Towards Generalizable Sentence Embeddings

Eleni Triantafillou, Jamie Ryan Kiros, Raquel Urtasun, Richard Zemel
Department of Computer Science
University of Toronto
{eleni, rkiros, urtasun, zemel}@cs.toronto.edu

Abstract

In this work, we evaluate different sentence encoders with emphasis on examining their embedding spaces. Specifically, we hypothesize that a "high-quality" embedding aids in generalization, promoting transfer learning as well as zero-shot and one-shot learning. To investigate this, we modify Skipthought vectors to learn a more generalizable space by exploiting a small amount of supervision. The aim is to introduce an additional notion of similarity in the embeddings, rendering the vectors informative for different tasks requiring less adaptation. Our embeddings capture human intuition on similarity favorably than competing models, while we also show positive indications of transfer from the task of natural language inference to paraphrase detection and paraphrase ranking. Further, our model's behaviour on paraphrase detection when trained with an increasing amount of labelled data is indicative of a generalizable model. Finally, we support our hypothesis on generalizability of our embeddings through inspection of their statistics.

1 Introduction

Natural language is an integral part of numerous applications, such as web search, information retrieval, and automatic text summarization, to name just a few. Therefore, constructing high-quality text representations is very important. In addition, despite having well-established methods to construct word representations, it remains an open problem to capture the semantics of larger pieces of text in a vector that is useful for different tasks with minimal adaptation. In this paper we report on our efforts towards building such generalizable sentence representations.

Representing a sentence as a vector can be thought of as "embedding" it into a high-dimensional space. Therefore, a meaningful representation relies on a function which sends "related" sentences to neighbouring points in this vector space. There are, however, many possible notions of closeness that may be desirably reflected in the embeddings. For instance, two sentences could be considered similar if they are likely to be found in the same context ("distributional similarity"), or if the second is entailed from the first, or if they are paraphrases of each other.

We hypothesize that an embedding space which adheres to multiple of these notions can host more generalizable vectors. For instance our hypothesis is that in a "generalizable" space, two sentences that are likely to be found in the same context and also entail each other are closer that two other sentences which are also likely to be found in the same context but contradict each other.

Moreover, we believe that "supervised evaluation" of sentence encoders is not informative of the embedding quality: a classifier is trained on top of the sentence embeddings and then the accuracy for the task is computed, and is used as a proxy for the quality of the embeddings. This approach has the disadvantage that it hides the embedding properties due to the extra training which allows to mend its potential shortcomings. We instead focus our attention on directly inspecting the model space.

In this work, we introduce a sentence encoder that is learned by injecting supervised information from the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) in the commonly used Skipthought embeddings. The aim is to enhance the embeddings with an additional notion of similarity, rendering them more generalizable. We experiment with this model and with

Proceedings of the 1st Workshop on Representation Learning for NLP, pages 239–248,
Berlin, Germany, August 11th, 2016. ©2016 Association for Computational Linguistics

sentence encoders of different training objectives in order to compare both their performance in the commonly used supervised fashion (Kiros et al., 2015; Collobert et al., 2011; Mikolov et al., 2013a; Wieting et al., 2015) for reference, but more importantly their embedding quality. We perform supervised evaluation on paraphrase detection, semantic relatedness, natural language inference and various classification benchmarks. We then evaluate the embeddings through paraphrase ranking, correlation of their similarity notion with human judgements (Hill et al., 2016; Hill et al., 2015; Levy et al., 2015; Baroni et al., 2014), through paraphrase detection with little or no training, and through examination of embedding statistics.

2 Related Work

The Skip-gram model for word embeddings (Mikolov et al., 2013a; Mikolov et al., 2013b), is trained on a text corpus with the objective of predicting the vectors of the surrounding words of a given word, when conditioned on its vector representation. Its success inspired Kiros et al. (2015) to create its sentence analogue, Skipthought vectors, which are trained by predicting the surrounding sentences when conditioned on the current one. Despite this simple objective, Skipthoughts perform remarkably well on various tasks: semantic relatedness, paraphrase detection, image-sentence ranking, and a number of classification benchmarks. In this paper we investigate how we can improve their embedding space through injecting small amounts of supervised information.

Aside from Skipthoughts, there are numerous sentence encoders. (Socher et al., 2013; Yin and Schütze, 2015; Wang and Nyberg, 2015; Socher et al., 2014) create sentence encoders which are optimized for a specific task of interest. On the other hand, methods which aim at constructing ”universal” embeddings include (Le and Mikolov, 2014; Socher et al., 2011; Li et al., 2015; Pham et al., 2015). Le and Mikolov (2014) learn paragraph embeddings by predicting sentences within a paragraph when conditioned on its representation, Pham et al. (2015) predict context in all levels of a syntactic tree, whereas Socher et al. (2011) and Li et al. (2015) present autoencoder-type models.

Hill et al. (2016) presented an extensive evaluation of unsupervised sentence encoders. They showed that Bag of Words (BOW) models on average perform on par with non-BOW models. Our results agree with this and we provide a possible explanation through examining the statistics of the datasets used for evaluation. An important distinction between our work and theirs is that Hill et al. (2016) focused on models that were trained in an unsupervised fashion, whereas we also present models finetuned or trained on SNLI for natural language inference.

Wieting et al. (2015) learned “universal” sentence vectors by exploiting a database of paraphrases: they optimize an objective which encourages paraphrases to lie closer to each other in space than to negative examples. Similarly to their work, we also use supervised information to construct informative embeddings, but our supervision comes from the task of natural language inference.

Transfer learning is the process of exploiting knowledge from one task or domain in order to benefit from it for a different (“target”) task or domain. This method has enjoyed considerable success in computer vision applications (notably the use of features derived from neural networks trained for object classification such as (Krizhevsky et al., 2012) for other tasks) but is less successful in language applications. Collobert and Weston (2008) perform mutli-task learning on various natural language processing tasks and report a very small gain for each task. Mou et al. (2016) presented negative results on their effort to transfer from the task of natural language inference to paraphrase detection. In this work, we show positive results on transfer from natural language inference to paraphrase detection and to the related task of paraphrase ranking.

3 Models

The success of Skipthoughts verify that predicting the “context” of a sentence is a valuable objective. However, sentences are also a nontrivial function of the words comprising them, so we believe that explicitly capturing their “content” in addition to their “context” can yield more informative embeddings.

Therefore, we experimented with several ways of capturing content, and evaluate the quality of content-only encoders as well as that of an encoder that combines a content with a context objective. The content-only models we use are two autoencoders (AEs): a BOW one which we refer to as “BOW AE”, and a Recurrent Neural Netowrk (RNN)-based one, referred to as “RNN AE”. BOW AE is the model proposed in (Lauly et al., 2014) which encodes the sentence as a vector that indicates which vocabulary words are present in the sentence, irrespective of their order. The objective is to reconstruct this indicator vector when given a nonlinear function of the sum of the embeddings of the present words according to the indicator. RNN AE, on the other hand, uses an RNN encoder with GRU units to represent the sentence and a similar RNN decoder which is trained to predict the same sentence when conditioned on its encoded representation. We trained these models on the Toronto book corpus (Zhu et al., 2015).

We also made use of the SNLI dataset for the purpose of capturing content, and created 3 “SNLI models”: a BOW and an RNN-based “content SNLI” models, which we refer to as “SNLI BOW” and “SNLI RNN”, respectively, as well as a finetuned version of Skipthoughts which we argue has encoded a combination of context and content and refer to as “SNLI-finetuned Skipthoughts”. These 3 SNLI models are illustrated in Figure 1.

SNLI is comprised of pairs of sentences with a label of “entailment”, “contradiction”, or “neutral” associated with each pair. Each SNLI model creates the representation of each sentence of the given pair separately (but using the same encoder), and then concatenates the two sentence embeddings, and feeds these into a single hidden layer neural network, with a softmax on top for the three-way classification of SNLI. We backpropagate through the encoder and the word embeddings as well. In the case of SNLI-finetuned Skipthoughts, the encoder is initialized from Skipthoughts, and subsequently finetuned to add “content-based” SNLI information. On the other hand, the encoder of SNLI BOW merely corresponds to the sum of the word embeddings (which are initialized from Skipthought word embeddings and modified during training), while the encoder of SNLI RNN is an RNN which is initialized “from scratch”.

An overview of the model space is presented in figure 2.

4 Training Details

The Skipthought model that we compare with in the experiments is the model which is referred to as combine-skip in (Kiros et al., 2015). This model is created from the combination of two separate encoders: a “uni-directional” and a “bi-directional” one. The uni-directional model is comprised of an RNN encoder with GRU units whose hidden state consists of 4800 dimensions. The bi-directional model is the concatenation of a 1200-dimensional GRU RNN encoder which reads the sentence in forward order (from left to right) and an equally sized GRU RNN encoder which reads the sentence in reverse order. After the separate training of the uni-directional and bi-directional models, their representations are combined for the creation of the 4800-dimensional combine-skip embedding.

In order to fairly compare with the combine-

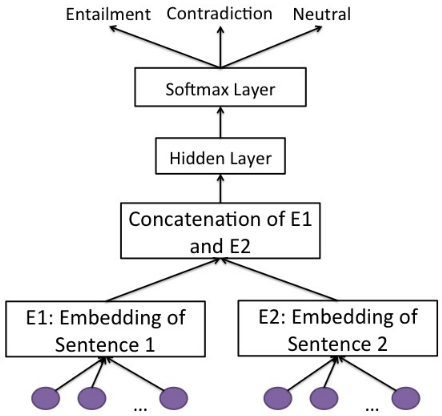

Entailment
Contradiction
Neutral
Softmax Layer
Hidden Layer
Concatenation of E1 and E2
E1: Embedding of Sentence 1
E2: Embedding of Sentence 2
...
...

Figure 1: The 3 SNLI models are all formed from the same formulation, illustrated above. For SNLI-finetuned Skipthoughts, E1 and E2 are initialized to the Skipthought Encoder, for SNLI BOW E1 and E2 are the sum of the word embeddings and for SNLI RNN E1 and E2 are an RNN encoder which is initialized from scratch.

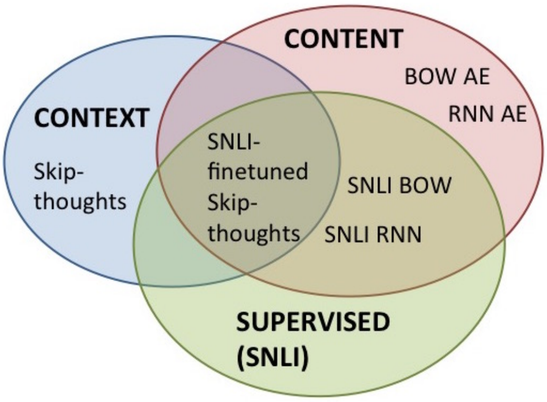


Figure 2: An overview of the models used in the experiments of this paper.

skip model, we have created the analogous SNLI-finetuned Skipthoughts model. The unidirectional and bi-directional Skipthought models were finetuned separately using the architecture mentioned in the previous section (Figure 1) and subsequently combined to yield 4800-dimensional embeddings. This is the model which we refer to as SNLI-finetuned Skipthoughts in the remainder of this paper.

The word embeddings of all 3 SNLI models and the 2 autoencoder models was initialized from the Skipthought word embeddings, which are 620-dimensional. The RNN of SNLI RNN, and the encoder and decoder RNNs of RNN AE have GRU units and their hidden state is 2400-dimensional. We initialized these recurrent weights with orthogonal initialization (Saxe et al., 2013). The non-recurrent weights of the hidden and softmax layers for the SNLI models are initialized from a uniform distribution in the range [-0.1, 0.1].

Adam optimizer (Kingma and Ba, 2014) was used for training all of these models.

In the following sections we present our results in two evaluation settings: Firstly, we perform supervised evaluation (Section 5), and then more importantly we directly evaluate the embedding space of the different models (Section 6).

5 Supervised Evaluation

In this section we present results on a number of supervised tasks. These results are obtained by encoding the sentence at hand (or each sentence of the pair when applicable) and using this encoding as the features of a logistic regression which is trained for the given task, following the approach in (Kiros et al., 2015). For the tasks involving pairs of sentences, the features that were given to the logistic classifier were computed as follows: the element-wise product and absolute difference between the two sentence embeddings were computed and then concatenated, resulting in a 9600-dimensional vector, as was also done in (Kiros et al., 2015). In the next paragraph we briefly describe the tasks that we report experiments on.

Paraphrase detection (MSRP dataset) is the task where given pairs of sentences, the goal is to assign a binary label indicating whether the sentences of each pair are paraphrases. For semantic relatedness we use SICK (Marelli et al., 2014), and the objective is to assign a score of relatedness in the range 1-5 to pairs of sentences. Natural language inference is the task of predicting a label of “entailment”, “contradiction”, or “neutral” for each pair. Note that SNLI is a dataset for this task, but the results we report here are on SICK, which has both relatedness scores as well as these 3-way classifications labels for each pair. TREC is a dataset for (6-way) question-type classification and finally, MR and SUBJ come from a movie review dataset and they are binary classification tasks for sentiment polarity (MR) and subjectivity status (SUBJ).

The results are shown in tables 1, 2, 3, and 4. In all tables, we use the following abbreviations for model names. ST: Skipthoughts, FT-ST: SNLI-FineTuned Skipthoughts, BOW: SNLI-BOW, RNN: SNLI-RNN. Results which outperform or perform on par with Skipthoughts are shown in bold, since Skipthoughts have shown to perform remarkably well in this supervised evaluation setting as demonstrated in detail in (Kiros et al., 2015), and verified in (Hill et al., 2016).

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
test acc	0.73	**0.75**	0.70	0.71	0.71	0.67
test f1	0.82	**0.83**	0.80	0.81	0.80	0.80

Table 1: Results on paraphrase detection (MSRP).

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
test acc	0.80	**0.83**	**0.81**	**0.82**	0.79	0.75

Table 2: Results on natural language inference (SICK). Note that this is the same task as SNLI, but different dataset.

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
test PR	0.84	**0.85**	0.81	0.82	0.79	0.70
test SR	0.78	**0.79**	0.76	0.77	0.72	0.64
test SE	0.30	**0.28**	0.36	0.34	0.39	0.52

Table 3: Results on semantic relatedness (SICK). PR, SR, SE: Pearson, Spearman correlation coefficient and mean squared error, resp. between model scores and human scores.

Overall, the most important observation is that a lot of these results are very comparable, with the reported numbers being within a small range in most cases, despite the very different nature of these models. For example, the fact that SNLI-RNN performs comparably with Skipthoughts on Semantic Relatedness is surprising given their training objectives. Recall that the encoder in SNLI-RNN was initialized from scratch. This fact

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
MR	0.76	**0.79**	**0.76**	0.71	0.75	0.65
SUBJ	0.94	**0.94**	0.93	0.89	0.92	0.86
TREC	0.92	**0.92**	0.86	0.87	0.85	0.82

Table 4: Results on various classification tasks. Each row stores test accuracies of the corresponding dataset.

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
accuracy@1	0.63	0.77	0.87	0.56	**0.90**	0.33
accuracy@10	0.74	0.86	0.93	0.67	**0.96**	0.40
accuracy@100	0.86	0.96	0.99	0.84	**1**	0.54
MRC	93	15	6	97	**2**	455

Table 5: Results on paraphrase ranking. accuracy@k is the proportion of sentences for which the true paraphrase received rank at most k. MRC is the Mean Rank of the Correct paraphrase.

could suggest that SNLI information, when injected into an RNN encoder and fed into a logistic regression classifier, is adequate in order to perform reasonably well on paraphrase detection and semantic relatedness. But we believe that this observation underlines the weakness of this method of evaluation.

The 2% improvment of SNLI-Finetuned Skipthoughts over skipthoughts for paraphrase detection is an indication of transfer from SNLI to MSRP, on which (Mou et al., 2016) presented negative results. Our results on transfer to the related task of paraphrase ranking which also uses MSRP (Section 6.1) are even more encouraging.

Further, on natural language inference all SNLI models (slightly) outperform the non-SNLI ones. This is not surprising given their training objective, and constitutes a less impressive sign of transfer between these two datasets of the same task.

Finally, we note that SNLI-finetuned Skipthoughts perform either better or on par with Skipthoughts on all tasks considered. This shows that the added SNLI information does not hurt Skipthoughts' performance on this evaluation while outperforming it in terms of embedding space quality as demonstrated in the next section, which we argue is more important.

6 Evalutating Embedding Spaces directly

In this section we evaluate the embedding quality directly. We do this firstly through paraphrase ranking, secondly though correlating embedding similarity with human scores, thirdly through paraphrase detection with few or no labelled examples, and finally through examination of the statistics of the embeddings.

6.1 Paraphrase Ranking

Paraphrase ranking is the task of assigning a rank to each sentence from a pool S, representing how likely they are to be paraphrases of a given sentence. To compute the ranks that S_1 assigns to the sentences, $sim(v_1, v_2)$ is computed $\forall S_2 \neq S_1 \in S$, where sim stands for cosine similarity and v_1 and v_2 are the embeddings of S_1 and S_2, respectively. Ranks are then assigned by sorting these similarities in decreasing order. In this setting, the sentence to which S_1 assigns rank 1 is predicted to be its paraphrase.

For this, we used the sentences from the MSRP dataset, which is comprised of pairs of sentences with a binary label indicating whether or not they are paraphrases. We "break" the pair ties and treat all sentences as members of a large pool, making use of the (binary) labels of MSRP in order to yield the (non-binary) label for this new task. We use both the training and test set of MSRP for this, totalling over 11000 sentences. The evaluation metrics we used are "Mean Rank of the Correct paraphrase", referred to as MRC from now on, and accuracy@k which is the proportion of sentences for which the true paraphrase is contained in the top k ranked sentences.

The results are shown in Table 5. We observe that BOW AE outperforms Skipthoughts by a large margin, with BOW following closely behind. In fact, it is not a coincidence that BOW models perform well on this task. To investigate this effect even further, we created a very simple BOW model which represents the sentence as the sum of its word embeddings, which are randomly generated 620-length vectors. Its performance is shown in Table 6. This model is in no way informative of the semantics, syntax, structure, or any useful property of the sentence whatsoever, and yet it outperforms Skipthoughts for example. This problematic behavior may be due to the fact that sentences in pairs with positive labels in MSRP have a very high word overlap. This suggests that any BOW model has an unfair advantage when evaluated on this dataset. We elaborate on this in the Discussion section, and provide statistics from the datasets to support this hypothesis.

However, the comparison between

Skipthoughts and SNLI-finetuned Skipthoughts here is valuable. The superiority of the latter model constitutes positive results of transfer from SNLI to paraphrase ranking using MSRP, supporting our conjecture regarding the generalizability of the SNLI-finetuned Skipthought space.

	random BOW
accuracy@1	79
accuracy@10	88
accuracy@100	96
MRC	21

Table 6: Very simple baseline for paraphrase ranking. random BOW is no way capturing anything informative about the sentence (see section 5.1 for description). These results suggest that BOW models may have an unfair advantage for MSRP.

6.2 Semantic Relatedness

SICK is comprised of pairs of sentences, each associated with a relatedness score in the range from 1 to 5. In order to directly evaluate the merit of the embeddings in capturing semantics, we used cosine similarity to estimate the relatedness of each pair. These similarity scores were then correlated with the human-annotated scores using Pearson's and Spearman's correlation coefficients and mean squared error. The results are shown in Table 7.

	ST	FT-ST	BOW	RNN	BOW-AE	RNN-AE
test PR	0.50	0.57	0.69	0.62	0.64	0.39
test SR	0.48	0.56	0.65	0.59	0.57	0.39
test SE	1.53	1.10	1.12	0.98	1.21	1.84

Table 7: Results on semantic relatedness (SICK) based on cosine distances. PR, SR, SE: Pearson, Spearman correlation coefficient and mean squared error, resp. between model scores and human scores.

As was the case for the MSRP dataset, we believe that SICK offers an unfair advantage to BOW models, therefore we do not believe that the success of BOW AE and SNLI BOW is necessarily indicative of their quality.

We observe that SNLI-Finetuned Skipthoughts outperform Skipthoughts on this task as well, supporting the conjecture that adding supervision through SNLI has lead to a more informative space. Moreover, the performance of SNLI RNN is impressive, outperforming both Skipthought-based models. Finally, out of the BOW models, the SNLI one performs better than the AE one. These are indications that SNLI information can aid in inducing a notion of similarity which is compatible with human intuition.

6.3 Towards Zero Shot Paraphrase Detection

We claimed earlier that learning an informative embedding space would facilitate zero-shot and one-shot learning applications. The aim of this section is to investigate whether SNLI-finetuned Skipthoughts are a more appropriate model for this purpose than the other models we explored.

By zero-shot paraphrase detection, we refer to the task of predicting a binary label for the "paraphrase status" of a pair of sentences without performing any training for this task. This amounts to choosing a threshold so that a pair is classified positively if and only if the similarity of its sentence embeddings surpasses this threshold. Since the choice of such a threshold is not obvious, we present the precision-recall curve in Figure 3 which corresponds to multiple thresholds.

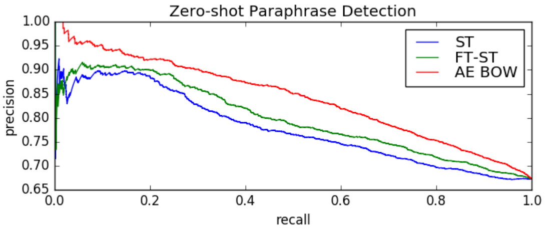


Figure 3: Precision-Recall curve for zero-shot paraphrase detection. (FT-)ST stands for (SNLI-Finetuned) Skipthoughts.

In figure 3 we have included the best-performing model for this task for reference, which is BOW AE, but we are more interested in the comparison between Skipthoughts and SNLI-Finetuned Skipthoughts. This is because we believe that the success of BOW AE on this task does not necessarily reflect its merit as a sentence encoder, as we elaborate on in the Discussion section.

The superior performance of SNLI-finetuned Skipthoughts in Figure 3 advocates for the generalizability of the former model since it requires less adaptation for paraphrase detection compared to Skipthoughts.

It is also interesting to investigate the behavior of our models when given various amounts of training data for the task of paraphrase detection.

For this, we plot in Figure 4 how the test set accuracy increases as more data is fed into the models.

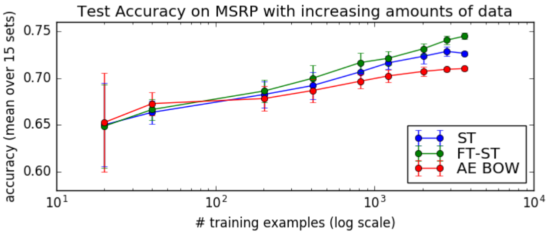


Figure 4: Test accuracy when an increasing amount of data was used for training. (FT-)ST stands for (SNLI-Finetuned) Skipthoughts.

We notice that AE BOW is less “data hungry” in that its performance ceases to increase significantly with the increase of data. SNLI-Finetuned Skipthoughts reach higher accuracy than Skipthoughts when given the same amount of data, supporting its aptness for one shot learning.

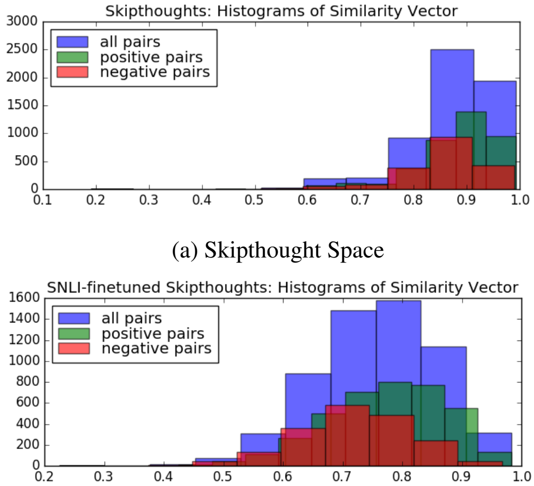
Figure 4 justifies the fact that Skipthought-based models outperform BOW AE in the supervised evaluation setting even though these roles are reversed in the zero-shot setting of Section 5. In particular, it appears that BOW AE is less capable of taking advantage of training data to improve its quality.

6.4 Diving into Embedding Space

In this section, we use the sentences from MSRP and examine their relationships in model space. The histograms in Figure 5 show the distribution of the pairwise-similarity for these sentences in Skipthought and SNLI-Finetuned Skipthought spaces.

Crucially, we observe that in Skipthought space, pairwise sentence similarities are significantly higher than in its SNLI-Finetuned variant, and there is less variation. This behavior can be attributed to their training objective. In particular, two sentences are neighbours in this space if they are likely to be found in the same contexts, resulting to sentences such as “I love sushi”, “I really really like sushi” and “I hate sushi” to be possibly equidistant neighbours.

On the other hand, the histogram for SNLI-Finetuned Skipthoughts contains more variation, which we conjecture is due to the fact that a second notion of relatedness is introduced, which pushes sentences with contradictory meanings further away from each other, in order to keep sentences which entail each other close.



(a) Skipthought Space

(b) SNLI-Finetuned Skipthought Space

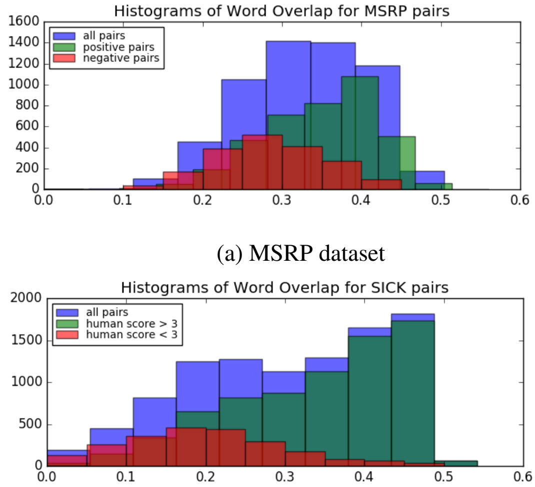
Figure 5: Distribution of Pairwise Similarities in Embedding Space (best viewed in color). Green and red denote positive and negative pairs, respectively.

Moreover, it is of interest to examine the red and green histograms, corresponding to the distributions of pair-wise similarities for positively and negatively labelled pairs, respectively. In both cases the similarity for green tends to be higher than that for red, as desired. However, in the case of SNLI-Finetuned Skipthoughts this separation is more prominent, possibly justifying the better performance shown in the precision recall curve in Figure 3.

7 Discussion

In this section we discuss a limitation of the datasets used for evaluation. Specifically, both MSRP and SICK have a high mean proportion of common words between the two sentences of the pairs. Further, this average is significantly higher for “positive pairs” (ie. labelled as paraphrases in the case of MSRP, or assigned a high human relatedness score, for the case of SICK), than it is for “negative” pairs. Figure 6 shows the histograms for the distribution of word overlap between pairs of sentences from these datasets, where word overlap for a pair is the proportion of words that are common between its two sentences.

Therefore, it may be inappropriate to draw conclusions on the quality of BOW models merely from their superior performance on these datasets. For example, models like the baseline in Table 6 are not expected to generalize to other tasks. We



(a) MSRP dataset

(b) SICK dataset

Figure 6: Distribution of Pairwise Word Overlap in MSRP and SICK (best viewed in color)

remind the reader that this model represents a sentence as a sum of its word embeddings, which are randomly-generated vectors. We do not expect this to capture any meaningful aspects of sentences, but yet it performs well in this setting. On the other hand, more complicated models like Skipthoughts, SNLI-finetuned Skipthoughts, SNLI RNN, and RNN AE attempt to capture a "latent" notion of relatedness, which does not rely on identifying common words. These models are unfairly penalized in this setting, rendering comparisons between these and BOW models irrelevant.

However, we believe that comparisons between models from within this "more complicated" group are still valid. For example, it is appropriate to compare Skipthoughts with SNLI-Finetuned Skipthoughts on these datasets: they both make the same effort to capture the more abstract sense of relatedness, and are in this sense equally penalized when evaluated on these datasets.

The histograms in Figure 6 underline the need for creation of high-quality datasets to evaluate a model's understanding of "latent" relatedness. For this, we have put together a small number of sentences, grouped into two semantically contradictory groups, as shown in Figure 7. Each model was used to compute the similarities of all pairs and assign ranks in the same way as for paraphrase ranking. The aim is to assign lower ranks (higher similarities) to sentences from the same group as the current sentence, than to sentences from the contradictory group.

Group 1

- There is steam coming out of my soup
- My tongue got burned when I tasted my soup
- I just heated up my soup

Group 2

- My soup is very cold
- My tongue did not get burned when I tasted my soup
- My soup is not hot anymore

Figure 7: Our small dataset for evaluating how well the encoders have captured "latent" relatedness (which does not rely on common word identification)

The results for the ranks and corresponding similarities that sentence "My soup is not hot anymore" assigns to the rest of the sentences are displayed in Table 8.

SNLI-Skipthoughts, SNLI BOW, SNLI RNN and BOW AE all assign the two highest ranks to the remaining sentences of Group 2, as desired. This means that they all have perfect average precision, outperforming Skipthoughts. However, it is more important to examine the "closeness" of these sentences in space (columns "sim" of Table 8). BOW AE, for example, produces a perfect ranking but sentences with ranks 2, 3 and 4 have approximately equal similarity with the "reference" sentence, casting doubts on the quality of the embeddings.

With this in mind, maybe the best performing method for this small group of sentences is SNLI RNN: the similarities between sentences of Group 2 with the reference sentence are much higher than those between Group 1 sentences and the reference sentence. Specifically, there is a gap of 13% between the lowest similarity with a Group 2 sentence and the highest with a Group 1 sentence. SNLI BOW is second best according to this metric, followed by SNLI-Skipthoughts.

This dataset is far too small to draw confident conclusions from but these results may serve as a preliminary indication of the benefit of SNLI information for separating semantically contradictory sentences and understanding "latent" relatedness.

8 Conclusion

In conclusion, we have exploited supervised information from SNLI to enrich the model

Skipthoughts		SNLI-Skipthoughts		SNLI BOW		SNLI RNN		BOW AE	
sim	Sentence	sim	Sentence	sim	Sentence	sim	Sentence	sim	Sentence
0.61	**My soup is very cold**	0.48	**My soup is very cold**	0.59	**My tongue did not get burned ...**	0.51	**My tongue did not get burned ...**	0.51	**My soup is very cold**
0.43	There is steam ...	0.32	**My tongue did not get burned ...**	0.45	**My soup is very cold**	0.47	**My soup is very cold**	0.35	**My tongue did not get burned ...**
0.39	I just heated ...	0.29	I just heated ...	0.39	My tongue got burned ...	0.34	My tongue got burned ...	0.35	I just heated ...
0.39	**My tongue did not get burned ...**	0.29	There is steam ...	0.37	I just heated ...	0.30	I just heated ...	0.35	There is steam ...
0.38	My tongue got burned ...	0.28	My tongue got burned ...	0.32	There is steam ...	0.28	There is steam ...	0.31	My tongue got burned ...

Table 8: Results of the ranking task for the reference sentence **My soup is not hot anymore**. sim refers to the similarity between the reference sentence and the sentence of the corresponding row in embedding space. Sentences which are ”relevant” to this one (**Group 2**), and thus should receive lower ranks, are shown in bold.

space of Skipthoughts, inducing SNLI-Finetuned Skipthoughts. Aside from performing better or on par with Skipthoughts on the supervised evaluations, this model exhibits properties of a superior embedding space. We report results on transfer from SNLI to MSRP in Table 1, and more encouraging results from SNLI to paraphrase ranking in Table 5. We also showed that SNLI-induced embedding spaces capture human intuition about relatedness favorably to other models. Finally, SNLI-finetuned Skipthoughts perform better than its competitors when few or no labelled examples are available for paraphrase detection.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2015. Learning to understand phrases by embedding the dictionary. *arXiv preprint arXiv:1504.00548*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of ALC*.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. *ACL, July*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.