Natural Language Inference

Adam Kovacs

TU-Wien

adam.kovacs@tuwien.ac.at

May 24, 2021

Overview

Natural Language Inference

Natural Language Inference

- Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*)

Natural Language Inference

- Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*)
- The *premise* can entail, contradict or be neutral to the hypothesis

Natural Language Inference

- Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*)
- The *premise* can entail, contradict or be neutral to the hypothesis
- We mean entailment when a *human reading* the premise would infer that the hypothesis is true (Dagan, Glickman, and Magnini, 2006)

Natural Language Inference

- Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*)
- The *premise* can entail, contradict or be neutral to the hypothesis
- We mean entailment when a *human reading* the premise would infer that the hypothesis is true (Dagan, Glickman, and Magnini, 2006)
- Increasing popularity in creating high-performing datasets

Natural Language Inference

- Natural Language Inference (NLI) is the task of defining the semantic relation between a *premise* and a *conclusion* (or *hypothesis*)
- The *premise* can entail, contradict or be neutral to the hypothesis
- We mean entailment when a *human reading* the premise would infer that the hypothesis is true (Dagan, Glickman, and Magnini, 2006)
- Increasing popularity in creating high-performing datasets
- Necessary step towards Reasoning and Natural Language Understanding (NLU) (Condoravdi et al., 2003; Nangia et al., 2017)

Natural Language Inference

entailment
A young family enjoys feeling ocean waves lap at their feet. A family is at the beach
contradiction
There is no man wearing a black helmet and pushing a bicycle One man is wearing a black helmet and pushing a bicycle
neutral
An old man with a package poses in front of an advertisement. A man poses in front of an ad for beer.

Table: NLI examples

Datasets - Early - FraCas

- FraCas (Cooper et al., 1996) (only 874 unique sentences, and the data is constructed)
- It contains 346 "problem" types
- But covers lot of inference classes
- The examples are mostly logical inference cases

Datasets - Early - FraCas

"Yes" examples

1. *premise* - Just one accountant attended the meeting.
2. *hypothesis* - Some accountant attended the meeting.

Datasets - Early - FraCas

"Yes" examples

1. *premise* - Just one accountant attended the meeting.
2. *hypothesis* - Some accountant attended the meeting.

"No" examples

1. *premise* - Exactly two lawyers and three accountants signed the contract.
2. *hypothesis* - Six lawyers signed the contract.

Datasets - Early - FraCas

"Yes" examples

1. *premise* - Just one accountant attended the meeting.
2. *hypothesis* - Some accountant attended the meeting.

"No" examples

1. *premise* - Exactly two lawyers and three accountants signed the contract.
2. *hypothesis* - Six lawyers signed the contract.

"Unknown" examples

1. *premise* - Either Smith, Jones or Anderson signed the contract.
2. *hypothesis* - Jones signed the contract.

Datasets - Early - PASCAL

Datasets - Early - PASCAL

- The seven Recognizing Textual Entailment (RTE) challenge: (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan, Dolan, et al., 2010; Luisa Bentivogli et al., 2009; L. Bentivogli et al., 2011)

Datasets - Early - PASCAL

- The seven Recognizing Textual Entailment (RTE) challenge: (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan, Dolan, et al., 2010; Luisa Bentivogli et al., 2009; L. Bentivogli et al., 2011)
- Naturally occuring data, and then hypothesis based on the premise

Datasets - Early - PASCAL

- The seven Recognizing Textual Entailment (RTE) challenge: (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan, Dolan, et al., 2010; Luisa Bentivogli et al., 2009; L. Bentivogli et al., 2011)
- Naturally occuring data, and then hypothesis based on the premise
- They suffer from incorrect inference (Zaenen, Karttunen, and Crouch, 2005)

Datasets - Early - PASCAL

- The seven Recognizing Textual Entailment (RTE) challenge: (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan, Dolan, et al., 2010; Luisa Bentivogli et al., 2009; L. Bentivogli et al., 2011)
- Naturally occuring data, and then hypothesis based on the premise
- They suffer from incorrect inference (Zaenen, Karttunen, and Crouch, 2005)
- Still very small (around 1000 pairs)

Datasets - Early - PASCAL

- The seven Recognizing Textual Entailment (RTE) challenge: (Dagan, Glickman, and Magnini, 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Dagan, Dolan, et al., 2010; Luisa Bentivogli et al., 2009; L. Bentivogli et al., 2011)
- Naturally occuring data, and then hypothesis based on the premise
- They suffer from incorrect inference (Zaenen, Karttunen, and Crouch, 2005)
- Still very small (around 1000 pairs)
- First step towards including "non-logical" inferences and presupossed information

Datasets - Early - PASCAL

Entailment

1. *premise* - Bill murdered John.
2. *hypothesis* - Bill killed John.

Datasets - Early - PASCAL

Entailment

1. *premise* - Bill murdered John.
2. *hypothesis* - Bill killed John.

Not entailment

1. *premise* - Bill didn't murder John
2. *hypothesis* - Bill didn't kill John

Datasets - Early - PASCAL

Entailment

1. *premise* - Bill murdered John.
2. *hypothesis* - Bill killed John.

Not entailment

1. *premise* - Bill didn't murder John
2. *hypothesis* - Bill didn't kill John

Entailment

1. *premise* - Bill didn't kill John
2. *hypothesis* - Bill didn't murder John.

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - Green cards are becoming more difficult to obtain.
2. *hypothesis* - Green card is now difficult to receive.

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - Green cards are becoming more difficult to obtain.
2. *hypothesis* - Green card is now difficult to receive.

entailment

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - Green cards are becoming more difficult to obtain.
2. *hypothesis* - Green card is now difficult to receive.

entailment

1. *premise* - Hippos do come into conflict with people quite often
2. *hypothesis* - Hippopotamus attacks human

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - Green cards are becoming more difficult to obtain.
2. *hypothesis* - Green card is now difficult to receive.

entailment

1. *premise* - Hippos do come into conflict with people quite often
2. *hypothesis* - Hippopotamus attacks human

entailment

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - South Korean's deputy foreign minister says his country won't change its plan to send 3000 soldiers to Iraq.
2. *hypothesis* -South Korea continues to send troops

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - South Korean's deputy foreign minister says his country won't change its plan to send 3000 soldiers to Iraq.
2. *hypothesis* -South Korea continues to send troops

entailment

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - South Korean's deputy foreign minister says his country won't change its plan to send 3000 soldiers to Iraq.
2. *hypothesis* -South Korea continues to send troops

entailment

1. *premise* - The White House failed to act on the domestic threat from al Qaida prior to September 11, 2001.
2. *hypothesis* - White House ignored the threat of attack

Datasets - Early - PASCAL - Problems (Zaenen, Karttunen, and Crouch, 2005)

1. *premise* - South Korean's deputy foreign minister says his country won't change its plan to send 3000 soldiers to Iraq.
2. *hypothesis* -South Korea continues to send troops

entailment

1. *premise* - The White House failed to act on the domestic threat from al Qaida prior to September 11, 2001.
2. *hypothesis* - White House ignored the threat of attack

entailment

Datasets - SICK Marelli et al., 2014

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)
- Don't require dealing with named entities, temporal phenomena, etc..

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)
- Don't require dealing with named entities, temporal phenomena, etc..
- They made an effort to reduce the needed encyclopedic world-knowledge

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)
- Don't require dealing with named entities, temporal phenomena, etc..
- They made an effort to reduce the needed encyclopedic world-knowledge
- It was created from captions of pictures

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)
- Don't require dealing with named entities, temporal phenomena, etc..
- They made an effort to reduce the needed encyclopedic world-knowledge
- It was created from captions of pictures
- Sentences were normalized

Datasets - SICK Marelli et al., 2014

- English corpus of 9840 sentence pairs (rich in syntactic and semantic phenomena)
- Dataset for Distributional Semantic Models (DSMs)
- Don't require dealing with named entities, temporal phenomena, etc..
- They made an effort to reduce the needed encyclopedic world-knowledge
- It was created from captions of pictures
- Sentences were normalized
- In (Kalouli, Real, and Paiva, 2017) they showed the logical fallacies in the SICK dataset

Datasets - SICK - The process

Annotation process

Each normalized sentence was used to generate three new sentences based on a set of rules, such as adding passive or active voice, adding negations, etc. Each sentence was then paired with all of those three generated sentences. A native speaker eliminated odd and ungrammatical sentences.

Datasets - SICK

The turtle followed the fish -> The turtle is following the fish

Datasets - SICK

The turtle followed the fish -> The turtle is following the fish

Sentences were expanded to

The turtle is following the red fish
The turtle isn't following the fish
The fish is following the turtle.

Datasets - SICK - Problems (Kalouli, Real, and Paiva, 2017)

- Annotators were not given strict guidelines
- They were not told the origin of the sentences
- Contradictions in logic should be symmetric (if A is contradictory to B then B must be contradictory to A)
- 611 pairs of 9840 are annotated with logical fallacies
- A entails B -> B contradicts A is found

Datasets - SICK - Problematic sentences

1. A motorcycle is riding standing up on the seat of the vehicle.
2. The black and white dog isn't running and there is no person standing behind

Example from SICK

premise - An Asian woman in a crowd is not carrying a black bag
hypothesis - An Asian woman in a crowd is carrying a black bag

A **contradicts** B but B is **neutral** to A

Datasets - SICK - Problems

1. *premise* - The lady is cracking an egg into a bowl.
2. *hypothesis* - The lady is cracking an egg into a dish.

Datasets - SICK - Problems

1. *premise* - The lady is cracking an egg into a bowl.
2. *hypothesis* - The lady is cracking an egg into a dish.

A **entails** B, but B is **contradictory** to A

Datasets - SICK - Problems

1. *premise* - The lady is cracking an egg into a bowl.
2. *hypothesis* - The lady is cracking an egg into a dish.

A **entails** B, but B is **contradictory** to A

1. *premise* - The man is aiming a gun.
2. *hypothesis* - The man is drawing a gun.

Datasets - SICK - Problems

1. *premise* - The lady is cracking an egg into a bowl.
2. *hypothesis* - The lady is cracking an egg into a dish.

A **entails** B, but B is **contradictory** to A

1. *premise* - The man is aiming a gun.
2. *hypothesis* - The man is drawing a gun.

A **entails** B, but B is **contradictory** to A

Datasets - SNLI, MultiNLI

Datasets - SNLI, MultiNLI

- More recents set have exploded to some hundred thousand examples

Datasets - SNLI, MultiNLI

- More recents set have exploded to some hundred thousand examples
- Enabling the training of Deep Neural Models

Datasets - SNLI, MultiNLI

- More recents set have exploded to some hundred thousand examples
- Enabling the training of Deep Neural Models
- SNLI (S. R. Bowman et al., 2015)

Datasets - SNLI, MultiNLI

- More recents set have exploded to some hundred thousand examples
- Enabling the training of Deep Neural Models
- SNLI (S. R. Bowman et al., 2015)
- Multi-NLI (Williams, Nangia, and S. Bowman, 2018)

Datasets - SNLI, MultiNLI

- More recents set have exploded to some hundred thousand examples
- Enabling the training of Deep Neural Models
- SNLI (S. R. Bowman et al., 2015)
- Multi-NLI (Williams, Nangia, and S. Bowman, 2018)
- These training sets contain annotation artifacts (Gururangan et al., 2018)

Datasets - SNLI, MultiNLI

- The Stanford Natural Language Inference (SNLI) contains 570k human-written sentence
- The Multi-Genre Natural Language Inference (MultiNLI) corpus consists of 433k sentence pairs

Datasets - SNLI, MultiNLI

- The Stanford Natural Language Inference (SNLI) contains 570k human-written sentence
- The Multi-Genre Natural Language Inference (MultiNLI) corpus consists of 433k sentence pairs
- MultiNLI contains pairs from ten distinct genres
  - matched - from same genres
  - mismatched - from other genres

Datasets - SNLI, MultiNLI

- The Stanford Natural Language Inference (SNLI) contains 570k human-written sentence
- The Multi-Genre Natural Language Inference (MultiNLI) corpus consists of 433k sentence pairs
- MultiNLI contains pairs from ten distinct genres
  - matched - from same genres
  - mismatched - from other genres
- In contrary of the SICK dataset the annotators were given the freedom to write themselves a conclusion sentence
- They also knew the context of the dataset (it comes from image captions)

Examples from SICK, SNLI, Multi-NLI (Talman and Chatzikyriakidis, 2019)

	entailment
SICK	*A person, who is riding a bike, is wearing gear which is black* *A biker is wearing gear which is black*
SNLI	*A young family enjoys feeling ocean waves lap at their feet.* *A family is at the beach.*
MultiNLI	*Kal tangled both of Adrin's arms, keeping the blades far away.* *Adrin's arms were tangled, keeping his blades away from Kal.*
	contradiction
SICK	*There is no man wearing a black helmet and pushing a bicycle* *One man is wearing a black helmet and pushing a bicycle*
SNLI	*A man with a tattoo on his arm staring to the side with vehicles and buildings behind him.* *A man with no tattoos is getting a massage.*
MultiNLI	*Also in Eustace Street is an information office and a cultural center for children, The Ark .* *The Ark, a cultural center for kids, is located in Joyce Street.*
	neutral
SICK	*A little girl in a green coat and a boy holding a red sled are walking in the snow* *A child is wearing a coat and is carrying a red sled near a child in a green and black coat*
SNLI	*An old man with a package poses in front of an advertisement.* *A man poses in front of an ad for beer.*
MultiNLI	*Enthusiasm for Disney's Broadway production of The Lion King dwindles.* *The broadway production of The Lion King was amazing, but audiences are getting bored.*

Table 2: Example sentence pairs from the three datasets.

Annotation Artifacts in Natural Language Inference Data (Gururangan et al., 2018)

- The paper showed that the data leaves clues about the labels
- It makes it possible to identify the label from the hypothesis
- Simple classification models -> 67% of SNLI and 53% of MultiNLI
- Linguistic phemomena like negation and vagueness correlates with the classes

Annotation Artifacts in Natural Language Inference Data

Criteria

Entailment - h is definitely true given p
Neutral - h might be true given p
Contradiction - h is definitely not true given p

Annotation Artifacts in Natural Language Inference Data

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

Annotation Artifacts in Natural Language Inference Data

	Entailment		Neutral		Contradiction	
SNLI	outdoors	2.8%	tall	0.7%	nobody	0.1%
	least	0.2%	first	0.6%	sleeping	3.2%
	instrument	0.5%	competition	0.7%	no	1.2%
	outside	8.0%	sad	0.5%	tv	0.4%
	animal	0.7%	favorite	0.4%	cat	1.3%
MNLI	some	1.6%	also	1.4%	never	5.0%
	yes	0.1%	because	4.1%	no	7.6%
	something	0.9%	popular	0.7%	nothing	1.4%
	sometimes	0.2%	many	2.2%	any	4.1%
	various	0.1%	most	1.8%	none	0.1%

Table 4: Top 5 words by PMI(*word*, *class*), along with the proportion of *class* training samples containing *word*. MultiNLI is abbreviated to MNLI.

Premise - Two dogs are running through a field

Entailment -
There are *animals outdoors.*

Neutral -
Some puppies are running *to catch a stick.*

Contradiction -
The pets are *sitting on a couch*

Annotation Artifacts in Natural Language Inference Data

Model	SNLI			MultiNLI Matched			MultiNLI Mismatched		
	Full	*Hard*	*Easy*	*Full*	*Hard*	*Easy*	*Full*	*Hard*	*Easy*
DAM	84.7	69.4	92.4	72.0	55.8	85.3	72.1	56.2	85.7
ESIM	85.8	71.3	92.6	74.1	59.3	86.2	73.1	58.9	85.2
DIIN	86.5	72.7	93.4	77.0	64.1	87.6	76.5	64.4	86.8

Table 5: Performance of high-performing NLI models on the full, *Hard*, and *Easy* NLI test sets.

Testing the Generalization Power of Neural Network Models across NLI Benchmarks

- Discussed in (Talman and Chatzikyriakidis, 2019)
- Conference paper on **BlackboxNLP**[1]

[1] `https://blackboxnlp.github.io/`

Testing the Generalization Power of Neural Network Models across NLI Benchmarks

- Discussed in (Talman and Chatzikyriakidis, 2019)
- Conference paper on **BlackboxNLP**[1]
- Current SOTA systems are over 90% accuracy on SICK, SNLI, Multi-NLI

[1] `https://blackboxnlp.github.io/`

Testing the Generalization Power of Neural Network Models across NLI Benchmarks

- Discussed in (Talman and Chatzikyriakidis, 2019)
- Conference paper on **BlackboxNLP**[1]
- Current SOTA systems are over 90% accuracy on SICK, SNLI, Multi-NLI
- The goal of the paper is to show that these results are benchmark specific

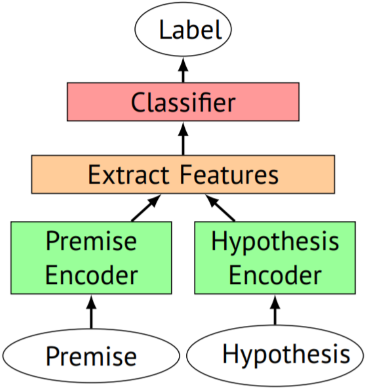
[1] https://blackboxnlp.github.io/

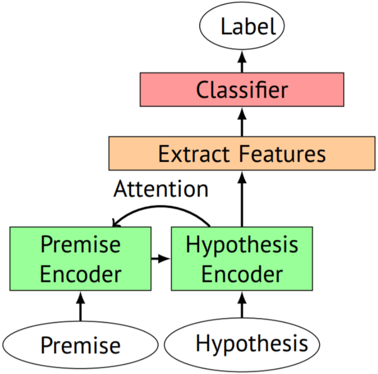
Testing the Generalization Power of Neural Network Models across NLI Benchmarks

- Discussed in (Talman and Chatzikyriakidis, 2019)
- Conference paper on **BlackboxNLP**[1]
- Current SOTA systems are over 90% accuracy on SICK, SNLI, Multi-NLI
- The goal of the paper is to show that these results are benchmark specific
- They trained six SOTA neural models
- They showed that each of them has problems generalizing

[1] https://blackboxnlp.github.io/

Neural NLI models (Glockner, Shwartz, and Goldberg, 2018)





The trained models

Model	Model type
BiLSTM-max (Conneau et al., 2017)	Sentence encoding
HBMP (Talman et al., 2018)	Sentence encoding
ESIM (Chen et al., 2017)	Cross-sentence attention
KIM (Chen et al., 2018)	Cross-sentence attention
ESIM + ELMo (Peters et al., 2018)	Pre-trained language model
BERT-base (Devlin et al., 2019)	Cross-sentence attention + pre-trained language model

Table 3: Model architectures used in the experiments.

- BiLSTM-max - Standard BiLSTM architecture with max pooling
- Hierarchical BiLSTM Max Pooling Architecture (HBMP)
- Enhanced Sequential Inference Model(ESIM) - Enhanced LSTM architecture with Attention
- Knowledge-based InferenceModel (KIM) enriches ESIM with external knowledge
- ESIM + ELMo - ESIM architecture with ELMo contextualized embeddings
- BERT-base - Fine tuning BERT

Combinations of the models

Train data	Test data	Size of the training set	Size of the test set
SNLI	**SNLI**	550,152	10,000
SNLI	MultiNLI	550,152	20,000
SNLI	SICK	550,152	9,840
MultiNLI	**MultiNLI**	392,702	20,000
MultiNLI	SNLI	392,702	10,000
MultiNLI	SICK	392,702	9,840
SNLI + MultiNLI	**SNLI**	942,854	10,000
SNLI + MultiNLI	SICK	942,854	9,840

Table 1: Dataset combinations used in the experiments. The rows in bold are baseline experiments, where the test data comes from the same benchmark as the training and development data.

SNLI models

Train data	Test data	Test accuracy	Δ	Model
SNLI	**SNLI**	**86.1**		**BiLSTM-max** (our baseline)
SNLI	**SNLI**	**86.6**		**HBMP** (Talman et al., 2018)
SNLI	**SNLI**	**88.0**		**ESIM** (Chen et al., 2017)
SNLI	**SNLI**	**88.6**		**KIM** (Chen et al., 2018)
SNLI	**SNLI**	**88.6**		**ESIM + ELMo** (Peters et al., 2018)
SNLI	**SNLI**	**90.4**		**BERT-base** (Devlin et al., 2019)
SNLI	MultiNLI-m	55.7*	-30.4	BiLSTM-max
SNLI	MultiNLI-m	56.3*	-30.3	HBMP
SNLI	MultiNLI-m	59.2*	-28.8	ESIM
SNLI	MultiNLI-m	61.7*	-26.9	KIM
SNLI	MultiNLI-m	64.2*	-24.4	ESIM + ELMo
SNLI	MultiNLI-m	75.5*	-14.9	BERT-base
SNLI	SICK	54.5	-31.6	BiLSTM-max
SNLI	SICK	53.1	-33.5	HBMP
SNLI	SICK	54.3	-33.7	ESIM
SNLI	SICK	55.8	-32.8	KIM
SNLI	SICK	56.7	-31.9	ESIM + ELMo
SNLI	SICK	56.9	-33.5	BERT-base

MultiNLI models

MultiNLI	**MultiNLI-m**	**73.1***		**BiLSTM-max**
MultiNLI	**MultiNLI-m**	**73.2***		**HBMP**
MultiNLI	**MultiNLI-m**	**76.8***		**ESIM**
MultiNLI	**MultiNLI-m**	**77.3***		**KIM**
MultiNLI	**MultiNLI-m**	**80.2***		**ESIM + ELMo**
MultiNLI	**MultiNLI-m**	**84.0***		**BERT-base**
MultiNLI	SNLI	63.8	-9.3	BiLSTM-max
MultiNLI	SNLI	65.3	-7.9	HBMP
MultiNLI	SNLI	66.4	-10.4	ESIM
MultiNLI	SNLI	68.5	-8.8	KIM
MultiNLI	SNLI	69.1	-11.1	ESIM + ELMo
MultiNLI	SNLI	80.4	-3.6	BERT-base
MultiNLI	SICK	54.1	-19.0	BiLSTM-max
MultiNLI	SICK	54.1	-19.1	HBMP
MultiNLI	SICK	47.9	-28.9	ESIM
MultiNLI	SICK	50.9	-26.4	KIM
MultiNLI	SICK	51.4	-28.8	ESIM + ELMo
MultiNLI	SICK	55.0	-29.0	BERT-base

SNLI+MultiNLI models

SNLI + MultiNLI	**SNLI**	**86.1**		**BiLSTM-max**
SNLI + MultiNLI	**SNLI**	**86.1**		**HBMP**
SNLI + MultiNLI	**SNLI**	**87.5**		**ESIM**
SNLI + MultiNLI	**SNLI**	**86.2**		**KIM**
SNLI + MultiNLI	**SNLI**	**88.8**		**ESIM + ELMo**
SNLI + MultiNLI	**SNLI**	**90.6**		**BERT-base**
SNLI + MultiNLI	SICK	54.5	-31.6	BiLSTM-max
SNLI + MultiNLI	SICK	55.0	-31.1	HBMP
SNLI + MultiNLI	SICK	54.5	-33.0	ESIM
SNLI + MultiNLI	SICK	54.6	-31.6	KIM
SNLI + MultiNLI	SICK	57.1	-31.7	ESIM + ELMo
SNLI + MultiNLI	SICK	59.1	-31.5	BERT-base

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences (Glockner, Shwartz, and Goldberg, 2018)

- The authors constructed a new test set
- The premise remained the same from SNLI
- In the hypothesis they replaced a single term from the premise

Contradiction

The man is holding a *saxophone* -> The man is holding an *electric guitar*

Neutral

A little girl is very *sad* -> A little girl is very *unhappy*

Entailment

A couple drinking *wine* $\rightarrow$ A couple drinking *champagne*

Results (Glockner, Shwartz, and Goldberg, 2018)

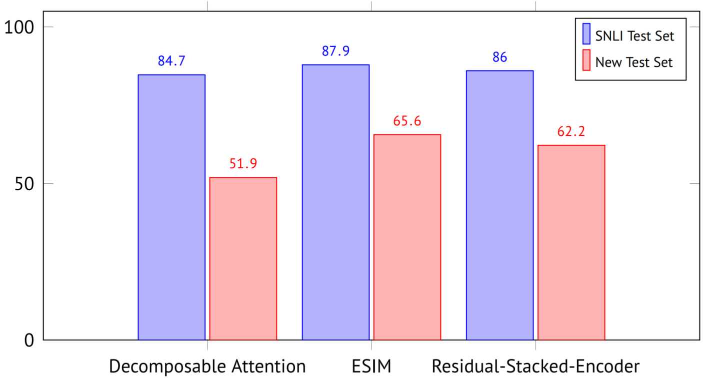


Figure: Training models on SNLI and testing on the new test set. Big drop in the performance.

Results (Glockner, Shwartz, and Goldberg, 2018)

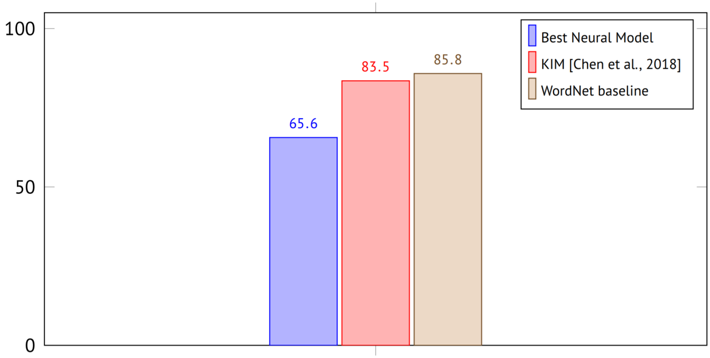


Figure: WordNet models solve the problem better.

Outlook to other datasets - Lexical Entailment - Semeval

- Lexical Entailment is a relaxed version of NLI, where we are only concerned with IS_A relations
- Semeval task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment" (**Glavas:2020**)
- From HyperLex (**Vulic:2017b**)
- Candidate word pairs for human annotation were gathered from the USF (**Nelson:2004**) and WordNet (**Miller:1995**) databases.
- *mole -> animal*

Outlook to other datasets - SherLlic

- More challenging dataset -> SherLlic dataset of lexical inference in context (**Schmitt:2019**)
- Extracting inference candidates from the ClueWeb corpus (**Gabrilovich:2013**)
- The pairs are chosen based on distributional evidence
- This makes them completely novel
- *run* entails *lead* if PERSON and COMPANY (e.g., *Bezos runs Amazon*)
- Does not if COMPUTER and SOFTWARE, as in *my mac runs macOS*.

Bibliography I

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo (Jan. 2006). "The second PASCAL recognising textual entailment challenge". In: *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Bentivogli, L., Peter Clark, I. Dagan, and Danilo Giampiccolo (2011). "The Seventh PASCAL Recognizing Textual Entailment Challenge". In: *Theory and Applications of Categories*.

Bentivogli, Luisa, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini (2009). "The Fifth PASCAL Recognizing Textual Entailment Challenge". In: *In Proc Text Analysis Conference (TAC'09*.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (Sept. 2015). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://www.aclweb.org/anthology/D15-1075.

Condoravdi, Cleo, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow (2003). "Entailment, intensionality and text understanding". In: *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pp. 38–45. URL: https://www.aclweb.org/anthology/W03-0906.

Cooper, Robin et al. (Mar. 1996). "Using the Framework". In:

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth (Jan. 2010). "The Fourth Pascal Recognizing Textual Entailment Challenge". In: *Journal of Natural Language Engineering*. URL: https://www.microsoft.com/en-us/research/publication/the-fourth-pascal-recognizing-textual-entailment-challenge/.

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2006). "The PASCAL Recognising Textual Entailment Challenge". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Ed. by Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 177–190. ISBN: 978-3-540-33428-6.

Bibliography II

Giampiccolo, Danilo, Bernardo Magnini, Ido Dagan, and Bill Dolan (2007). "The Third PASCAL Recognizing Textual Entailment Challenge". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. USA: Association for Computational Linguistics, pp. 1–9.

Glockner, Max, Vered Shwartz, and Yoav Goldberg (July 2018). "Breaking NLI Systems with Sentences that Require Simple Lexical Inferences". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 650–655. DOI: 10.18653/v1/P18-2103. URL: https://www.aclweb.org/anthology/P18-2103.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (June 2018). "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://www.aclweb.org/anthology/N18-2017.

Kalouli, Aikaterini-Lida, Livy Real, and Valeria de Paiva (2017). "Textual Inference: getting logic from humans". In: *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*. URL: https://www.aclweb.org/anthology/W17-6915.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli (May 2014). *The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment*. Zenodo. DOI: 10.5281/zenodo.2787612. URL: https://doi.org/10.5281/zenodo.2787612.

Nangia, Nikita, Adina Williams, Angeliki Lazaridou, and Samuel Bowman (Sept. 2017). "The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations". In: *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1–10. DOI: 10.18653/v1/W17-5301. URL: https://www.aclweb.org/anthology/W17-5301.

Bibliography III

Talman, Aarne and Stergios Chatzikyriakidis (Aug. 2019). "Testing the Generalization Power of Neural Network Models across NLI Benchmarks". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 85–94. DOI: 10.18653/v1/W19-4810. URL: https://www.aclweb.org/anthology/W19-4810.

Williams, Adina, Nikita Nangia, and Samuel Bowman (June 2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://www.aclweb.org/anthology/N18-1101.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch (June 2005). "Local Textual Inference: Can it be Defined or Circumscribed?" In: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 31–36. URL: https://www.aclweb.org/anthology/W05-1206.

Thank you