Enforcing Compositionality in Sentence Embedding Models

Arseny Moskvichev
Department of Cognitive Sciences
University of California, Irvine
amoskvic@uci.edu

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine
mark.steyvers@uci.edu

Abstract

Compositionality is considered to be among the defining properties of language, and yet, capturing compositionality remains a struggle for existing sentence embedding models. We present preliminary results on a data augmentation method that helps sentence embedding models to learn the recursive compositional property of language. We add an element to the training objective, which we call compositionality loss: we artificially increase the level of recursive depth in a sentence while maintaining its meaning, and then penalize the encoder model if the embeddings of the augmented and original sentences differ. The proposed method is flexible and can be applied with minimal adjustments to most existing sentence embedding methods. As a proof of concept, we demonstrate the effects of our approach on an example of Sequential Denoising Autoencoder model, where it allows to improve the model's performance on a number of transfer tasks. Although much work is still to be done, we believe that the proposed approach shows promise and will be of interest to the community.

1 Introduction

Sentence-level sequence to sequence models play an important role in a wide range of Natural Language Processing (NLP) applications, including machine translation, plagiarism detection, sentiment analysis, and many other domains. Most of these models operate by constructing a distributed representation of the input sentence, which is subsequently used to solve the task at hand. The problem of constructing broadly applicable sentence embeddings has therefore received considerable attention. The main goal of the present article is to improve general purpose sentence embeddings.

Despite the abundance of impressive practical results, there is still a question of whether existing models are able to understand and use the fundamental properties of language, in particular, compositionality (Marcus, 2018). Compositionality of language refers to the notion that the meaning of a sentence can be acquired by combining the meanings of its constituents. This idea is often believed to be among the cornerstone features underlying rich expressive capabilities of language in general (Fodor, 2001).

Unfortunately, recent findings suggest that existing recurrent neural network-based models still fail to capture compositionality (Lake and Baroni, 2017). In particular, it was shown that recurrent neural network (RNN) architectures, while being able to infer the meaning of examples composed of elements close to those seen during training, still fail to do so when the models have to generalize to instances that more distant from the training data.

Having models that can use compositionality is not only desirable because their behaviour would agree with our intuitions about how language is and should be used. Also, it might prevent NLP systems from failing in many situations. One example of a failure related to compositionality is the precipitous drop in accuracy of the question-answering systems after appending an adversarily constructed irrelevant fact to the end of the paragraph (Jia and Liang, 2017). Such a behaviour can be viewed as an instance of failing to understand the compositional structure of a paragraph. Indeed, the model is looking for a certain word pattern, without being aware that this pattern is, in general, located in an irrelevant part of the paragraph. In our work, we focus on the sentence level compositionality, but a similar approach can be ap-

plied on the paragraph level as well.

Considering the importance of such a property, we believe that if compositionality does not emerge naturally, it might be reasonable to enforce it directly.

One potential solution is to develop new architectures that would by design tend to capture the required property. This approach has already been explored with some success. For example, it is possible to represent each node in a parse-tree as a word-matrix pair, corresponding to the meaning of a node and the effect of the node on the meaning of its neighbours (Socher et al., 2012, 2013). The limitation is, however, that such models tend to be task-specific, so the proposed modifications could not always be readily adapted to other domains and architectures.

Our proposed approach is to rely on existing broadly applicable architectures, but to change the task so that it requires learning compositionality. In a way, we are transforming the task from extrapolation into interpolation. Results obtained by B. Lake (Lake and Baroni, 2017) clearly illustrate a well known idea that for a neural network, it is much easier to generalize to examples that do not differ in systematic ways from what was seen during training. Therefore, if we want the model to learn the compositional property of language, we need to provide plenty of examples, where applying compositionality is necessary to succeed.

Because we expect a phrase "I saw a [boy]" to have the same meaning as "I saw a [male human child]", we might directly penalize the model proportionally to the differences between embeddings of these two sentences. Alternatively, we might augment the data by replacing the instances of "I saw a boy" with "I saw a male human child". In our work, we test both of these alternatives on an example of Sequential Denoising Autoencoder models.

In a way, we want to directly impose a constraint that the meaning of a sentence is independent from its phrasing. A similar attempt was undertaken by Wieting (Wieting et al., 2015), where a paraphrase corpus was used as a source of sentences that express the same meaning in a different form so that the embeddings of corresponding sentences might be constrained to lie closer to each other. Another example is the DictRep method, where a model was trained to map dictionary definitions to the words being defined (Hill et al., 2015). Although demonstrating good performance in many applications, these methods do not directly target compositionality and, more importantly, they rely on the supervised datasets they are trained on, which might limit their generalization ability.

Overall, most successful sentence embedding models take advantage of the vast amounts of available unsupervised data (Kiros et al., 2015; Hill et al., 2016; Wieting et al., 2015). Nevertheless, it might be the case that some important features are not used often enough in natural language for the models to be able to capture these regularities (Hill et al., 2016). It is also possible that not every aspect of meaning might be extracted in an unsupervised manner. This view might be corroborated by the successes of embeddings trained on relatively small amounts of supervised data (Conneau et al., 2017; Triantafillou et al., 2016; Hill et al., 2015; Wieting et al., 2015). We believe that it might be most beneficial to combine supervised and unsupervised knowledge in a flexible manner, and this is exactly what our approach allows to do.

2 Method

2.1 Data augmentation and compositionality loss

The main idea is to constrain the model so that the embeddings of sentences like "I ate an [apple]" (original) and "I ate a [slightly sour round and green fruit]" (expanded) are similar to each other. If we had such an expanded example for every sentence in our training set, we might then simply add a compositionality penalty to our loss function: $L_c \sim d(emb_o, emb_e)$. Where $d(v, u)$ is a cosine distance between vectors v and u, and emb_o and emb_e denote embeddings of the original and expanded sentences.

Alternatively, however, we might simply include the expanded sentences in the dataset and thus force the desired correspondence in an indirect manner.

In order to acquire such pairs of sentences, we used a thesaurus-based data-augmentation approach. Thus, we replaced a certain percentage of words by their thesaurus-based definitions. A similar approach was applied before (leading to a marginal test set performance improvement), but it was limited to replacing certain words with synonyms (Zhang and LeCun, 2015). Our approach has two key differences:

- We replace the words with a complete definition, which means that the resulting sentences are dramatically different in length, while still maintaining the original meaning
- We directly force the embeddings to be the same, instead of only enlarging the dataset

Because many words have more than one meaning, a special care must be taken to ensure that the resulting sentences are coherent and that their meaning remains the same. It is possible to use the disambiguated corpora to ensure correct substitutions, but to keep the procedure as broadly applicable as possible, we only used a part of speech tagger to ensure that the parts of speech are correctly replaced by corresponding definitions from the WordNet database (Fellbaum, 1998). We also excluded names and stop words from potential replacements.

The resulting sentences are far from being grammatically and stylistically perfect, but usually it is still possible to understand their meaning.

Original sentence	Expanded sentence
My dear Mr. Bennet, replied his wife, how can you be so tiresome!	My dear Mr. Bennet, replied his wife, how can you be so lacking in interest as to cause mental weariness!
I started to grab my phone but stopped.	I take the first step or steps in carrying out an action to grab my phone but stopped.
I flipped open the pad and wrote: walking home.	I flipped open the number of sheets of paper fastened together along one edge and produce a literary work: walking home.

Table 1: Expanded sentence examples

As can be seen from the Table 1, there are considerable limitations of such an approach. For example, the WordNet definitions for different verb forms (like "go" and "went") are the same, while in order to correctly expand the sentences, we need to account for the verb inflection.

Moreover, since there is only one definition for every word sense, the model is likely to simply remember to map these definitions to the corresponding words. This issue is, however, mitigated, if we add noise (swapping/removing some words) after original augmentation.

To solve the problem in the long term, we started collecting a crowdsourced database of embeddable definitions, sensitive to such detail. To this means, we constructed an online app which could be accessed at speaktoai.com. We plan to explore the effects of using more nuanced augmentation in subsequent work.

2.2 Performance Evaluation

We evaluated the effects of our proposed data augmentation approach by applying it to the Sequential Denoising Autoencoder (SDAE) model (Hill et al., 2016), with pre-trained GloVe (Pennington et al., 2014) word embeddings. We trained this model on the Toronto Books Corpus [1] for 72 hours. We used the accompanying code for the work by F. Hill and colleagues (Hill et al., 2016) as a basis of our implementation and trained the embeddings using a GPU. Nevertheless, perhaps due to hardware differences or different batch size settings, our model was running considerably slower than reported in (Hill et al., 2016) and thus was trained for less than one full epoch. In particular, for every tested model, we performed 300000 gradient updates with a batch size of 16 sentences.

We then evaluated the resulting embeddings on a range of transfer tasks, using the SentEval library (Conneau et al., 2017).

We tested two variations of our approach:

- using sentence expansion as another noise model in SDAE
- using the compositionality constraint, but retaining the original noize model in SDAE

3 Results

When treated as a regularizer, our approach performed poorly, almost universally resulting in a decreased performance. We, therefore, provide results for the second approach, namely using our augmentation method as a noise model for SDAE.

In the Table 2, comparison with original SAE results (Hill et al., 2016) on unsupervised evaluations is provided. Our model outperforms it in most cases, except for the WordNet. Since we largely rely on WordNet definitions, we believe that this particular metric is not truly representative of the model's performance.

[1] http://yknzhu.wixsite.com/mbweb

	SAE + augment	SAE
Forum	.32/.32	.22/.23
News	.57/.57	.52/.54
Headlines	.52/.50	.41/.41
Images	.64/.63	.64/.64
WordNet*	.48/.52	.60/.55
Twitter	.63/.62	.60/.60
All (weighted avg.)	.54/.54	.42/.43
All (average)	.52/.53	-

Table 2: Unsupervised evaluation results

	SAE + augment	SDAE
MSRP	66.5 / 78.3	73.7 / 80.7
MR	74.0	74.6
CR	77.9	78.0
SUBJ	91.1	90.8
MPQA	87.7	86.9
TREC	79.0	78.4

Table 3: Supervised evaluation results

From the Table 3, we can see that in most cases, performance on the supervised transfer tasks changes only slightly. At the same time, the pattern is different for MSRP (paraphrase detection) and TREC (question type identification) datasets, where the model experiences a dramatic drop in accuracy. Although unfortunate, this result might provide some insight into the features that might be important for these tasks.

4 Conclusions and further research

We proposed a thesaurus-based data-augmentation method and a related regularizer. Together they allow to impose a soft compositionality constraint on the encoder model.

When treated as a noise model for Sequential Denoising Autoencoders, the proposed data augmentation method improves performance on a range of standard transfer tasks. In particular, it leads to improvements on the unsupervised similarity judgements evaluations, which might indicate the increase in interpretability of the acquired embeddings. At the same time, supervised evaluations only show marginal improvements or even a significant drop in performance.

When treated as a regularizer, our approach did not yield many benefits, and, on the contrary, reduced the transfer performance in most cases. Regularization is known to require careful parameter tuning, however, and since every training iteration requires considerable time, we hope to report on the optimal values in subsequent research.

Despite only a partial success, we find these results highly encouraging. Indeed, even by only using WordNet dictionary definitions that lack some of the desired properties, we were able to achieve consistent increase in performance for the SDAE model. A proper implementation of our data augmentation method requires a small amount of supervised data which we are currently collecting through a web application (speaktoai.com). We find it highly plausible that with additional data and minor modifications, the proposed method might become a universally applicable regularization, allowing to introduce compositionality constraints in a broad range of sentence embedding models.

References

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* .

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Jerry Fodor. 2001. Language, thought and compositionality. *Royal Institute of Philosophy Supplements* 48:227–242.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *CoRR* abs/1602.03483. http://arxiv.org/abs/1602.03483.

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

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *CoRR* abs/1707.07328. http://arxiv.org/abs/1707.07328.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *CoRR* abs/1506.06726. http://arxiv.org/abs/1506.06726.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *CoRR* abs/1711.00350. http://arxiv.org/abs/1711.00350.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, pages 1201–1211.

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

Eleni Triantafillou, Jamie Ryan Kiros, Raquel Urtasun, and Richard Zemel. 2016. Towards generalizable sentence embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. pages 239–248.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR* abs/1511.08198. http://arxiv.org/abs/1511.08198.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710* .