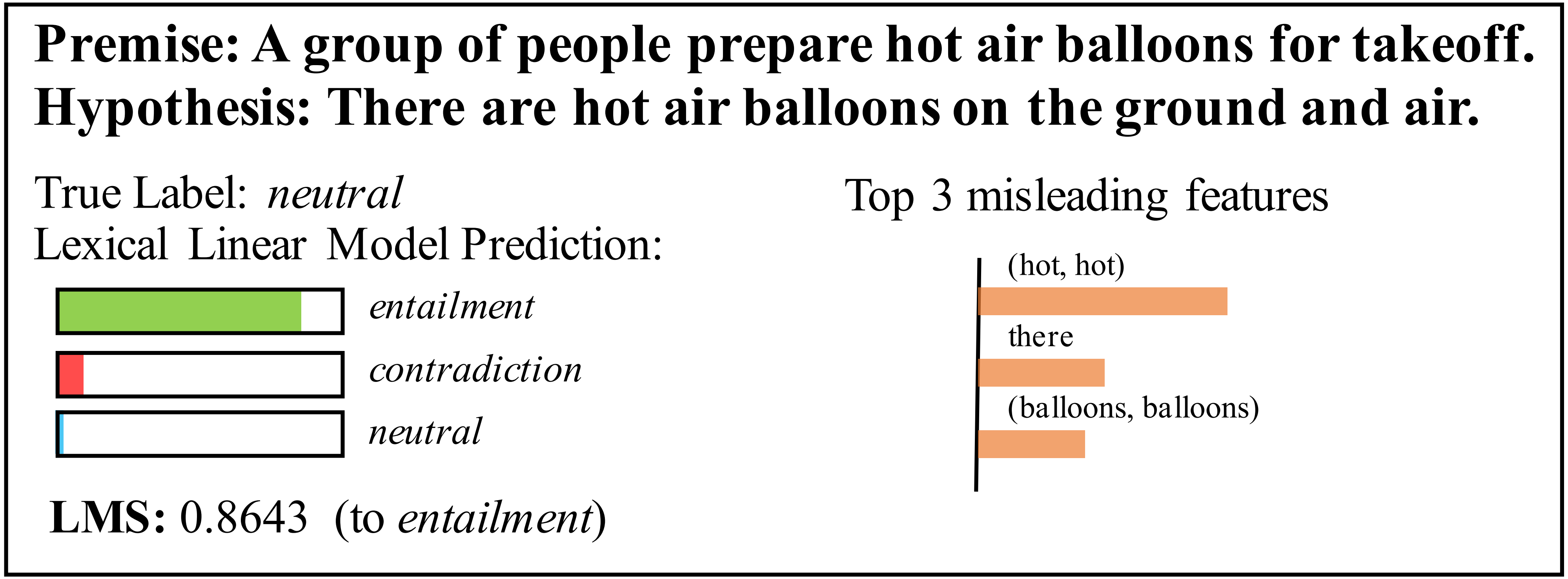
where c^* is the ground truth label, $p(c \mid x)$ is the probability generated by our regression model, and $L = \{\text{entailment, contradiction, neutral}\}$ is the label set. In other

Lexically-Misleading Score



Correct prediction for this example requires recognizing that ‘not standing’ and ‘sitting’ are the same state, rather than focusing on the superficial lexical clues such as ‘not’ and the cross unigram (‘sitting’, ‘standing’) that both mislead to ‘contradiction’.

Lexically-Misleading Score



For this example, word-overlap misleads the classifier to predict ‘entailment.

Compositionality-Sensitivity Testing

Given a standard evaluation set and associated ‘ground-truth’ labels, $D = \{(x_i, c_i)\}_{i=1}^{N}$, we create CS_λ, the compositionality-sensitivity evaluation set of confidence λ:

$$\mathrm{CS}_\lambda = \{(x_i, c_i) \in D \mid f_{LMS}(x_i, c_i) \geq \lambda\}$$

Compositionality-Sensitivity Results

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Compositionality-Sensitivity Results

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Compositionality-Sensitivity Results

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Compositionality-Sensitivity Results

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Compositionality-Sensitivity Results

UNC NLP

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Compositionality-Sensitivity Results

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	$\mathbf{CS_{0.7}}$
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

Table 5: Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottem on the table.

Thanks

Yixin Nie
yixin1@cs.unc.edu
www.cs.unc.edu/~yixin1

Yicheng Wang
yicheng@cs.unc.edu
www.cs.unc.edu/~yicheng

Mohit Bansal
mbansal@cs.unc.edu
www.cs.unc.edu/~mbansal

Acknowledgment: Verisk, Google, Facebook

UNC NLP