Natural Language Inference (and the representation of sentences)

Computational Semantics 2021

Adam Ek

Plan

- Part 1: Natural Language Inference
- Part 2: Sentence embeddings and Natural Language Inference Models

$\rightarrow$ 15 min break $\leftarrow$

- Part 3: Where are we?
- Part 4: Moving forward?

Languages in this course

English

Natural Language Inference (1)

- The task of Natural Language Inference (NLI) is finding the relationship between two sentences

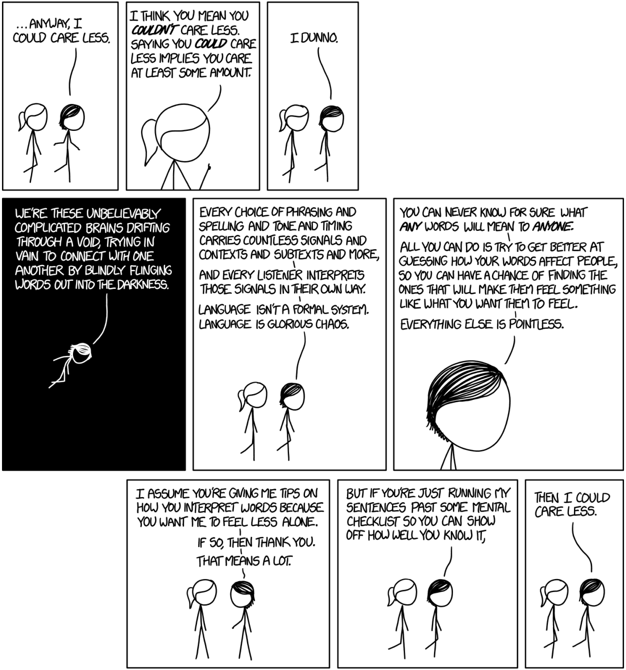
Natural Language Inference (1)

- The task of Natural Language Inference (NLI) is finding the relationship between two sentences
- The relations we are interested in are:
  - Entailment: the hypothesis is true given the premise
  - Contradiction: the hypothesis is false given the premise
  - Neutral: the hypothesis may be true given the premise

Natural Language Inference (1)

- The task of Natural Language Inference (NLI) is finding the relationship between two sentences
- The relations we are interested in are:
  - Entailment: the hypothesis is true given the premise
  - Contradiction: the hypothesis is false given the premise
  - Neutral: the hypothesis may be true given the premise

Some NLP taxonomy/terminology: NLI is a subfield of NLU (Nautral Language Understanding)


…ANYWAY, I COULD CARE LESS.
I THINK YOU MEAN YOU *COULDN'T* CARE LESS. SAYING YOU *COULD* CARE LESS IMPLIES YOU CARE AT LEAST SOME AMOUNT.
I DUNNO.
WE'RE THESE UNBELIEVABLY COMPLICATED BRAINS DRIFTING THROUGH A VOID, TRYING IN VAIN TO CONNECT WITH ONE ANOTHER BY BLINDLY FLINGING WORDS OUT INTO THE DARKNESS.
EVERY CHOICE OF PHRASING AND SPELLING AND TONE AND TIMING CARRIES COUNTLESS SIGNALS AND CONTEXTS AND SUBTEXTS AND MORE,
AND EVERY LISTENER INTERPRETS THOSE SIGNALS IN THEIR OWN WAY.
LANGUAGE ISN'T A FORMAL SYSTEM. LANGUAGE IS GLORIOUS CHAOS.
YOU CAN NEVER KNOW FOR SURE WHAT *ANY* WORDS WILL MEAN TO *ANYONE*.
ALL YOU CAN DO IS TRY TO GET BETTER AT GUESSING HOW YOUR WORDS AFFECT PEOPLE, SO YOU CAN HAVE A CHANCE OF FINDING THE ONES THAT WILL MAKE THEM FEEL SOMETHING LIKE WHAT YOU WANT THEM TO FEEL.
EVERYTHING ELSE IS POINTLESS.
I ASSUME YOU'RE GIVING ME TIPS ON HOW YOU INTERPRET WORDS BECAUSE YOU WANT ME TO FEEL LESS ALONE.
IF SO, THEN THANK YOU.
THAT MEANS A LOT.
BUT IF YOU'RE JUST RUNNING MY SENTENCES PAST SOME MENTAL CHECKLIST SO YOU CAN SHOW OFF HOW WELL YOU KNOW IT,
THEN I COULD CARE LESS.

Natural Language Inference (2)

- A NLI example contains a *premise* (P) and a *hypothesis* (H)

Natural Language Inference (2)

- A NLI example contains a *premise* (P) and a *hypothesis* (H)
- The task is to determine the relationship between H and P

Natural Language Inference (2)

- A NLI example contains a *premise* (P) and a *hypothesis* (H)
- The task is to determine the relationship between H and P
- For example, what is the relationship of the H given P below:

P A cat and a dog are playing hockey

H1 Two pets are playing hockey

H2 Two animals are playing hockey

H3 Three animals are playing hockey

Natural Language Inference (3)

- Approaches
  - First-order logic, lambda calculus etc
  - Statistical and Neural methods

Natural Language Inference (3)

- Approaches
  - First-order logic, lambda calculus etc
  - Statistical and Neural methods
- "Tasks" needed to perform NLI (very broadly)
  - Common-sense knowledge
  - Syntactic understanding
  - Semantic understanding
  - How two sentences relate to each other

Neural NLI

- To tackle this problem with neural networks we will use *sentence representations* of the *hypothesis* and *premise* to determine which class the pair belongs to.
- Mainly then, we need to construct good *sentence representations* and *combine* them.

Sentence representations

- We've seen one sentence representation already: the final *hidden state*

Sentence representations

- We've seen one sentence representation already: the final *hidden state*

What you can cram into a single $&!#* vector:
Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

Marco Baroni
Facebook AI Research
mbaroni@fb.com

- Setup: Train models on different NLP tasks and investigate how well they predict linguistic properties

Sentence representations

- We've seen one sentence representation already: the final *hidden state*

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

Marco Baroni
Facebook AI Research
mbaroni@fb.com

- Setup: Train models on different NLP tasks and investigate how well they predict linguistic properties
- Use the last *hidden state* or *max pooling* (similar to what we did in the Word2Vec lab)

Predicting linguistic properties

Task	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
Baseline representations										
Majority vote	20.0	0.5	17.9	5.0	50.0	50.0	50.0	50.0	50.0	50.0
Hum. Eval.	100	100	84.0	84.0	98.0	85.0	88.0	86.5	81.2	85.0
Length	**100**	0.2	18.1	9.3	50.6	56.5	50.3	50.1	50.2	50.0
NB-uni-tfidf	22.7	**97.8**	24.1	41.9	49.5	77.7	68.9	64.0	38.0	50.5
NB-bi-tfidf	23.0	95.0	24.6	53.0	**63.8**	75.9	69.1	65.4	39.9	**55.7**
BoV-fastText	66.6	91.6	**37.1**	**68.1**	50.8	**89.1**	**82.1**	**79.8**	**54.2**	54.8
BiLSTM-last encoder										
Untrained	36.7	43.8	28.5	76.3	49.8	84.9	84.7	74.7	51.1	64.3
AutoEncoder	**99.3**	23.3	35.6	78.2	62.0	84.3	84.7	82.1	49.9	65.1
NMT En-Fr	83.5	**55.6**	42.4	81.6	62.3	88.1	89.7	89.5	52.0	71.2
NMT En-De	83.8	53.1	42.1	81.8	60.6	88.6	89.3	87.3	51.5	**71.3**
NMT En-Fi	82.4	52.6	40.8	81.3	58.8	88.4	86.8	85.3	52.1	71.0
Seq2Tree	94.0	14.0	**59.6**	**89.4**	**78.6**	**89.9**	**94.4**	**94.7**	49.6	67.8
SkipThought	68.1	35.9	33.5	75.4	60.1	89.1	80.5	77.1	**55.6**	67.7
NLI	75.9	47.3	32.7	70.5	54.5	79.7	79.3	71.3	53.3	66.5
BiLSTM-max encoder										
Untrained	73.3	**88.8**	46.2	71.8	70.6	89.2	85.8	81.9	73.3	68.3
AutoEncoder	**99.1**	17.5	45.5	74.9	71.9	86.4	87.0	83.5	73.4	71.7
NMT En-Fr	80.1	58.3	51.7	81.9	73.7	89.5	90.3	89.1	73.2	75.4
NMT En-De	79.9	56.0	52.3	82.2	72.1	90.5	90.9	89.5	73.4	**76.2**
NMT En-Fi	78.5	58.3	50.9	82.5	71.7	90.0	90.3	88.0	73.2	75.4
Seq2Tree	93.3	10.3	**63.8**	**89.6**	**82.1**	**90.9**	**95.1**	**95.1**	73.2	71.9
SkipThought	66.0	35.7	44.6	72.5	73.8	90.3	85.0	80.6	**73.6**	71.0
NLI	71.7	87.3	41.6	70.5	65.1	86.7	80.7	80.3	62.1	66.8

Predicting acceptability

- But for semantics (and NLI) we are interested in more than linguistic properties

Predicting acceptability

- But for semantics (and NLI) we are interested in more than linguistic properties
- In particular, we hope that a model gives us representations that allow the model to predict how *acceptable* (or reasonable) humans consider sentences.

Predicting acceptability

Language Modeling with Syntactic and Semantic Representation for Sentence Acceptability Predictions

Adam Ek **Jean-Phillipe Bernardy** **Shalom Lappin**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek, jean-philippe.bernardy, shalom.lappin}@gu.se

Predicting acceptability

Language Modeling with Syntactic and Semantic Representation for Sentence Acceptability Predictions

Adam Ek **Jean-Phillipe Bernardy** **Shalom Lappin**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek, jean-philippe.bernardy, shalom.lappin}@gu.se

- We investigate if the probabilities a LM assign correlate with human judgments
  $SLOR_M = \frac{log(P_M(s)) - log(P_U(s))}{len(s)}$

Can we really predict acceptability

- In particular, we investigate whether semantic or syntactic information helps us:

Can we really predict acceptability

- In particular, we investigate whether semantic or syntactic information helps us:

Table 2: Weighted Pearson correlation between prediction from different models on the SMOG1 dataset. * indicates that the tags have been shuffled.

	HUMAN	LSTM	+SYN	+SYN*	+SEM	+SEM*	+DEPTH	+DEPTH*
HUMAN	1.00							
LSTM	**0.58**	1.00						
+SYN	0.55	0.96	1.00					
+SYN*	0.39	0.76	0.75	1.00				
+SEM	0.54	0.81	0.78	0.61	1.00			
+SEM*	0.52	0.81	0.78	0.63	0.96	1.00		
+DEPTH	0.56	0.97	0.97	0.74	0.79	0.79	1.00	
+DEPTH*	0.46	0.87	0.85	0.73	0.72	0.72	0.86	1.00

Natural Language Inference for Neural Networks

- A problem with neural networks is that we need **a lot** of data for systems to work well

Natural Language Inference for Neural Networks

- A problem with neural networks is that we need **a lot** of data for systems to work well
- Previous NLI datasets were small with carefully selected examples (FraCas), which made them unfit for neural networks (but excellent for evaluation)

Natural Language Inference for Neural Networks

- In 2015, we got SNLI with 550k examples and in 2018 MNLI with 440k examples

A large annotated corpus for learning natural language inference

Samuel R. Bowman[*†]
sbowman@stanford.edu

Gabor Angeli[†‡]
angeli@stanford.edu

Christopher Potts[*]
cgpotts@stanford.edu

Christopher D. Manning[*†‡]
manning@stanford.edu

[*]Stanford Linguistics [†]Stanford NLP Group [‡]Stanford Computer Science

Natural Language Inference for Neural Networks

- In 2015, we got SNLI with 550k examples and in 2018 MNLI with 440k examples

A large annotated corpus for learning natural language inference

Samuel R. Bowman[*,†]
sbowman@stanford.edu

Gabor Angeli[†,‡]
angeli@stanford.edu

Christopher Potts[*]
cgpotts@stanford.edu

Christopher D. Manning[*,†,‡]
manning@stanford.edu

[*]Stanford Linguistics [†]Stanford NLP Group [‡]Stanford Computer Science

- and in 2018 MNLI with 440k examples

A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

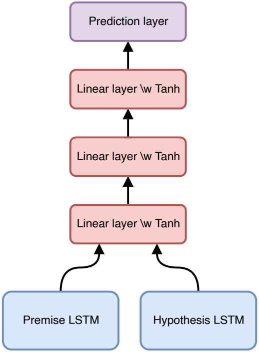
Adina Williams[1]
adinawilliams@nyu.edu

Nikita Nangia[2]
nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

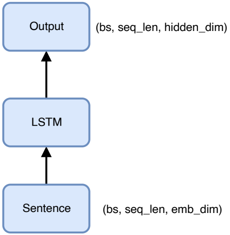
Neural Networks for Natural Language Inference

- A general architecture for NLI problems was first proposed in the SNLI paper:



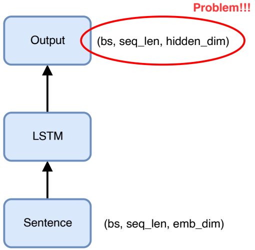
But what is a sentence representation...

- If we use a LSTM to encode a sentence, we get one representation for each token:



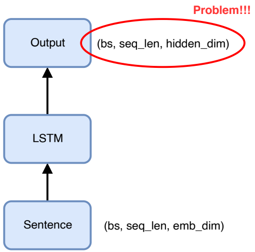
But what is a sentence representation...

- If we use a LSTM to encode a sentence, we get one representation for each token:



But what is a sentence representation...

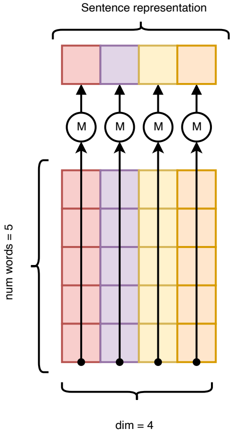
- If we use a LSTM to encode a sentence, we get one representation for each token:



- to predict a class we need *one* embedding...

Compressing a sequence of token representations

- To solve this issue, we can use some form of *pooling*:



- Max
- Sum
- Mean
- ...

Neural Networks for Natural Language Inference (2)

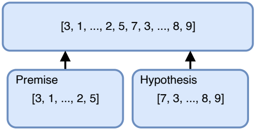
- In general, a neural network will produce one representation of the premise and one of the hypothesis.

Neural Networks for Natural Language Inference (2)

- In general, a neural network will produce one representation of the premise and one of the hypothesis.
- But to predict in a neural net, we need one representation! So we have to combine them somehow.

Neural Networks for Natural Language Inference (2)

- In general, a neural network will produce one representation of the premise and one of the hypothesis.
- But to predict in a neural net, we need one representation! So we have to combine them somehow.
- The model presented previously simply *concatenate* the premise and hypothesis representations



Universal Sentence Representations from NLI

Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Alexis Conneau
Facebook AI Research
aconneau@fb.com

Douwe Kiela
Facebook AI Research
dkiela@fb.com

Holger Schwenk
Facebook AI Research
schwenk@fb.com

Loïc Barrault
LIUM, Université Le Mans
loic.barrault@univ-lemans.fr

Antoine Bordes
Facebook AI Research
abordes@fb.com

Universal Sentence Representations from NLI

Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Alexis Conneau Facebook AI Research aconneau@fb.com

Douwe Kiela Facebook AI Research dkiela@fb.com

Holger Schwenk Facebook AI Research schwenk@fb.com

Loïc Barrault LIUM, Université Le Mans loic.barrault@univ-lemans.fr

Antoine Bordes Facebook AI Research abordes@fb.com

Model	**dim**	**NLI dev**	**NLI test**
LSTM	2048	81.9	80.7
GRU	4096	82.4	81.8
BiGRU-last	4096	81.3	80.9
BiLSTM-Mean	4096	79.0	78.2
Inner-attention	4096	82.3	82.5
HConvNet	4096	83.7	83.4
BiLSTM-Max	4096	**85.0**	**84.5**

- Evaluate different model architectures on the same data
- This gives us a "estimation" of which architecture produce the best representations

A more advanced approach to NLI

- Instead of just concatenating H and P, we also consider the element-wise subtraction (vector contraction) and multiplication (vector scaling)

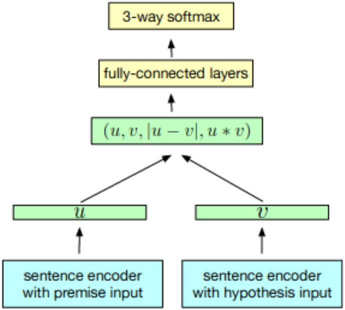
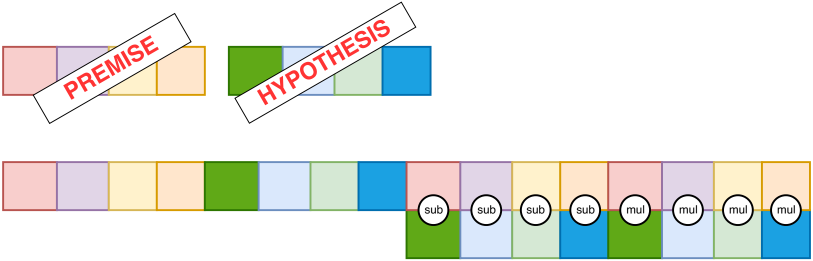
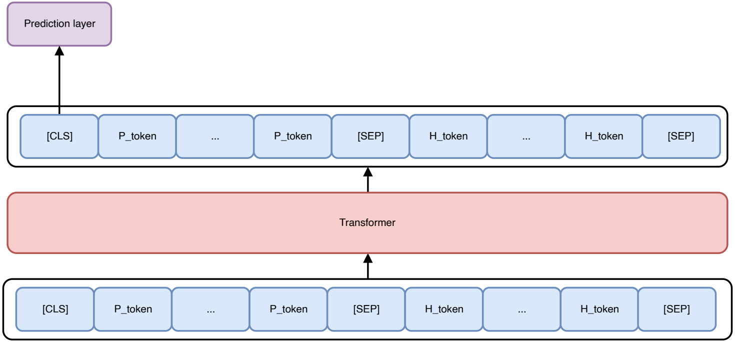


Figure 1: **Generic NLI training scheme.**

Combining sentences



Transformers for Natural Language Inference



- We take the representation of *both sentences*, the CLS-token and use it to predict the relationship between the premise and hypothesis.

Transformers for Natural Language Inference

- Why don't we combine the representation of the premise and the hypothesis in the transformer?

Transformers for Natural Language Inference

- Why don't we combine the representation of the premise and the hypothesis in the transformer?
- In the transformer we consider both sentences *jointly*, i.e. the token representations are already conditioned on each other.

Next up...

Meta-overview

- We train some neural network to predict some classes over a large dataset
- But semantic problems (or tasks) usually contain complex reasoning involving commonsense knowledge
- To annotate datasets of NLI (and other NLU tasks) we use human labor

The state of NLI

- We've seen fancy models of inference (BERT/LSTM)
- They appear to work well, getting over 80-90% accuracy on NLI datasets

The state of NLI

- We've seen fancy models of inference (BERT/LSTM)
- They appear to work well, getting over 80-90% accuracy on NLI datasets
- They appear to be great! We've solved NLI!

The state of NLI

- We've seen fancy models of inference (BERT/LSTM)
- They appear to work well, getting over 80-90% accuracy on NLI datasets
- They appear to be great! We've solved NLI!
- But people started to look at the predictions and noticed, we can't explain why the system does x, y or z

The state of NLI

- We've seen fancy models of inference (BERT/LSTM)
- They appear to work well, getting over 80-90% accuracy on NLI datasets
- They appear to be great! We've solved NLI!
- But people started to look at the predictions and noticed, we can't explain why the system does x, y or z
- and people noted that "easy" examples were not solved by NLI systems, but they appear to solve more "difficult" examples

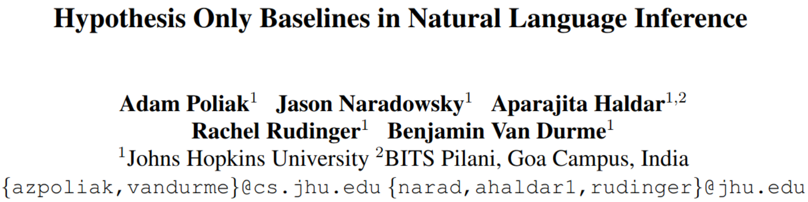
Hypothesis-only baseline

Hypothesis Only Baselines in Natural Language Inference

Adam Poliak[1] **Jason Naradowsky**[1] **Aparajita Haldar**[1,2]
Rachel Rudinger[1] **Benjamin Van Durme**[1]
[1]Johns Hopkins University [2]BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

- What happens if we only consider the hypothesis?

Hypothesis-only baseline


Hypothesis Only Baselines in Natural Language Inference

Adam Poliak[1] Jason Naradowsky[1] Aparajita Haldar[1,2]
Rachel Rudinger[1] Benjamin Van Durme[1]
[1]Johns Hopkins University [2]BITS Pilani, Goa Campus, India
{azpoliak,vandurme}@cs.jhu.edu {narad,ahaldar1,rudinger}@jhu.edu

- What happens if we only consider the hypothesis?

P None

H The cats are playing

Label Contradiction

Does considering just the hypothesis work?

	DEV				TEST					
Dataset	Hyp-Only	MAJ	$\|\Delta\|$	$\Delta\%$	Hyp-Only	MAJ	$\|\Delta\|$	$\Delta\%$	Baseline	SOTA
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

Does considering just the hypothesis work?

	DEV				TEST					
Dataset	Hyp-Only	MAJ	$\lvert\Delta\rvert$	$\Delta\%$	Hyp-Only	MAJ	$\lvert\Delta\rvert$	$\Delta\%$	Baseline	SOTA
SNLI	69.17	33.82	+35.35	+104.52	69.00	34.28	+34.72	+101.28	78.2	89.3
MNLI-1	55.52	35.45	+20.07	+56.61	–	35.6	– –		72.3	80.60
MNLI-2	55.18	35.22	+19.96	+56.67	–	36.5	–	–	72.1	83.21

- The results show that it's possible to predict the relation based only on the hypothesis. But how does this make sense? (it shouldn't work at all)
- The hypothesis contain implicit signals that can be used to predict its class

Sensitivity to word-order

- The success of the transformer models for NLU has been attributed to it's ability to model complex *syntactic and semantic dependencies* between words

Sensitivity to word-order

- The success of the transformer models for NLU has been attributed to it's ability to model complex *syntactic and semantic dependencies* between words
- We can test this by pre-training a transformer on permuted data, and test on several downstream tasks

Sensitivity to word-order

- The success of the transformer models for NLU has been attributed to it's ability to model complex *syntactic and semantic dependencies* between words
- We can test this by pre-training a transformer on permuted data, and test on several downstream tasks

Masked Language Modeling and the Distributional Hypothesis:
Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]

Adina Williams[†] **Douwe Kiela**[†]

[†] Facebook AI Research; [‡] McGill University / Montreal Institute of Learning Algorithms
{koustuvs,adinawilliams,dkiela}@fb.com

Some results

- Setup: train RoBERTa (Liu et al. 2019) on a dataset containing sentence-permutations: permute n-grams of size 1, 2, 3, and 4

Some results

- Setup: train RoBERTa (Liu et al. 2019) on a dataset containing sentence-permutations: permute *n*-grams of size 1, 2, 3, and 4

  The cat is super tall and fancy $\rightarrow$ The cat (super tall) is fancy

Some results

- Setup: train RoBERTa (Liu et al. 2019) on a dataset containing sentence-permutations: permute *n*-grams of size 1, 2, 3, and 4

  The cat is super tall and fancy $\rightarrow$ The cat (super tall) is fancy
- Evaluate on the GLUE benchmark

GLUE benchmark

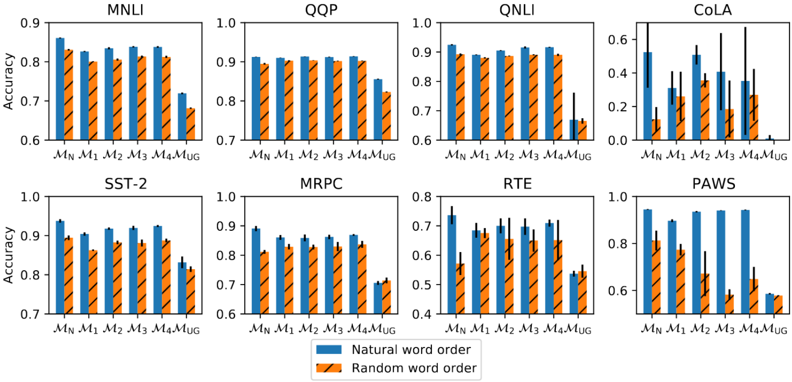
- GLUE is a collection of datasets used to measure "success" in a variety of NLU tasks (including NLI)

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

Some results

- Setup: train RoBERTa (Liu et al. 2019) on a dataset containing sentence-permutations: permute *n*-grams of size 1, 2, 3, and 4
  The cat is super tall and fancy → The cat (super tall) is fancy
- Evaluate on the GLUE benchmark



Wut!?

- So, pre-training on permuted data is possible and yield good results on downstream tasks, but how is this possible?
- For example in English, word order gives *meaning* rather than morphology (like Turkish)

Wut!?

- So, pre-training on permuted data is possible and yield good results on downstream tasks, but how is this possible?
- For example in English, word order gives *meaning* rather than morphology (like Turkish)
- The authors show that models use higher-order *distributional statistics* to construct *meaning*

Wut!?

- So, pre-training on permuted data is possible and yield good results on downstream tasks, but how is this possible?
- For example in English, word order gives *meaning* rather than morphology (like Turkish)
- The authors show that models use higher-order *distributional statistics* to construct *meaning*
- Consequently, it appears that transformers don't really consider the classical NLP pipeline (using human syntactic and semantic mechanisms)

Wut!?

- So, pre-training on permuted data is possible and yield good results on downstream tasks, but how is this possible?
- For example in English, word order gives *meaning* rather than morphology (like Turkish)
- The authors show that models use higher-order *distributional statistics* to construct *meaning*
- Consequently, it appears that transformers don't really consider the classical NLP pipeline (using human syntactic and semantic mechanisms)
- We need better evaluation datasets, that can't be solved by distributional statistics.

Let's replace some words!

- What happens if we remove entire word-classes from the dataset?

Let's replace some words!

- What happens if we remove entire word-classes from the dataset?

NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance

Aarne Talman[*†], Marianna Apidianaki[*], Stergios Chatzikyriakidis[‡], Jörg Tiedemann[*]

[*]Department of Digital Humanities, University of Helsinki
{name.surname}@helsinki.fi
[†]Basement AI
[‡]CLASP, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
{name.surname}@gu.se

Let's replace some words!

- What happens if we remove entire word-classes from the dataset?

NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance

Aarne Talman[*†], Marianna Apidianaki[*], Stergios Chatzikyriakidis[‡], Jörg Tiedemann[*]

[*]Department of Digital Humanities, University of Helsinki
{name.surname}@helsinki.fi
[†]Basement AI
[‡]CLASP, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
{name.surname}@gu.se

- Setup: Train a BERT model on a corrupted version of the dataset and test it

Data corruption

	Premise	Hypothesis
Contradiction	*He was hardly more than five ~~feet~~, four ~~inches~~, but carried himself with great ~~dignity~~.*	*The ~~man~~ was 6 ~~foot~~ ~~tall~~.*
Entailment	*Two ~~plants~~ died on the long ~~journey~~ and the third one found its way to ~~Jamaica~~ exactly how is still shrouded in ~~mystery~~.*	*The third ~~plant~~ was a different ~~type~~ from the first two.*
Neutral	*In a ~~couple~~ of ~~days~~ the ~~wagon~~ ~~train~~ would head on north to ~~Tucson~~, but now the ~~activity~~ in the ~~plaza~~ was a ~~mixture~~ of ~~market~~ ~~day~~ and ~~fiesta~~.*	*They were ~~south~~ of ~~Tucson~~.*

Table 1: Sentence pairs from a corrupted MNLI training dataset where nouns have been removed.

What happens :(

Data	CORRUPT-TRAIN	Δ	CORRUPT-TEST	Δ	CORRUPT-TRAIN AND TEST	Δ
MNLI-NUM	82.37%	-1.37	81.71%	-2.03	81.87%	-1.87
MNLI-CONJ	83.09%	-0.65	82.75%	-0.99	83.10%	-0.64
MNLI-ADV	80.21%	-3.53	72.41%	-11.33	75.69%	-8.05
MNLI-PRON	83.27%	-0.47	81.98%	-1.75	82.65%	-1.09
MNLI-ADJ	81.67%	-2.07	74.61%	-9.13	76.44%	-7.30
MNLI-DET	83.15%	-0.59	79.29%	-4.44	81.32%	-2.42
MNLI-VERB	81.40%	-2.34	73.96%	-9.78	76.30%	-7.44
MNLI-NOUN	80.72%	-3.02	69.80%	-13.94	73.38%	-10.35
MNLI-NOUN-PRON	79.74%	-4.00	68.41%	-15.33	72.14%	-11.60
NOUN+PRON+VERB	72.55%	-11.19	54.59%	-29.15	62.18%	-21.56
NOUN+ADV+VERB	67.58%	-16.16	62.58%	-21.16	67.58%	-16.16
NOUN+VERB	71.14%	-12.60	52.90%	-30.84	61.31%	-22.43
NOUN+VERB+ADJ	75.54%	-8.20	61.90%	-21.84	68.20%	-15.54
NOUN+VERB+ADV+ADJ	79.81%	-3.93	71.81%	-11.93	76.29%	-7.45

Table 2: Prediction accuracy (%) for the BERT-`base` model fine-tuned on CORRUPT-TRAIN and tested on the original MNLI-matched evaluation (dev) set (columns 2 and 3); fine-tuned on the original MNLI data and tested on CORRUPT-TEST; fine-tuned on CORRUPT-TRAIN and tested on CORRUPT-TEST (columns 6 and 7). The delta shows the difference in accuracy compared to the model fine-tuned on the original MNLI training set and evaluated on the MNLI-matched development set (83.74%).

The case of punctuation

How does Punctuation Affect Neural Models in Natural Language Inference

Adam Ek **Jean-Philippe Bernardy** **Stergios Chatzikyriakidis**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek,jean-philippe.bernardy,stergios.chatzikyriakidis}@gu.se

- Is BERT and LSTM models for NLI sensitive to punctuation?

The case of punctuation

How does Punctuation Affect Neural Models in Natural Language Inference

Adam Ek **Jean-Philippe Bernardy** **Stergios Chatzikyriakidis**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek,jean-philippe.bernardy,stergios.chatzikyriakidis}@gu.se

- Is BERT and LSTM models for NLI sensitive to punctuation?
- LSTMs are very sensitive to any punctuations
- BERT doesn't care at all about punctuation

The case of punctuation

How does Punctuation Affect Neural Models in Natural Language Inference

Adam Ek **Jean-Philippe Bernardy** **Stergios Chatzikyriakidis**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{adam.ek,jean-philippe.bernardy,stergios.chatzikyriakidis}@gu.se

- Is BERT and LSTM models for NLI sensitive to punctuation?
- LSTMs are very sensitive to any punctuations
- BERT doesn't care at all about punctuation

MODEL	TEST	MA	MM
BiLSTM_{orig}		.724	.723
BiLSTM_p	p	.723	.724
BiLSTM_p	$\neg p$	.428	.414
$\text{BiLSTM}_{\neg p}$	$\neg p$	.714	.727
$\text{BiLSTM}_{\neg p}$	p	.424	.430
HBMP_{orig}		.729	.733
HBMP_p	p	.728	.729
HBMP_p	$\neg p$	.430	.408
$\text{HBMP}_{\neg p}$	$\neg p$	.729	.732
$\text{HBMP}_{\neg p}$	p	.436	.427
BERT_{orig}		.833	.839
BERT_p	p	.835	.837
BERT_p	$\neg p$	.816	.822
$\text{BERT}_{\neg p}$	$\neg p$	.819	.820
$\text{BERT}_{\neg p}$	p	.830	.833

Taking a step back

- There are many ways to break NLI datasets and cast doubt on the performance of various "good" models

Taking a step back

- There are many ways to break NLI datasets and cast doubt on the performance of various "good" models
- The issue arise (mainly) from *poorly* constructed datasets (SNLI/MNLI)

Taking a step back

- There are many ways to break NLI datasets and cast doubt on the performance of various "good" models
- The issue arise (mainly) from *poorly* constructed datasets (SNLI/MNLI)
- These datasets are constructed from *crowdsourcing*

Taking a step back

- There are many ways to break NLI datasets and cast doubt on the performance of various "good" models
- The issue arise (mainly) from *poorly* constructed datasets (SNLI/MNLI)
- These datasets are constructed from *crowdsourcing*
- Annotators have biases and use shortcuts when annotating (such as "give-away")

Taking a step back

- There are many ways to break NLI datasets and cast doubt on the performance of various "good" models
- The issue arise (mainly) from *poorly* constructed datasets (SNLI/MNLI)
- These datasets are constructed from *crowdsourcing*
- Annotators have biases and use shortcuts when annotating (such as "give-away")
- Models exploit this!

Adversarial datasets

- One solution proposed to this problem is *adversarial* datasets
  E.g. ANLI
  (`https://github.com/facebookresearch/anli`)
- Humans generate a dataset with examples that fool the model

Adversarial datasets

- One solution proposed to this problem is *adversarial* datasets E.g. ANLI (`https://github.com/facebookresearch/anli`)
- Humans generate a dataset with examples that fool the model
- we then train and evaluate on this and construct new models that "solve" these adversarial examples

Is adversarial datasets the solution?

What Will it Take to Fix Benchmarking in Natural Language Understanding?

Samuel R. Bowman
New York University
bowman@nyu.edu

George E. Dahl
Google Research, Brain Team
gdahl@google.com

- Quote: *Evaluation for many natural language understanding (NLU) tasks is broken*

Taking a step back

- Large-scale evaluation and models trained on huge dataset deviate from classical linguistics

Taking a step back

- Large-scale evaluation and models trained on huge dataset deviate from classical linguistics
  - Solutions tried so far: adversarial (ANLI) and out-of-domain test sets (MNLI)
  - but these methods inevitably obscure the models abilities

Taking a step back

- Large-scale evaluation and models trained on huge dataset deviate from classical linguistics
  - Solutions tried so far: adversarial (ANLI) and out-of-domain test sets (MNLI)
  - but these methods inevitably obscure the models abilities
- Where do we go from here??? That's currently a work-in-progress :)

  (i.e. a great time to get into NLP and NLU)