HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Natural Language Inference with Hierarchical BiLSTM Architecture

Aarne Talman, Anssi Yli-Jyrä and Jörg Tiedemann

September 28, 2018

University of Helsinki

FoTran Found in Translation
ACADEMY OF FINLAND

Outline

Background of the Current Project

PhD Project: *Natural Language Inference with Multilingual Grounding*

Main phases of the project:

1. **Baseline system development and experiments** with different architectures
   - Aarne Talman, Anssi Yli-Jyrä and Jörg Tiedemann, *Natural Language Inference with Hierarchical BiLSTM Max Pooling Architecture* (Talman et al., 2018)
     - `https://arxiv.org/abs/1808.08762`
     - `https://github.com/Helsinki-NLP/HBMP`
2. Multilingual NLI and application of language independent meaning representations to NLI

Natural Language Inference

Natural Language Inference (NLI) is the problem of determining whether a natural language hypothesis can be inferred from a natural language premise.

- A simple example:
  - *p* *A group of people are standing on steps in front of a building.*
  - *h* *A group of people are standing in front of a building.*
- A typical NLI task involves classification of such hypothesis-premise pairs into entailments, contradictions or neutral.
- NLI is relatively easy for humans, but has turned out to be quite hard for computers – even when the data is presented in nicely organised sentence pairs.
- Some well known NLI tasks and datasets include Recognizing Textual Entailment (RTE), Stanford Natural Language Inference (SNLI), Multi-genre Natural Language Inference (MultiNLI), SciTail...

Sentence Encoding Based Architecture for NLI

- Our current models are based on the sentence encoding approach.
- Both the premise and hypothesis are encoded separately.
- Encoded sentences are passed to a multilayer perceptron classifier.

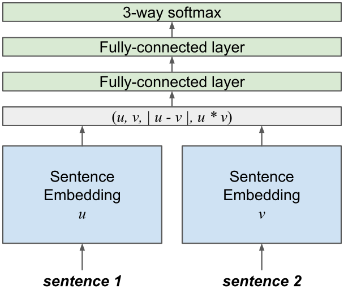

Figure 1: Sentence encoding architecture for NLI based on Bowman et al. (2015)

FOTRAN Found in Translation

ACADEMY OF FINLAND

Hierarchical BiLSTM Max Pooling Architecture (HBMP)

- The architecture is motivated by the good results with simple BiLSTM Max Pooling encoder (InferSent) by Conneau et al. (2017).
- The idea behind the HBMP architecture is to allow all BiLSTM layers to re-read the input sentences, while preserving the hiddent and cell states from the previous layer.
- Our hypothesis is that each layer learns additional semantic information not present on the previous layer.

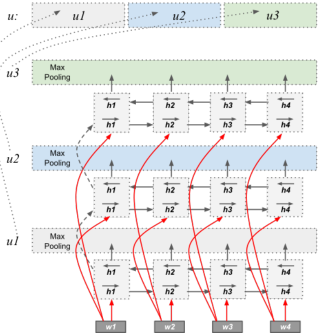

Figure 2: HBMP architecture for sentence encodings (Talman et al., 2018)

Experimental Results – NLI

SNLI (Bowman et al., 2015)

Model	**Accuracy**
BiLSTM Max Pool (InferSent)[a]	84.5
600D BiLSTM with generalized pooling[b]	86.6
600D Dynamic Self-Attention Model[c]	86.8
2400D Multiple-Dynamic Self-Attention Model[c]	**87.4**
Our HBMP	86.6

Table 1: SNLI test accuracies (%). Results marked with [a] by Conneau et al. (2017), [b] by Chen et al. (2018) and [c] by Yoon et al. (2018).

SciTail (Khot et al., 2018)

Model	**Accuracy**
DecompAtt[a]	72.3
ESIM[a]	70.6
Ngram[a]	70.6
DGEM w/o edges[a]	70.8
DGEM[a]	77.3
CAFE[b]	83.3
Our LSTM	67.3
Our BiLSTM max pooling	84.9
Our HBMP	**86.0**

Table 2: SciTail test accuracies (%). Results marked with [a] are baseline results reported by Khot et al. (2018) and [b] by Tay et al. (2018).

FoTRAN Found in Translation

ACADEMY OF FINLAND

Experimental Results – NLI

Model	Accuracy (MultiNLI-m)	Accuracy (MultiNLI-mm)
CBOW[a]	66.2	64.6
BiLSTM[a]	67.5	67.1
BiLSTM + enh embed + max pooling[b]	70.7	70.8
BiLSTM + Inner-attention[c]	72.1	72.1
Deep Gated Attn. BiLSTM encoders[d]	73.5	**73.6**
Shortcut-Stacked BiLSTM[e]	**74.5**	73.5
Our HBMP	73.7	73.0

Table 3: MultiNLI test accuracies (%). Results marked with [a] are baseline results by Williams et al. (2018), [b] by Vu (2017), [c] by Balazs et al. (2017), [d] by Chen et al. (2017) and [e] by Nie and Bansal (2017). Our results for the MultiNLI test sets were obtained by submitting the predictions to the respective Kaggle competitions.

Experimental Results – Transfer Learning with SentEval

SentEval downstream tasks (Conneau et al., 2017)

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS14
InferSent	81.1	86.3	92.4	90.2	**84.6**	88.2	76.2/83.1	**0.884**	**86.3**	.70/.67
SkipThought	79.4	83.1	**93.7**	89.3	82.9	88.4	-	0.858	79.5	.44/.45
Our 600D HBMP	81.5	86.4	92.7	89.8	83.6	86.4	74.6/82.0	0.876	85.3	.70/.66
Our 1200D HBMP	**81.7**	**87.0**	**93.7**	**90.3**	84.0	**88.8**	**76.7/83.4**	0.876	84.7	**.71/.68**

Table 4: Transfer learning test results for the HBMP model on a number of SentEval downstream sentence embedding evaluation tasks. InferSent and SkipThought results as reported by Conneau et al. (2017).

SentEval probing tasks (Conneau et al., 2018)

Model	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
InferSent	71.7	**87.3**	41.6	70.5	65.1	86.7	80.7	80.3	**62.1**	66.8
Our 600D HBMP	**75.9**	84.1	42.9	76.6	64.3	86.2	83.7	79.3	58.9	68.5
Our 1200D HBMP	75.0	85.3	**43.8**	**77.2**	**65.6**	**88.0**	**87.0**	**81.8**	59.0	**70.8**

Table 5: SentEval probing task results (accuracy %). InferSent results are BiLSTM Max (NLI) results as reported by Conneau et al. (2018).

SentEval website: `https://github.com/facebookresearch/SentEval`

FOTRAN Found in Translation

ACADEMY OF FINLAND

Latest Negative Results

Joint work with Stergios Chatzikyriakidis (CLASP)

NLI systems break down when training and testing on different datasets...

Train	Dev	Test	Test Accuracy	Model details
SNLI	**SNLI**	**SNLI**	**86.14**	**BiLSTM-max**
SNLI	SNLI	SICK	54.50	BiLSTM-max
SNLI	SNLI	RTE	53.07	BiLSTM-max
SNLI	SNLI	MultiNLI-m	55.71	BiLSTM-max
SNLI	SNLI	SciTail	60.16	BiLSTM-max

Table 6: Test accuracies (%) for models trained on SNLI.

Train	Dev	Test	Test Accuracy	Model details
MultiNLI	**MultiNLI-m**	**MultiNLI-m**	**73.07**	**BiLSTM-max**
MultiNLI	MultiNLI-m	SNLI	63.83	BiLSTM-max
MultiNLI	MultiNLI-m	SICK	54.12	BiLSTM-max
MultiNLI	MultiNLI-m	RTE	59.60	BiLSTM-max
MultiNLI	MultiNLI-m	SciTail	70.60	BiLSTM-max

Table 7: Test accuracies (%) for models trained on MultiNLI.

Bottom line: NLI systems not able to generalise

FOTRAN Found in Translation ACADEMY OF FINLAND

Thank You!

References I

Balazs, J., Marrese-Taylor, E., Loyola, P., and Matsuo, Y. (2017). Refining raw sentence representations for textual entailment recognition via attention. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 51–55. Association for Computational Linguistics.

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.

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

References II

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40. 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., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

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

References III

Nie, Y. and Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45. Association for Computational Linguistics.

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

Tay, Y., Tuan, L. A., and Hui, S. C. (2018). A compare-propagate architecture with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.

Vu, H. (2017). Lct-malta's submission to repeval 2017 shared task. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 56–60. Association for Computational Linguistics.

References IV

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*.