Inside Outside Recursive Neural Network for Unsupervised Compositional Semantics Learning

Phong Le

December 4, 2013

1 Introduction

There are many reasons why distributional *lexical* semantics has become popular in computational linguistics. Firstly, it has strong support from not only linguistics but also cognitive science (Lenci, 2008). Secondly, that kind of semantics can be *learned* from unannotated data, which are redundant and free on the Internet, thanks to the flourish of unsupervised learning techniques, such as Latent semantic analysis (Landauer et al., 1998), neural network language modelling (Collobert et al., 2011; Huang et al., 2012), Brown clustering algorithm (Brown et al., 1992), and spectral learning (Dhillon et al., 2012). And finally, it has many applications such as in information retrieval, sentiment analysis, and syntactic parsing.

However, its sister, distributional *compositional* semantics, is still very challenging since available techniques for distributional lexical semantics are not applicable. The core problem lies in the the sparsity of data: there are not enough data (and I believe that we will never be able to collect enough data) since the number of semantically plausible phrases is infinitive. In this report, I will firstly point out hypotheses that are used in many approaches in order to solve the challenge, in Section 2. Then, I will propose a new framework, namely Inside Outside Recursive Neural Network (IORNN), in Section 3, and Dialogue Context Inside Outside Recursive Neural Network (DC-IORNN), in Section 4. In Section 5, I will point out three important applications for that framework.

2 Hypotheses

The *distributional hypothesis* states that (Lenci, 2008)

> The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear.

This hypothesis can be rewritten, in machine learning terms, for lexical semantics learning as follows

> *Hypothesis 1:* The agreement between words and contexts provides enough evidence for lexical semantics learning.

and for phrasal semantics learning as follows

> *Hypothesis 2:* The agreement between phrases and contexts provides enough evidence for phrasal semantics learning.

Unfortunately, it turns out that the hypothesis 2 is useless if it stands alone: we will never collect enough data for learning since the number of well-grammatical phrases is infinitive. Hence, people looked for another direction, which relies on the *principle of compositionality* (Partee, 1995)

> The meaning of a complex expression is a function of the meaning of its parts and of the syntactic rules by which they are combined.

However, the principle itself doesn't point out how to construct compositionality functions, which are the target of many research on distributional compositional semantics.

The most simple approach is use vector addition and multiplication as compositionality functions (Mitchell and Lapata, 2008). Those functions, although no parameters need to be optimized, are too simple to capture real compositionality.

Socher and colleagues propose two neural network frameworks: recursive auto encoder (RAE) (Socher et al., 2011b), (and latter, unfolding RAE (Socher et al., 2011a)) for unsupervised learning, and recursive neural network for supervised learning with task-based training signal (Socher et al., 2010) (e.g., for sentiment analysis, the training signal is the sentiment given by voters) (and later, MV-RNN (Socher et al., 2012a) and RNTN (Socher et al., 2013b)). The key idea of the RAE framework is that: a compositionality function is a compression function, such that an input is able to be recovered from the output by a decompression function. In other words, the following hypothesis is used

> *Hypothesis 3:* Phrases themselves, along with their syntactic structures, provide enough evidence for distributional compositional semantics learning.

Baroni et al. (2012), Grefenstette et al. (2013) and others attempt the challenge in a different way. They use tensors to represent functor words (i.e., verbs, adjectives, etc.), linear maps as compositionality functions, and use contexts for estimating tensors' elements and functions' parameters. Hence, we can say that those approaches make use of Hypothesis 2. Those approaches, although relying on the principle of compositionality, also suffer the sparsity of data because there are too many parameters to optimize.

From those approaches, we can see that there are three main key factors for a successful approach: (i) the expressiveness of compositionality functions, (ii) the ability to overcome the sparsity of data, and (iii), the most important one, the strength of training signal.

In this report, I propose a new framework based on the following hypothesis

> *Hypothesis 4:* The agreement between words and contexts provides enough evidence for compositional semantics learning.

(Note the difference between Hypothesis 4 and Hypothesis 1.) The key idea here is that, compositionality functions are used for composing the meanings of contexts, under the constraint that those meanings are able to be used to predict the target words. This comes from the observation that a human being can guess the meaning of an unknown word given a context around it. In order to do that, he must use the meaning of the context to infer the meaning of the word. Therefore, if he successfully infers the meaning of that word, we can say that, at some degree of certainty, he comprehends the context.

There are three important points. Firstly, although many of those approaches also make use of context, they rely on the 'flat' relationship between words and contexts, which is stored in *co-occurrence* matrices. My framework, in contrast, focuses on the meaning of context and how to construct it. Hence, it can capture stronger constraints between words and contexts and make use of data better in order to overcome the problem of sparsity of data.

Secondly, what my framework gives to a phrase is not only its meaning without context but also the meaning of its context (we call them *inner meaning* and *outer meaning* respectively). Hence, when combining the two kinds of meaning, we can come up with phrasal semantics in context, which is important in information retrieval (Korkontzelos et al., 2013). In addition, we can also compute the semantic plausibility of a phrase in a specific context, which is useful for syntactic parsing (Lazaridou et al., 2013).

Finally, my framework is general in the sense that, it can make use of any neural net framework proposed by Socher and colleagues. For the RAE framework, the training signal is now even stronger since it uses not only phrases and their syntactic structures but also the agreement between words and contexts. For the RNN framework, we now can train RNNs in an unsupervised learning manner. This framework, therefore, possesses those three key factors.

3 Inside Outside Recursive Neural Network (IO-RNN)

In this section, I will present my framework. In order to be simple, RNN is used. However, as I mentioned before, any networks which can process tree structures are also applicable.

3.1 Network Structure

Given a constituent and its tree structure (like the one in Figure 1), each node u is assigned two vectors $\mathbf{o}_u$ and $\mathbf{i}_u$. The first one, called *outer meaning*, denotes the meaning of the context; the second one, called *inner meaning*, denotes the meaning of the phrase that the node covers.

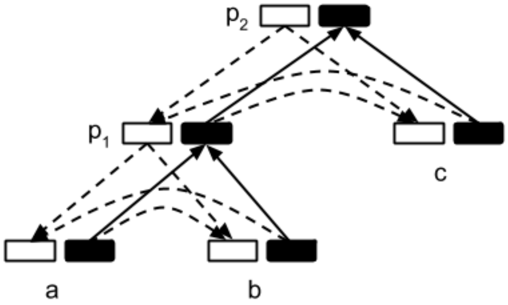

Figure 1: Inside-Outside Recursive Neural Network. Black rectangles correspond to inner meanings, white rectangles correspond to outer meanings.

Word embeddings (e.g., $\mathbf{i}_a$) Similar to Socher et al. (2010), and Collobert et al. (2011), given a string of binary representations of words $(a, b, ..., w)$ (i.e., all of the entries of w are zero except the one corresponding to the index of the word in the dictionary), we first compute a string of vectors $(\mathbf{i}_a, ..., \mathbf{i}_w)$ representing inner meanings of those words by using a look-up table (i.e., word embeddings) $\mathbf{L} \in \mathbb{R}^{n \times |V|}$, where $|V|$ is the size of the vocabulary and n is the dimensionality of the vectors. This look-up table $\mathbf{L}$ could be seen as a storage of lexical semantics where each column is a vector representation of a word. Hence,

$$\mathbf{i}_w = \mathbf{L}w \in \mathbb{R}^n \tag{1}$$

Computing inner meaning The inner meaning of a non-terminal node, say p_1, is given by

$$\mathbf{i}_{p_1} = f(\mathbf{W}_1^i \mathbf{i}_a + \mathbf{W}_2^i \mathbf{i}_b + \mathbf{b}^i) \tag{2}$$

where $\mathbf{W}_1^i, \mathbf{W}_2^i$ are $n \times n$ real matrices, $\mathbf{b}^i$ is a bias vector, and $f(.)$ is an activation function, e.g. *tanh* function. Intuitively, the inner meaning of a parent node is the function of the inner meanings of its children. This is similar to what Socher et al. (2010) call recursive neural network.

Computing outer meaning The outer meaning of the root node, $\mathbf{o}_{root}$, is initially set randomly, and then learnt later. To a node which is not the root, say p_1, the outer meaning is given by

$$\mathbf{o}_{p_1} = g(\mathbf{W}_1^o \mathbf{o}_{p_2} + \mathbf{W}_2^o \mathbf{i}_c + \mathbf{b}^o) \tag{3}$$

where $\mathbf{W}_1^o, \mathbf{W}_2^o$ are $n \times n$ real matrices, $\mathbf{b}^o$ is a bias vector, and $g(.)$ is an activation function, e.g. *tanh* function. Informally speaking, the outer meaning of a node (i.e., the meaning of its context) is the function of the outer meaning of its parent and the inner meaning of its sister.

The reader, if familiar with syntactic parsing, could recognizes the similarity between Equation 2, 3 and the inner, outer probabilities given a parse tree

$$P_{in}(A,r,t) = P(A \rightarrow B\ C)P_{in}(B,r,s)P_{in}(C,s,t) \tag{4}$$

$$P_{out}(B,r,s) = P(A \rightarrow B\ C)P_{out}(A,r,t)P_{in}(C,s,t) \tag{5}$$

Therefore, I name the framework Inside-Outside Recursive Neural Network.

3.2 Learning

The learning is based on the Hypothesis 4. That is there must be a strong correlation between $\mathbf{o}_w$ and $\mathbf{i}_w$ where w is any word in a given sentence. The simplest way to train the network is to force $\mathbf{o}_{w_j} = \mathbf{i}_{w_j}$; hence, learning is to minimize the following loss function

$$J(\theta) = \sum_{s \in D} \sum_{w \in s} \|\mathbf{o}_w - \mathbf{i}_w\| \tag{6}$$

where D is a set of training sentences and θ are the network parameters. However, that could be problematic because the meaning of context is not necessary the meaning of the target word.

Here, based on the observation that the meaning of context sets constraints on selecting a word to fill in the blank, I put a *softmax* neuron unit on the top of each $\mathbf{o}_w$ in order to predict what w is

$$\hat{w} = softmax(\mathbf{W}^l \mathbf{o}_w + \mathbf{b}^l) \tag{7}$$

where $\mathbf{W}^l$ is a $n \times |V|$ matrix, and

$$softmax((z_1, ..., z_n)^T) = \frac{1}{\sum_j e^{z_j}} \left(e^{z_1}, ..., e^{z_n}\right)^T \tag{8}$$

And therefore, the loss function needs to be minimized is a cross entropy loss function

$$J(\theta) = \sum_{s \in D} \sum_{w \in s} \sum_{j} -w_j log(\hat{w}_j) \tag{9}$$

Intuitively, the loss function computes the word prediction error of the net over all words of all training sentences. The gradient $\frac{\partial J}{\partial \theta}$ could be effectively computed thanks to backpropagation through the structure (Goller and Kuchler, 1996).

4 IO-RNN with Dialogue Context (DC-IO-RNN)

It turns out that is is easy to extend the framework above to dialogue context, and I call the new framework DC-IO-RNN. Figure 2 illustrates how to connect inner and outer meanings of sentences in a dialogue. Intuitively, the outer meaning of a sentence is the function of the inner meanings of its neighbour sentences.

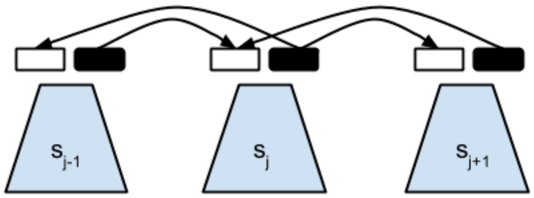

Figure 2: Inside-Outside Recursive Neural Network with Dialogue Context. Black rectangles correspond to inner meanings, white rectangles correspond to outer meanings.

5 Applications

5.1 Syntactic Parsing

Lazaridou et al. (2013) point out that *semantic plausibility* is useful for discriminating correct from incorrect syntactic structures. The question is what could be used to compute semantic plausibility. Here, I believe that the agreement between inner meaning and outer meaning could be used as a measure for that. And therefore, we can make use of reranker methods, such as ones proposed by Socher et al. (2013a) and Le et al. (2013).

5.2 Word Meaning in Context

Current research on word meaning in context (Huang et al., 2012; Thater et al., 2011; Dinu and Lapata, 2010; Erk and Padó, 2008; Dinu et al., 2012) focuses on 'flat' relations of target words and their context, e.g. relations of target words and individual neighbour words. In contrast, my framework computes the meaning of context and hence we can make use of the fact that context words interact with each other in order to shift the meaning of the target word.

5.3 Sentiment Analysis

Socher et al. (2012b) conclude that pretraining (in a unsupervised learning manner) a MV-RNN does not help to improve the final performance on sentiment analysis since "antonyms often get similar vectors [...] due to high similarity of local syntactic contexts". That is true if sentences are processed independently. However, if all sentences in a dialogue or paragraph are processed at the same time by DC-IO-RNN (see Section 4), antonyms could be well distinguishable. The reason is that, in a positive/negative paragraph, positive/negative words tend to occur together, though in different sentences, and DC-IO-RNN is hopfully able to capture that.

5.4 Paraphrasing

[???]

5.5 Dialogue Modeling with DC-IO-RNN

[???]

References

Baroni, M., Bernardi, R., and Zamparelli, R. (2012). Frege in space: A program for compositional distributional semantics.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:24932537.

Dhillon, P., Rodu, J., Foster, D., and Ungar, L. (2012). Two step cca: A new spectral method for estimating vector models of words. *arXiv preprint arXiv:1206.6403*.

Dinu, G. and Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172. Association for Computational Linguistics.

Dinu, G., Thater, S., and Laue, S. (2012). A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 611–615. Association for Computational Linguistics.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics.

Goller, C. and Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. In *Neural Networks, 1996., IEEE International Conference on*, volume 1, page 347352.

Grefenstette, E., Dinu, G., Zhang, Y.-Z., Sadrzadeh, M., and Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Korkontzelos, I., Zesch, T., Zanzotto, F. M., and Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lazaridou, A., Vecchi, E. M., and Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help np parsing. *Proceedings of EMNLP, Seattle, WA*.

Le, P., Zuidema, W., and Scha, R. (2013). Learning from errors: Using vector-based compositional semantics for parse reranking. In *Proceedings of the 1st ACL Workshop on Continuous Vector Space Models and their Compositionality*, page 11.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):131.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, page 236244.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311360.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013a). Parsing with compositional vector grammars. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011a). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems*, 24:801809.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012a). 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*, page 12011211.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012b). 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*, pages 1201–1211. Association for Computational Linguistics.

Socher, R., Manning, C. D., and Ng, A. Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011b). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 151161.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *IJCNLP*, pages 1134–1143.