Siamese Recurrent Architectures for Learning Sentence Similarity

Jonas Mueller and Aditya Thyagarajan

Association for the Advancement of Artificial Intelligence, 2016

Problems Addressed

1. The paper presents a simple model for computing the semantic similarity between sentences of variable length, which outperforms state of the art models with carefully handcrafted features.

Proposed Solution

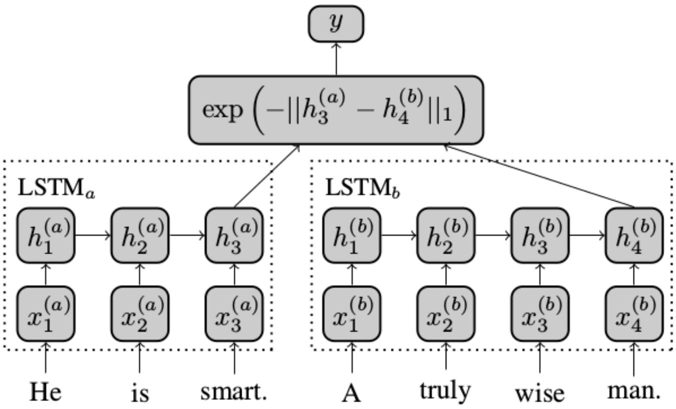


Figure 1: Our model uses an LSTM to read in word-vectors representing each input sentence and employs its final hidden state as a vector representation for each sentence. Subsequently, the similarity between these representations is used as a predictor of semantic similarity.

1. The LSTMs in this model function simply as the encoders of the seq2seq model.

2. The LSTMs learn a mapping from the space of variable length sequences of d_{in} - dimensional vectors into $R^{d_{rep}}$ ($d_{in} = 300, d_{rep} = 50$ in this work).

3. Since the similarity metric is wholly dependent on the hidden representations, the LSTM is forces to entirely capture the semantic differences during training.

4. The authors claim that using l_2 rather than l_1 norm in the similarity function can lead to undesirable plateaus in the overall objective function.

5. This is because during early stages of training, a l_2 - based model is unable to correct errors where it erroneously believes semantically different sentences to be nearly identical due to vanishing gradients of the Euclidean distance.

Experiments

1. Data augmentation is used to enable the model to generalize beyond the examples available in the datasets.

2. Invariance is encouraged over precise wording - dataset is expanded by employing thesaurus-based augmentation in which 10,022 additional training examples are generated by replacing random words with one of their synonyms found in Wordnet.

3. MSE is used as the loss function for this model, after rescaling the training-set relatedness labels to lie $\in [0, 1]$.

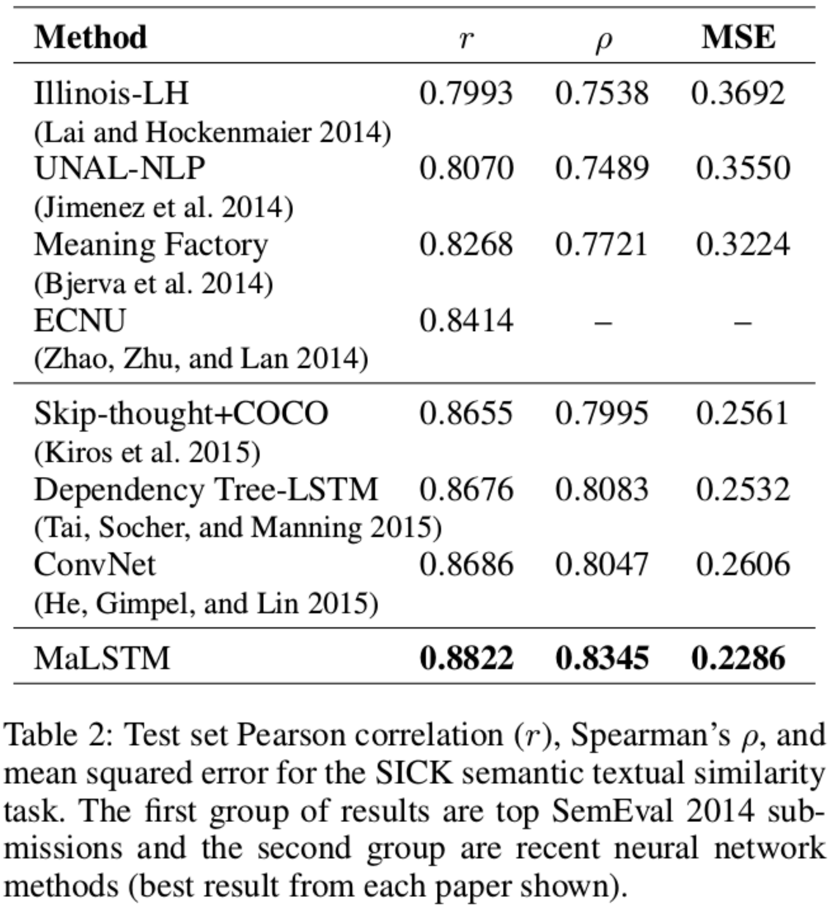
4. After training the model, we apply an additional nonparametric regression step to obtain better-calibrated predictions (with respect to MSE).

5. For experiments on the SICK dataset:

    - Over the training set, the given labels (under original [1, 5] scale) are regressed against the univariate MaLSTM g-predicted relatedness as the sole covariate.
    - The fitted regression function is evaluated on the MaLSTM-predicted relatedness of the test pairs to produce adjusted final predictions.
    - The classical local-linear estimator discussed is used, with bandwidth selected using leave-one-out cross-validation.
    - This calibration step serves as a minor correction for the restrictively simple similarity function (which is necessary to retain interpretability of the sentence representations).
    - First, the LSTM weights are initialized with small random Gaussian entries (and a separate large value of 2.5 for the forget gate bias to facilitate modeling of long range dependence).
    - Then, the MaLSTM is (pre)trained on sentence-pair data provided for the earlier SemEval 2013 Semantic Textual Similarity task.
    - The weights resulting from this pre-training thus form our starting point for the SICK data, which is markedly superior to a random initialization.

6. Adadelta is used as the optimizer for this model.

Results



Method	r	ρ	MSE
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

Table 2: Test set Pearson correlation (r), Spearman's ρ, and mean squared error for the SICK semantic textual similarity task. The first group of results are top SemEval 2014 submissions and the second group are recent neural network methods (best result from each paper shown).

- The sequential MaLSTM is slightly worse at identifying active-passive equivalence, while it is better at distinguishing verbs and objects than the compositional Tree-LSTM which often infers seemingly over-estimated relatedness scores.

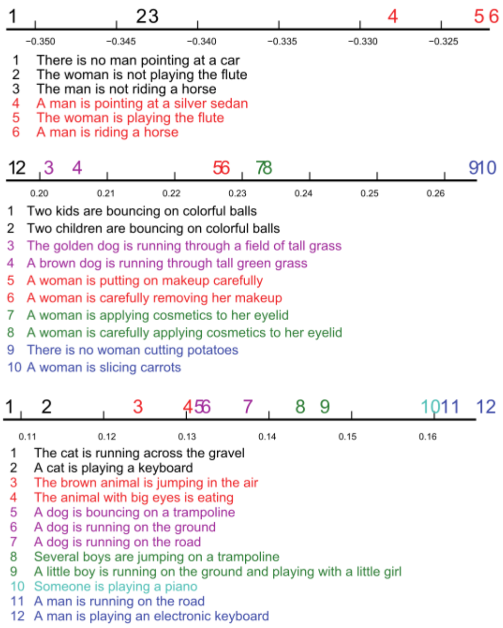
- For example, the ground truth labeling between "Tofu is being sliced by a woman" and "A woman is slicing butter" is only 2.7 in the SICK test set and substituting "potatoes" for "butter" should not greatly increase relatedness between the two statements).

- As the l_1 metric is the sum of element-wise differences, the authors hypothesize that by using specific hidden units (i.e. dimensions of the sentence representation) to encode particular characteristics of a sentence, the trained MaLSTM infers semantic similarity between sentences by simply aggregating their differences in various characteristics.

- For example, it is evident that the hidden unit shown at the top has learned to detect negation, separating sentences containing words like "no" or "not" from the rest, regardless of the other content in the text.

Ranking by Dependency Tree-LSTM Model	**Tree**	**M**
a woman is slicing potatoes		
a woman is cutting potatoes	4.82	4.87
potatoes are being sliced by a woman	4.70	4.38
tofu is being sliced by a woman	4.39	3.51
a boy is waving at some young runners from the ocean		
a group of men is playing with a ball on the beach	3.79	3.13
a young boy wearing a red swimsuit is jumping out of a blue kiddies pool	3.37	3.48
the man is tossing a kid into the swimming pool that is near the ocean	3.19	2.26
two men are playing guitar		
the man is singing and playing the guitar	4.08	3.53
the man is opening the guitar for donations and plays with the case	4.01	2.30
two men are dancing and singing in front of a crowd	4.00	2.33

Table 3: Most similar sentences (from 1000-sentence subsample) in the SICK test data according to the Tree-LSTM. Tree / M denote relatedness (with the sentence preceding each group) predicted by the Tree-LSTM / MaLSTM.

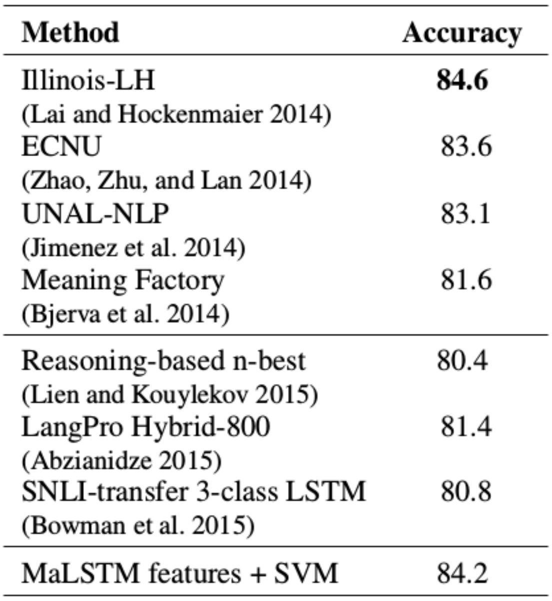


1 There is no man pointing at a car
2 The woman is not playing the flute
3 The man is not riding a horse
4 A man is pointing at a silver sedan
5 The woman is playing the flute
6 A man is riding a horse

1 Two kids are bouncing on colorful balls
2 Two children are bouncing on colorful balls
3 The golden dog is running through a field of tall grass
4 A brown dog is running through tall green grass
5 A woman is putting on makeup carefully
6 A woman is carefully removing her makeup
7 A woman is applying cosmetics to her eyelid
8 A woman is carefully applying cosmetics to her eyelid
9 There is no woman cutting potatoes
10 A woman is slicing carrots

1 The cat is running across the gravel
2 A cat is playing a keyboard
3 The brown animal is jumping in the air
4 The animal with big eyes is eating
5 A dog is bouncing on a trampoline
6 A dog is running on the ground
7 A dog is running on the road
8 Several boys are jumping on a trampoline
9 A little boy is running on the ground and playing with a little girl
10 Someone is playing a piano
11 A man is running on the road
12 A man is playing an electronic keyboard

Figure 2: MaLSTM representations of test set sentences depicted along three different dimensions of h_T (indices 1, 2, and 6). Each number above the axis corresponds to a sentence representation and its location represents the value this particular hidden unit assigns to the sentence (shown below).



Method	**Accuracy**
Illinois-LH (Lai and Hockenmaier 2014)	**84.6**
ECNU (Zhao, Zhu, and Lan 2014)	83.6
UNAL-NLP (Jimenez et al. 2014)	83.1
Meaning Factory (Bjerva et al. 2014)	81.6
Reasoning-based n-best (Lien and Kouylekov 2015)	80.4
LangPro Hybrid-800 (Abzianidze 2015)	81.4
SNLI-transfer 3-class LSTM (Bowman et al. 2015)	80.8
MaLSTM features + SVM	84.2

Table 4: Test set accuracy for the SICK semantic entailment classification. The first group of results are top SemEval 2014 submissions and the second are more recently proposed methods.

Bibliography

1. Abzianidze, L. 2015. A Tableau Prover for Natural Logic and Language. EMNLP 2492–2502.

2. Agirre, E., and Cer, D. 2013. SEM 2013 shared task: Se- mantic Textual Similarity. SemEval 2013.

3. Bengio, Y. 2012. Deep Learning of Representations for Un- supervised and Transfer Learning. JMLR W&CP: Proc. Un- supervised and Transfer Learning challenge and workshop 17–36.

4. Bjerva, J.; Bos, J.; van der Goot, R.; and Nissim, M. 2014. The meaning factory: Formal semantics for recognizing tex- tual entailment and determining semantic similarity. Se- mEval 2014.

5. Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. EMNLP 632–642.

6. Chen, K., and Salman, A. 2011. Extracting Speaker-Specific Information with a Regularized Siamese Deep Network. NIPS 298–306.

7. Cho, K.; Gulcehre, B. v. M. C.; Bahdanau, D.; Schwenk, F. B. H.; and Bengio, Y. 2014. Learning Phrase Representa- tions using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 1724–1734.

8. Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learn- ing a similarity metric discriminatively, with application to face verification. Computer Vision and Pattern Recognition 1:539–546.

9. Fan, J., and Gijbels, I. 1992. Variable bandwidth and lo- cal linear regression smoothers. The Annals of Statistics 20:2008–2036.

10. Graves, A. 2012. Supervised Sequence Labelling with Re- current Neural Networks. Studies in Computational Intelli- gence, Springer.

11. Greff, K.; Srivastava, R. K.; Koutnik, J.; Steunebrink, B. R.; and Schmidhuber, J. 2015. LSTM: A Search Space Odyssey. arXiv: 1503.04069.

12. He, H.; Gimpel, K.; and Lin, J. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Net- works. EMNLP 1576–1586.

13. Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. Neural Computation 9(8):1735–1780.

14. Jimenez, S.; Duenas, G.; Baquero, J.; Gelbukh, A.; Bátiz, A. J. D.; and Mendizábal, A. 2014. Unal-nlp: Combining soft cardinality features for semantic textual similarity, related- ness and entailment. SemEval 2014.

15. Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Skip-Thought Vectors. NIPS to appear.

16. Lai, A., and Hockenmaier, J. 2014. Illinois-lh: A deno- tational and distributional approach to semantics. SemEval 2014.

17. Le, Q., and Mikolov, T. 2014. Distributed Representations of Sentences and Documents. ICML 1188–1196.

18. Li, Y.; Xu, L.; Tian, F.; Jiang, L.; Zhong, X.; and Chen, E. 2015. Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective. IJ- CAI.

19. Lien, E., and Kouylekov, M. 2015. Semantic Parsing for Textual Entailment. International Conference on Parsing Technologies 40–49.

20. Marei, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; and Zamparelli, R. 2014. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. SemEval 2014.

21. Mihal ea, R.; Corley, C.; and Strapparava, C. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. AAAI Conference on Artificial Intelligence.

22. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. NIPS 3111–3119. Miller, G. A. 1995.

23. WordNet: A Lexical Database for English. Communications of the ACM 38(11):39–41.

24. Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. ICML 1310–1318.

25. Siegelmann, H. T., and Sontag, E. D. 1995. On the Computational Power of Neural Nets. Journal of Computer and System Sciences 50:132–150.

26. Socher, R. 2014. Recursive Deep Learning for Natural Lan- guage Processing and Computer Vision. Phd thesis, Stanford University.

27. Sutskever, I.; Vinyals, O.; and Le, Q. 2014. Sequence to sequence learning wit and Manning, C. D. 2015

28. Impred Semantic Representations From Tree-Structured Long Short-ional Data Using t-SNE. Journal of Machine Learning Research 9:2579–2605.

29. Yih, W.; T inative Projections for Text Similarity Measures. Proceedings of the Fifteenth Conference on Computational Natural Language Learning 247–256.Zeiler, M. D. 2012. ADADELTA: An Adaptive Learning Rate Method. arXiv:1212.57 01.. 2015. Character-level Convolutional Networks for Text Classification. arXiv:1509.01626.

30. Zhao, J.; Zhu, T. T.; and Lan, M. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. SemEval 2014.