HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

State-of-the-Art Natural Language Inference Systems Fail to Capture the Semantics of Inference

Aarne Talman

October 25, 2018

Department of Digital Humanities
University of Helsinki

FOTRAN Found in Translation
ACADEMY OF FINLAND

Outline

1. Introduction
2. We Have Solved the Problem of NLI with Neural Networks!
3. No We Haven't!
4. Now What?

Outline

1 Introduction

2 We Have Solved the Problem of NLI with Neural Networks!

3 No We Haven't!

4 Now What?

Natural Language Inference

- Natural language inference (NLI) tries to model the inferential relationship between two or more given sentences.
- Given two sentences, the premise *p* and the hypothesis *h*, the task is to determine whether:
  1. *h* is *entailed* by *p*
  2. the sentences are in *contradiction* with each other
  3. there is no inferential relationship between the sentences (*neutral*).
- Example:

Premise:	*A couple walk hand in hand down a street.*
Hypothesis:	*A couple is walking together.*
Label:	**entailment**

NLI Datasets and Benchmark Tasks

SNLI: Stanford Natural Language Inference (Bowman et al., 2015)

- The first large-scale human-written manually labeled dataset for NLI.
- Contains 550,152 training pairs, 10,000 development pairs 10,000 test pairs.
- Source: image captions taken from the Flickr30k corpus

MultiNLI: Multi-Genre Natural Language Inference (Williams et al., 2018)

- The same data collection method and definition of inference as SNLI.
- Contains 392,702 training pairs, 20,000 development pairs 20,000 test pairs.
- Sentence pairs drawn from ten distinct genres of written and spoken English.

SICK: Sentences Involving Compositional Knowledge (Marelli et al., 2014)

- Premises drawn from 8K ImageFlickr and STS MSRVideo Description datasets.
- Hypotheses automatically generated.
- Contains 9,840 labeled sentence pairs.
- Contains examples pertaining to logical inference (negation, conjunction, disjunction, relative clauses, etc.)

Other notable datasets include:

- FraCas (Cooper et al., 1996), RTE (Dagan et al., 2006), SciTail (Khot et al., 2018), XNLI (Conneau et al., 2018)

NLI Datasets and Benchmark Tasks

	Example sentence pairs (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.*

Table 1: Example sentence pairs from the three datasets.

Outline

1 Introduction

2 We Have Solved the Problem of NLI with Neural Networks!

3 No We Haven't!

4 Now What?

Neural Network Architectures for NLI

Sentence encoding models

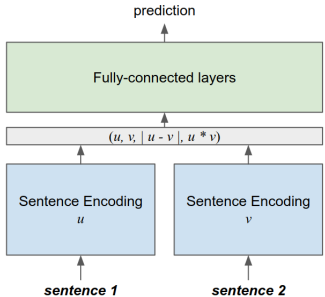

Figure 1: Sentence encoding architecture for NLI based on Bowman et al. (2015)

Cross-sentence attention models

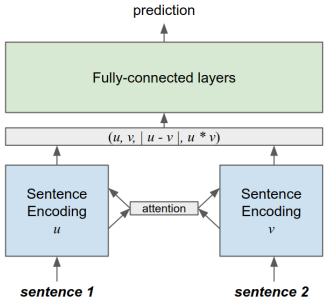

Figure 2: Simplified cross-sentence attention model for NLI

An Example Sentence Encoding Model

Hierarchical BiLSTM Max Pooling Architecture (HBMP) (Talman et al., 2018)

- Motivated by the good results with one-layer bidirectional LSTM max pooling encoder (InferSent) by Conneau et al. (2017).
- Main idea: allow all BiLSTM layers to re-read the input sentences, while preserving information from the previous layers.
- Our hypothesis is that each layer learns additional semantic information not present on the previous layer.
- Holds the current top score in SciTail NLI benchmark by AllenAI (Khot et al., 2018).

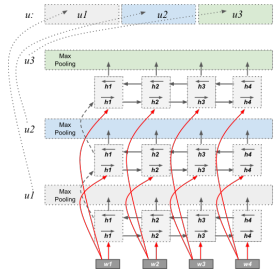

Figure 3: HBMP architecture for sentence encodings (Talman et al., 2018)

An Example Cross-Sentence Attention Model

Densely-Connected Recurrent and Co-Attentive Network (DRCN) (Kim et al., 2018)

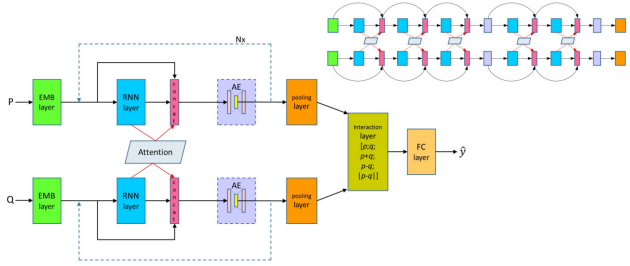

Figure 1: General architecture of our Densely-connected Recurrent and Co-attentive neural Network (DRCN). Dashed arrows indicate that a group of RNN-layer, concatenation and AE can be repeated multiple (N) times (like a repeat mark in a music score). The bottleneck component denoted as AE, inserted to prevent the ever-growing size of a feature vector, is optional for each repetition. The upper right diagram is our specific architecture for experiments with 5 RNN layers ($N = 4$).

Figure 4: Densely-Connected Recurrent and Co-Attentive Network architecture (Kim et al., 2018)

The Current State-of-the-Art Model for SNLI Achieves 90% Accuracy!

Sentence Encoding Models

Model	Accuracy
BiLSTM Max Pool (InferSent) (Conneau et al., 2017)	84.5
Distance-based Self-Attention (Im and Cho, 2017)	86.3
ReSA Shen et al. (2018)	86.3
600D BiLSTM with gen pooling (Chen et al., 2018)	86.6
600D Dynamic Self-Attention Model (Yoon et al., 2018)	86.8
2400D Multiple-Dynamic Self-Attention Model (Yoon et al., 2018)	**87.4**
Our HBMP (Talman et al., 2018)	86.6

Table 2: Sentence Encoding Model test accuracies (%).

Cross-Sentence Attention Models

Model	Accuracy
KIM Ensemble (Chen et al., 2017)	89.1
450D DR-BiLSTM Ensemble (Ghaeini et al., 2018)	89.3
300D CAFE Ensemble (Tay et al., 2017)	89.3
150D Multiway Attn Network Ensemble (Tan et al., 2018)	89.4
300D DMAN Ensemble (Pan et al., 2018)	89.6
Fine-Tuned LM-Pretrained Transformer (Radford et al., 2018)	89.9
DRCN Ensemble (Kim et al., 2018)	**90.1**

Table 3: Cross-Sentence Attention Model test accuracies (%).

Bottom line: Neural network models for NLI have become hugely successful!!

Outline

1 Introduction

2 We Have Solved the Problem of NLI with Neural Networks!

3 No We Haven't!

4 Now What?

Neural Network Models Fail to Capture Lexical Semantics

- Breaking NLI (Glockner et al., 2018): a test set of 8,193 sentence pairs constructed to highlight how poorly current neural network models for NLI can handle lexical meaning.
- Constructed by taking premises from the SNLI training set, creating several hypotheses from them by changing at most one word.
- Lexical items changed in the dataset include e.g. colors, instruments, ordinals, drinks, cardinals, rooms, vegetables, etc.

Premise:	*Several women stand on a platform near the **yellow** line.*
Hypothesis:	*Several women stand on a platform near the **green** line.*
Label:	**contradiction**

Table 4: Example sentence pair from the Breaking NLI dataset.

Neural Network Models Fail to Capture Lexical Semantics

	Baseline	Cross-Sentence Attention			Sentence Encoding	
Category	**WordNet** *	**Decomposable Attention***	**ESIM***	**KIM***	**InferSent****	**600D HBMP****
antonyms	95.5	41.6	70.4	86.5	51.6	54.7
antonyms(wordnet)	94.5	55.1	74.6	78.8	63.7	69.1
cardinals	98.6	53.5	75.5	93.4	49.4	58.8
colors	98.7	85.0	96.1	98.3	90.6	90.4
countries	100.0	15.2	25.4	70.8	77.2	81.2
drinks	94.8	52.9	63.7	96.6	85.1	81.3
instruments	67.7	96.9	90.8	96.9	98.5	96.9
materials	75.3	65.2	89.7	98.7	81.6	82.6
nationalities	78.5	37.5	35.9	73.5	47.3	49.8
ordinals	40.7	2.1	21.0	56.6	7.4	4.5
planets	100.0	31.7	3.3	5.0	75.0	45.0
rooms	89.9	59.2	69.4	77.6	76.3	72.1
synonyms	70.5	97.5	99.7	92.1	99.6	84.5
vegetables	86.2	43.1	31.2	79.8	39.5	40.4
Total	85.8	51.9	65.6	83.5	65.6	65.1

Table 5: Breaking NLI scores (accuracy %). Results marked with * as reported by Glockner et al. (2018) and ** by Talman et al. (2018). Scores highlighted with bold are top scores when comparing the InferSent and our HBMP model.

Neural Network Models Learn Annotation Artifacts

- Gururangan et al. (2018) show that datasets like SNLI and MultiNLI contain unintentional annotation artifacts which help NLI models in classification.
- A simple text categorization model (BoW + bigram) can correctly predict the gold label using the **hypotheses alone** in about 67% of SNLI and 53% of MultiNLI (majority class being 34% and 35% respectively).
- Examples:
  - Entailments often contain generalizations (e.g. *dog* $\rightarrow$ *animal*).
  - Neutrals contain modifiers (e.g. *tall*, *sad*) and superlatives (e.g. *first*, *most*).
  - Contradictions often contain negations. The word *cat* also appears frequently in contradictions.
- Conclusion: annotation artifacts inflate model performance in NLI tasks.

Neural Network Models for NLI Fail in Transfer Learning Between NLI Tasks

- Our recent experiments show that the success of neural network models for NLI is largely task specific (Talman and Chatzikyriakidis, 2018).
- We trained four state-of-the-art NLI models on SNLI, MultiNLI and SNLI+MultiNLI training data, and tested them on test data drawn from a different corpus.

Train	Dev	Test
SNLI	**SNLI**	**SNLI**
SNLI	SNLI	MultiNLI
SNLI	SNLI	SICK
MultiNLI	**MultiNLI**	**MultiNLI**
MultiNLI	MultiNLI	SNLI
MultiNLI	MultiNLI	SICK
SNLI+MultiNLI	**SNLI**	**SNLI**
SNLI+MultiNLI	SNLI	SICK

Table 6: List of all the combinations of data used in the experiments.

Neural Network Models for NLI Fail in Transfer Learning Between NLI Tasks

- The drop in accuracy is biggest when training with SNLI, possibly due to the simplicity of sentences in the dataset.
- However, remember how similar the SICK and SNLI examples were?

Train	Dev	Test	Test Accuracy	Δ	Model
SNLI	**SNLI**	**SNLI**	**86.1**		**600D BiLSTM-max**
SNLI	**SNLI**	**SNLI**	**86.6**		**600D HBMP (Talman et al., 2018)**
SNLI	**SNLI**	**SNLI**	**88.0**		**600D ESIM (Chen et al., 2017)**
SNLI	**SNLI**	**SNLI**	**88.6**		**300D KIM (Chen et al., 2018)**
SNLI	SNLI	MultiNLI-m	55.7*	-30.4	600D BiLSTM-max
SNLI	SNLI	MultiNLI-m	56.3*	-30.3	600D HBMP
SNLI	SNLI	MultiNLI-m	59.2*	-28.8	600D ESIM
SNLI	SNLI	MultiNLI-m	61.7*	-26.9	300D KIM
SNLI	SNLI	SICK	54.5	-31.6	600D BiLSTM-max
SNLI	SNLI	SICK	53.1	-33.5	600D HBMP
SNLI	SNLI	SICK	54.3	-33.7	600D ESIM
SNLI	SNLI	SICK	55.8	-32.8	300D KIM

Table 7: Test accuracies (%). For the baseline results highlighted in bold the training data includes examples from the same corpus as the test data. For the other models the training and test data are taken from separate corpora.

Neural Network Models for NLI Fail in Transfer Learning Between NLI Tasks

- The drop in accuracy is smallest when trained on MultiNLI and tested on SNLI.
- However, the drop is unexpectedly big given the same definition of inference and the same data collection method in SNLI and MultiNLI.

Train	Dev	Test	Test Accuracy	Δ	Model
MultiNLI	**MultiNLI-m**	**MultiNLI-m**	**73.1***		**600D BiLSTM-max**
MultiNLI	**MultiNLI-m**	**MultiNLI-m**	**73.2***		**600D HBMP**
MultiNLI	**MultiNLI-m**	**MultiNLI-m**	**76.8***		**600D ESIM**
MultiNLI	**MultiNLI-m**	**MultiNLI-m**	**77.3***		**300D KIM**
MultiNLI	MultiNLI-m	SNLI	63.8	-9.3	600D BiLSTM-max
MultiNLI	MultiNLI-m	SNLI	65.3	-7.9	600D HBMP
MultiNLI	MultiNLI-m	SNLI	66.4	-10.4	600D ESIM
MultiNLI	MultiNLI-m	SNLI	68.5	-8.8	300D KIM
MultiNLI	MultiNLI-m	SICK	54.1	-19.0	600D BiLSTM-max
MultiNLI	MultiNLI-m	SICK	54.1	-19.1	600D HBMP
MultiNLI	MultiNLI-m	SICK	47.9	-28.9	600D ESIM
MultiNLI	MultiNLI-m	SICK	50.9	-26.4	300D KIM

Table 8: Test accuracies (%). For the baseline results highlighted in bold the training data includes examples from the same corpus as the test data. For the other models the training and test data are taken from separate corpora.

Neural Network Models for NLI Fail in Transfer Learning Between NLI Tasks

- Combining SNLI and MultiNLI training data doesn't help when testing with SICK.

Train	Dev	Test	Test Accuracy	Δ	Model
SNLI+MultiNLI	**SNLI**	**SNLI**	**86.1**		**600D BiLSTM-max**
SNLI+MultiNLI	**SNLI**	**SNLI**	**86.1**		**600D HBMP**
SNLI+MultiNLI	**SNLI**	**SNLI**	**87.5**		**600D ESIM**
SNLI+MultiNLI	**SNLI**	**SNLI**	**86.2**		**300D KIM**
SNLI+MultiNLI	SNLI	SICK	54.5	-31.6	600D BiLSTM-max
SNLI+MultiNLI	SNLI	SICK	55.0	-31.1	600D HBMP
SNLI+MultiNLI	SNLI	SICK	54.5	-33.0	600D ESIM
SNLI+MultiNLI	SNLI	SICK	54.6	-31.6	300D KIM

Table 9: Test accuracies (%). For the baseline results highlighted in bold the training data includes examples from the same corpus as the test data. For the other models the training and test data are taken from separate corpora.

Getting 90% accuracy on a benchmark task is not enough if you're not able to apply the trained model outside of that benchmark!

Outline

1 Introduction

2 We Have Solved the Problem of NLI with Neural Networks!

3 No We Haven't!

4 Now What?

Conclusions and Future Research

Current state-of-the-art neural network models for NLI are not able to capture the semantics of NLI:

- They can be broken by small changes in lexical meaning.
- They learn annotation artifacts in the data.
- They fail in transfer learning between NLI tasks.

Tasks for the future:

- Better datasets with more diverse notion of inference, see (Chatzikyriakidis et al., 2017) for dicussion.
- Better NLI models that generalize across datasets.
  - Huge pretrained language models, e.g. AllenAI's ELMo (Peters et al., 2018), OpenAI's Finetuned Transformer (Radford et al., 2018) and Google's BERT (Devlin et al., 2018)?
  - What about multilingual NLI and multilingual models?

Thank You!

FOTRAN Found in Translation

ACADEMY OF FINLAND

References I

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chatzikyriakidis, S., Cooper, R., Dobnik, S., and Larsson, S. (2017). An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*.

Chen, Q., Ling, Z.-H., and Zhu, X. (2018). Enhancing Sentence Embedding with Generalized Pooling. *arXiv preprint arXiv:1806.09828*.

Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., and Wei, S. (2018). Neural natural language inference models enhanced with external knowledge. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

References II

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

References III

Cooper, R., Crouch, D., Eijck, J. V., Fox, C., Genabith, J. V., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). Using the framework. In *Technical report LRE 62-051r*. The FraCaS consortium.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Ghaeini, R., Hasan, S. A., Datla, V., Liu, J., Lee, K., Qadir, A., Ling, Y., Prakash, A., Fern, X. Z., and Farri, O. (2018). Dr-bilstm: Dependent reading bidirectional lstm for natural language inference.

References IV

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (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)*, pages 107–112. Association for Computational Linguistics.

Im, J. and Cho, S. (2017). Distance-based self-attention network for natural language inference. *arXiv preprint arXiv:1712.02047*.

Khot, T., Sabharwal, A., and Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. In *AAAI*.

References V

Kim, S., Hong, J.-H., Kang, I., and Kwak, N. (2018). Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC2014*.

Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., and He, X. (2018). Discourse marker augmented network with reinforcement learning for natural language inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

References VI

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Shen, T., Zhou, T., Long, G., Jiang, J., Wang, S., and Zhang, C. (2018). Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*.

Talman, A. and Chatzikyriakidis, S. (2018). Neural network models for natural language inference fail to capture the semantics of inference. *arXiv preprint arXiv:1808.08762*.

Talman, A., Yli-Jyrä, A., and Tiedemann, J. (2018). Natural language inference with hierarchical bilstm max pooling architecture. *arXiv preprint arXiv:1808.08762*.

References VII

Tan, C., Wei, F., Wang, W., Lv, W., and Zhou, M. (2018). Multiway attention networks for modeling sentence pairs. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4411–4417. International Joint Conferences on Artificial Intelligence Organization.

Tay, Y., Tuan, L. A., and Hui, S. C. (2017). Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference.

Williams, A., Nangia, N., and Bowman, S. R. (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 (NAACL)*. Association for Computational Linguistics.

Yoon, D., Lee, D., and Lee, S. (2018). Dynamic Self-Attention : Computing Attention over Words Dynamically for Sentence Embedding. *arXiv preprint arXiv:1808.07383*.