Direct Network Transfer for Semantic Similarity

Li Zhang, Steven R. Wilson, and Rada Mihalcea, University of Michigan
{zharry,steverw,mihalcea}@umich.edu

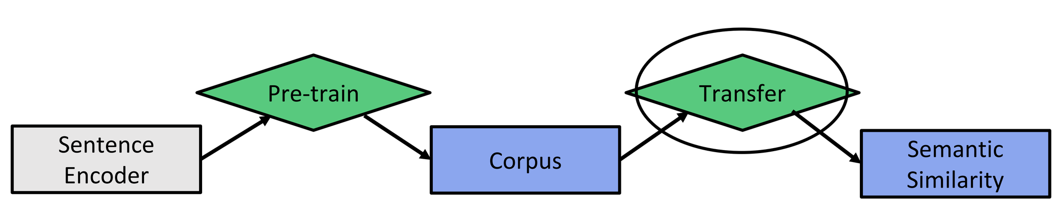
Introduction

- Sentence encoders map a sentence to a fixed-size vector:
  - BiLSTM-Avg [1]
  - GRAN [2]
  - InferSent [3]
- Semantic similarity tasks compute a scale of relatedness between two sentences:

Sentence 1	Sentence 2	Annotated Similarity
A man is cutting up a cucumber.	A man is slicing a cucumber.	4.2
A man is dancing.	A woman is exercising.	0.4
The Dow Jones industrial average .DJI ended up 56.79 points, or 0.67 percent, at 8,588.36 -- its highest level since January 17.	he Dow Jones Industrial Average ($DJ: news, chart, profile) rose 56 points, or 0.7 percent, to 8,588.	3.6

Table 1: Some examples from the STS Benchmark [4].

- Sentence encoders are applied to semantic similarity tasks by using transfer learning:



- We introduce a new transfer learning setting called *direct network transfer* with better performance overall and state-of-the-art in some datasets.

Data

- We evaluate on four semantic similarity dataset:

Human Activity [5]
- Pairs of phrases describing daily human activities in four relations
- 1000 - 375 - 1000 pairs in train-dev-test splits

SICK [6]
- A large number of sentence pairs that are rich in the lexical, syntactic and semantic phenomena
- 4439 - 495 - 4906 pairs in train-dev-test splits

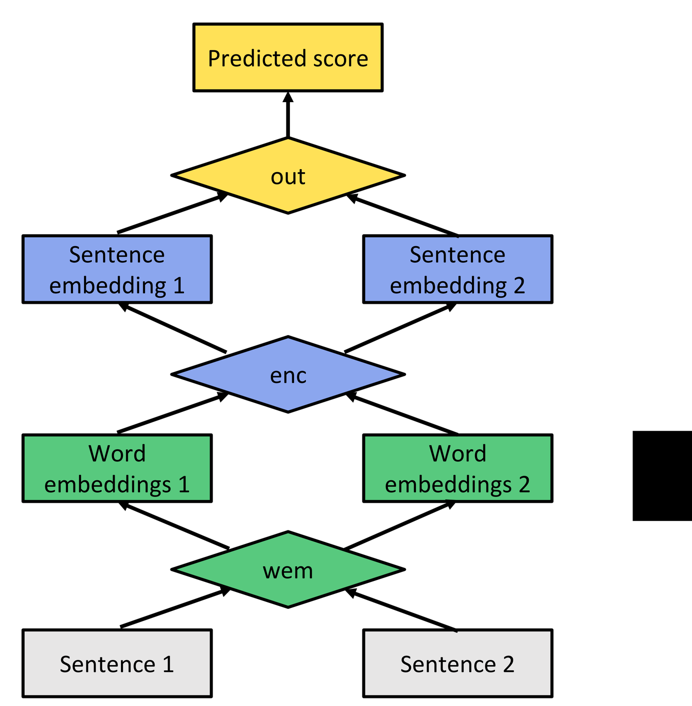
SemEval STS
- A selection of the datasets used in the STS tasks organized by SemEval
- STS Benchmark: 5749 - 1500 - 1379 pairs in train-dev-test splits
- STS 12 [7]: 2000 - 234 - 1959 pairs in train-dev-test splits

Short Answer Grading [8]
- A collection of student and instructor answers to questions on assignments and examinations
- 1460 - 500 - 552 pairs in train-dev-test splits

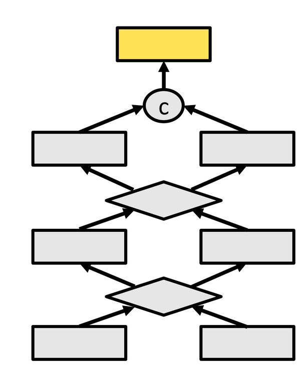
Activity 1	Activity 2	SIM	REL	MA	PAC
go jogging	lift weights	1.67	2.22	2.89	1.11
read to one's kids	go to a bar	0	0	0	-1.29
take transit to work	commute to work	3.38	3.5	3.38	0.5
make one's bed	organize one's desk	0.58	1.29	1.57	0.71

Table 2: Sample of scores assigned to pairs of activities in the Human Activity Dataset. SIM, REL, and MA scores are in the range [0,4] and PAC scores lie in [-2,2]. Scores are averaged across 10 annotators.

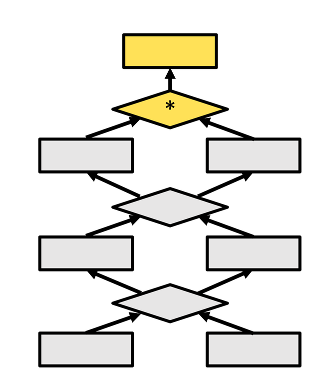
Sentence Encoder Transfer Settings



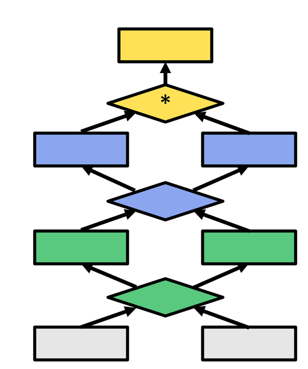
General Sentence Encoder Training Architecture
- Initial parameters are learned from a large *source* dataset.
- Afterwards, there are several ways to transfer the learned model to the *target* dataset (e.g., human activities).



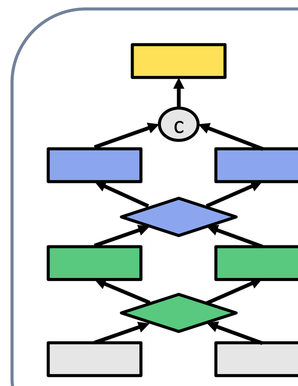
Unsupervised Evaluation
- No parameters are updated using the target dataset
- Labels are generated directly from sentence embeddings using (c) (cosine similarity)



Feature Transfer
- All embedding layers are frozen during target training
- A new logistic regression classifier (*) is trained to make predictions



Network Transfer
- Embedding parameters are allowed to be updated during target training
- A new logistic regression classifier (*) is trained to make predictions



Direct Network Transfer
- Embedding parameters are allowed to be updated during target training
- Labels are generated **directly** from sentence embeddings using (c) (cosine similarity)

Experimental Results

Datasets	STS Bench.	SICK	STS 12	SIM	REL	MA	PAC	SAG
BiLSTM-Avg [UE]	.791/.783	.735	.803	.649	.639	.603	.469	.450
BiLSTM-Avg [FT] MSE	.779/.746	.860	**.867**†	.534	.514	.474	.412	.761
BiLSTM-Avg [FT] KL	.797/.779	.861	.864	.518	.509	.461	.400	.774
BiLSTM-Avg [NT] MSE 🔒	.836/.810	.864	.860	.576	.575	.529	.456	.761
BiLSTM-Avg [NT] MSE 🔓	.833/.809	.864	.861	.571	.571	.526	.453	.806
BiLSTM-Avg [NT] KL 🔒	.840/.806	**.866**	.854	.559	.558	.515	.459	.801
BiLSTM-Avg [NT] KL 🔓	.837/.802	.864	.845	.556	.529	.512	.449	.813
BiLSTM-Avg [DNT] 🔒	**.852/.824**†	.856	.861	**.699**	**.688**	**.660**	**.470**	.816
BiLSTM-Avg [DNT] 🔓	**.851/.824**†	.859	.861	.691	.680	.646	.462	**.834**
GRAN [UE]	.688/.583	.703	.560	.644	.642	.596	**.444**	.323
GRAN [FT] MSE	.759/.693	.792	.651	.561	.576	.526	.392	.504
GRAN [FT] KL	**.771/.701**	.790	.649	.556	.577	.525	.398	.649
GRAN [NT] MSE 🔒	.710/.648	**.857**	**.734**	.575	.567	.523	.375	.742
GRAN [NT] MSE 🔓	.720/.653	.855	.726	.578	.560	.510	.385	.736
GRAN [NT] KL 🔒	.717/.643	**.857**	.731	.558	.574	.530	.401	.802
GRAN [NT] KL 🔓	.717/.644	.853	.718	.541	.537	.442	.415	.791
GRAN [DNT] 🔒	.749/.644	**.857**	.670	**.668**	.663	**.624**	.407	.792
GRAN [DNT] 🔓	.745/.641	**.857**	.663	**.668**	**.666**	**.623**	.413	**.807**
InferSent [UE]	.782/.738	.748	.607	**.701**†	.686	.652	.525	.209
InferSent [FT] MSE	.809/.757	**.884**†	**.792**	.655	.644	.608	.432	.738
InferSent [FT] KL	**.831/.783**	.882	.788	.688	.680	.642	.510	.735
InferSent [NT] MSE 🔒	.783/.744	.859	.777	.699	.692	.672	.537	.792
InferSent [NT] KL 🔒	.812/.763	.867	**.791**	.679	.668	.634	.484	.783
InferSent [DNT] 🔒	.802/.740	.854	.742	**.702**†	**.722**†	**.691**†	**.572**†	**.838**†

Table 3: The performance of transfer settings for three models across all datasets. Spearman's r is reported for Human Activity Phrase dataset including the four dimensions SIM, REL, MA and PAC, and Pearson's r for the rest, in accordance with the specification of the dataset to allow for direct comparison with previous results. The lock icon indicates freezing the word embedding matrix weights (wem), and the unlock icon indicates updating them. Note that wem of InferSent must be frozen due to its implementation constraints. For each dataset, the best transfer result per-model is listed in bold font, the best overall result is underlined, and the state-of-the-art result is marked by a dagger.

Analysis on Human Activities

- We distinguish between two types of pairs for which transfer is helpful and show some illustrative examples:

1. Pairs with scores that were initially overestimated

Phrase 1	Phrase 2
have dinner with friends	eat dinner by oneself
go to a party	go to bible study
play football	play basketball
go to the movie theater	go to office to work

2. Pairs with scores that were initially Underestimated

take long walks	go on a walk
take care of one's dogs	groom one's dog
read books	visit a bookstore
go to the doctor	see the doctor

- We use the leave-one-out ablation analysis as a basis for the following definition of the irrelevance:

$$irrelevance(w, p_1, p_2, m_1, m_2) = m_2(p_1^{\neg w}, p_2) - m_1(p_1^{\neg w}, p_2)$$

We explore the effect of PAC dimension. Two illustrative example are shown here.

have	dinner	with	friends
0.58	0.37	0.65	0.4

eat	dinner	by	oneself
0.54	0.4	0.64	0.35

go	to	a	party
0.22	0.31	0.33	0.13

go	to	bible	study	at	church
0.2	0.33	0.52	0.4	0.34	0.25

Conclusions

- Direct network transfer is the best transfer learning setting in most cases
- Direct network transfer with BiLSTM-Avg achieves state-of-the-art performance on STS Benchmark
- Direct network transfer with InferSent achieves state-of-the-art performance on Human Activity dataset
- The choice of transfer learning setting influences performance in most cases
- Freezing lower layers, choice of loss function and normalization of scores also influence performance and should be tuned as hyperparameters

References

[1] Wieting, J., and Gimpel, K. 2017a. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *arXiv preprint arXiv:1711.05732*.

[2] Wieting, J., and Gimpel, K. 2017b. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. *ArXiv eprints*.

[3] 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*.

[4] Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. Semeval-2017 task 1: Semantic textual similaritymultilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

[5] Wilson, S. and Mihalcea, R., 2017. Measuring Semantic Relations between Human Activities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Vol. 1, pp. 664-673).

[6] Marelli, M.; Menini, S.; Baroni, M.; Bentivogli, L.; Bernardi, R.; and Zamparelli, R. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC, 216–223*.

[7] Agirre, E.; Diab, M.; Cer, D.; and Gonzalez-Agirre, A. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task*, and *Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, 385–393*. Association for Computational Linguistics.

[8] Mohler, M.; Bunescu, R.; and Mihalcea, R. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 752–762*. Association for Computational Linguistics.