Adam Kovacs
TU-Wien
adam.kovacs@tuwien.ac.at

May 24, 2021
Overview

1. NLI introduction
2. Early Datasets
   - FraCas
   - PASCAL
3. Modern Datasets
   - SICK
   - SNLI, MultiNLI
4. Annotation Artifacts in the datasets
5. Generalization power of NLI models
6. Breaking NLI systems
7. Lexical Entailment
   - Semeval
   - SherLlic
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 is a necessary step towards Reasoning and Natural Language Understanding (NLU) (Condoravdi et al., 2003; Nangia et al., 2017).
Natural Language Inference (NLI) is the task of defining the semantic relation between a premise and a conclusion (or hypothesis).
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 (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 (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 (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)
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
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
"Yes" examples

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

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 occurring 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 presupposed information
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)
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 occurring data, and then hypothesis based on the premise
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 occurring data, and then hypothesis based on the premise

They suffer from incorrect inference (Zaenen, Karttunen, and Crouch, 2005)
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 occurring data, and then hypothesis based on the premise
- They suffer from incorrect inference (Zaenen, Karttunen, and Crouch, 2005)
- Still very small (around 1000 pairs)
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 occurring 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 presupposed information
Entailment

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


2. **hypothesis** - White House ignored the threat of attack
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**


2. **hypothesis** - White House ignored the threat of attack

   **entailment**
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 Marelli et al., 2014

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

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

More recent sets 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

- More recent sets have exploded to some hundred thousand examples
Datasets - SNLI, MultiNLI

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

Enabling the training of Deep Neural Models

SNLI (S. R. Bowman et al., 2015)
More recent sets 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)
More recent sets 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).
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).
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
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)

<table>
<thead>
<tr>
<th></th>
<th>SICK</th>
</tr>
</thead>
<tbody>
<tr>
<td>entailment</td>
<td>A person, who is riding a bike, is wearing gear which is black</td>
</tr>
<tr>
<td></td>
<td>A biker is wearing gear which is black</td>
</tr>
<tr>
<td>SNLI</td>
<td>A young family enjoys feeling ocean waves lap at their feet.</td>
</tr>
<tr>
<td></td>
<td>A family is at the beach.</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>Kal tangled both of Adrin’s arms, keeping the blades far away.</td>
</tr>
<tr>
<td></td>
<td>Adrin’s arms were tangled, keeping his blades away from Kal.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SICK</th>
</tr>
</thead>
<tbody>
<tr>
<td>contradiction</td>
<td>There is no man wearing a black helmet and pushing a bicycle</td>
</tr>
<tr>
<td></td>
<td>One man is wearing a black helmet and pushing a bicycle</td>
</tr>
<tr>
<td>SNLI</td>
<td>A man with a tattoo on his arm staring to the side with vehicles and buildings behind him.</td>
</tr>
<tr>
<td></td>
<td>A man with no tattoos is getting a massage.</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>Also in Eustace Street is an information office and a cultural center for children, The Ark.</td>
</tr>
<tr>
<td></td>
<td>The Ark, a cultural center for kids, is located in Joyce Street.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SICK</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>A little girl in a green coat and a boy holding a red sled are walking in the snow</td>
</tr>
<tr>
<td></td>
<td>A child is wearing a coat and is carrying a red sled near a child in a green and black coat</td>
</tr>
<tr>
<td>SNLI</td>
<td>An old man with a package poses in front of an advertisement.</td>
</tr>
<tr>
<td></td>
<td>A man poses in front of an ad for beer.</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>Enthusiasm for Disney’s Broadway production of The Lion King dwindles.</td>
</tr>
<tr>
<td></td>
<td>The broadway production of The Lion King was amazing, but audiences are getting bored.</td>
</tr>
</tbody>
</table>

Table 2: Example sentence pairs from the three datasets.
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 phenomena like negation and vagueness correlates with the classes.
## Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entailment</strong></td>
<td>( h ) is definitely true given ( p )</td>
</tr>
<tr>
<td><strong>Neutral</strong></td>
<td>( h ) might be true given ( p )</td>
</tr>
<tr>
<td><strong>Contradiction</strong></td>
<td>( h ) is definitely not true given ( p )</td>
</tr>
</tbody>
</table>
### Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol.

<table>
<thead>
<tr>
<th>Premise</th>
<th>A woman selling bamboo sticks talking to two men on a loading dock.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entailment</strong></td>
<td>There are at least three <em>people</em> on a loading dock.</td>
</tr>
<tr>
<td><strong>Neutral</strong></td>
<td>A woman is selling bamboo sticks to help provide for her family.</td>
</tr>
<tr>
<td><strong>Contradiction</strong></td>
<td>A woman is <em>not</em> taking money for any of her sticks.</td>
</tr>
</tbody>
</table>
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*
### Table 5: Performance of high-performing NLI models on the full, *Hard*, and *Easy* NLI test sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>SNLI</th>
<th>MultiNLI Matched</th>
<th>MultiNLI Mismatched</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Full</em></td>
<td><em>Hard</em></td>
<td><em>Easy</em></td>
</tr>
<tr>
<td>DAM</td>
<td>84.7</td>
<td>69.4</td>
<td>92.4</td>
</tr>
<tr>
<td>ESIM</td>
<td>85.8</td>
<td>71.3</td>
<td>92.6</td>
</tr>
<tr>
<td>DIIN</td>
<td>86.5</td>
<td>72.7</td>
<td>93.4</td>
</tr>
</tbody>
</table>
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)

Diagram: Two architectures for NLI models. The left side shows a model with separate encoders for the premise and hypothesis, followed by an extract features layer and a classifier. The right side shows a model with attention between the premise and hypothesis encoders before the classifier.
The trained models

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

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 Inference Model (KIM) enriches ESIM with external knowledge
- ESIM + ELMo - ESIM architecture with ELMo contextualized embeddings
- BERT-base - Fine tuning BERT
## Combinations of the models

<table>
<thead>
<tr>
<th>Train data</th>
<th>Test data</th>
<th>Size of the training set</th>
<th>Size of the test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNLI</td>
<td>SNLI</td>
<td>550,152</td>
<td>10,000</td>
</tr>
<tr>
<td>SNLI</td>
<td>MultiNLI</td>
<td>550,152</td>
<td>20,000</td>
</tr>
<tr>
<td>SNLI</td>
<td>SICK</td>
<td>550,152</td>
<td>9,840</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI</td>
<td>392,702</td>
<td>20,000</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>392,702</td>
<td>10,000</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>392,702</td>
<td>9,840</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>942,854</td>
<td>10,000</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>942,854</td>
<td>9,840</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>MultiNLI</th>
<th>MultiNLI-m</th>
<th>Score</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>73.1*</td>
<td>BiLSTM-max</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>73.2*</td>
<td>HBMP</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>76.8*</td>
<td>ESIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>77.3*</td>
<td>KIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>80.2*</td>
<td>ESIM + ELMo</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>MultiNLI-m</td>
<td>84.0*</td>
<td>BERT-base</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>63.8</td>
<td>BiLSTM-max</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>65.3</td>
<td>HBMP</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>66.4</td>
<td>ESIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>68.5</td>
<td>KIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>69.1</td>
<td>ESIM + ELMo</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SNLI</td>
<td>80.4</td>
<td>BERT-base</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>54.1</td>
<td>BiLSTM-max</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>54.1</td>
<td>HBMP</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>47.9</td>
<td>ESIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>50.9</td>
<td>KIM</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>51.4</td>
<td>ESIM + ELMo</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>SICK</td>
<td>55.0</td>
<td>BERT-base</td>
</tr>
<tr>
<td>Model</td>
<td>Dataset</td>
<td>Score</td>
<td>Baseline</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>86.1</td>
<td>BiLSTM-max</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>86.1</td>
<td>HBMP</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>87.5</td>
<td>ESIM</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>86.2</td>
<td>KIM</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>88.8</td>
<td>ESIM + ELMo</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SNLI</td>
<td>90.6</td>
<td>BERT-base</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>54.5</td>
<td>BiLSTM-max</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>55.0</td>
<td>HBMP</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>54.5</td>
<td>ESIM</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>54.6</td>
<td>KIM</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>57.1</td>
<td>ESIM + ELMo</td>
</tr>
<tr>
<td>SNLI + MultiNLI</td>
<td>SICK</td>
<td>59.1</td>
<td>BERT-base</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Contradiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man is holding a <em>saxophone</em> $\rightarrow$ The man is holding an <em>electric guitar</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>A little girl is very <em>sad</em> $\rightarrow$ A little girl is very <em>unhappy</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>A couple drinking <em>wine</em> $\rightarrow$ A couple drinking <em>champagne</em></td>
</tr>
</tbody>
</table>
Figure: Training models on SNLI and testing on the new test set. Big drop in the performance.
Figure: WordNet models solve the problem better.
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
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.


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


Thank you