Continuous Word and Phrase Representation with Semantic Structure

Kushal Arora, karora@cise.ufl.edu

March 3, 2015

Abstract

A lot of recent supervised NLP systems use the unsupervised word embeddings to improve generalization. Most of these models use single word representation which fails to capture both the compositionality and longer context dependencies which are inherent to languages. One of the ways to overcome this lacunae is to learns both word and phrase embeddings as well as their compositional semantics. This proposal puts forward a recursive model which learns these phrase embedding as well as compositional operator and probability measure on them. To capture the context dependecies, the model proposes explicit usage based clustering of words and phrases similar to class based models. The model will be evaluated for regularities using analogy task and componsitionality through perplexity on unknown sentences.

Motivation

What are the characteristics of a ideal continous space language model? This paper proposes the following 1.) good generalization for unknown sentences, 2.) compositionality i.e. ability to discover phrase embedding using sub-phrases embedding and 3.) linguistic regularities among embeddings like clustering of semantically similar words and phrases and symmetry for syntactically similar ones. We will now consider each one of these characteristics in details and see how the proposed model enforces these properties.

Linguistic Regularity

Most of the continuous models are able to capture the linguistic regularities to some extent. These properties have been evaluated by [14, 9, 5] using syntactic and semantic analogy tasks. These models learn regularities implicitly with focus on improving next word prediction. Class based models as proposed by [2] learn regularities explicitly in attempt to improve generalization of n-gram models. Proposed model attempts to capture the essence of the class based models in NNLM setting. This is done by clustering words and phrases which

are used in same context. These constraints would lead to representation which should performs better on analogy tasks and provide better results for standard NLP tasks like NER, Chunking and POS tagging.

Compositionality

Languages are recursive in nature. Words combine to build phrases and phrases, to build sentences. The proposed model attempts to capture this recursive compositionality by learning operators on words and phrases. This can help us in representing unknown sentences and phrases in embedding space and assign them a probability. At the same time regularity constraint on phrases can help us infer the meaning of the phrase using the neighboring embeddings. Similar compositional and multi-word embedding approach has been tried by Socher et al[11, 13, 12]. The proposed model differs from their approach mainly in two ways. Firstly, their model focuses on supervised sentiment classification whereas the attempt here is to learn unsupervised word embeddings. Secondly, they learnt compositionality and embedding of only those phrases which maximized their objective. On the contrary, we attempt to embed all sub-phrases in the continous space. The proposed model is more general in its attempt and produces embeddings and operators which can be used for supervised tasks including sentiment classification.

Generalization

The standard n-gram models represented words and phrases in a discrete space. This prevents the true interpolation of probabilities of unseen n-grams since a change in this word space can result in an arbitrary change of the n-gram probability. Continuous space approaches tries to solve this problem by projecting words in continuous space and using a probability estimator on it. This leads to better generalization as shown by [1, 7, 10]. This paper proposes a way to improve generalization further using the regularities and compositionality. By assigning propobabilities to multi-word embeddings and using compositionality to learn probability on unknown sentences, the proposed model can improve state of the art results in language modelling.

Model Architecture

The approach is to build a model that can be trained layer-wise to get best representation for the composed phrases. Each layer in model corresponds to phrases of length l built by appending a word to the sub-phrase of length $l-1$. Other way of seeing this is that, at each layer a composition operator $g_l(\cdot)$ uses two sub-phrases of length $l-1$ and 1 to compose a phrase of length l. Each operator returns two things an embedding for the phrase and its probability. The probability of words at layer 0 are computed from the corpus. Figure 1 shows the composition architecture for an example sentence.

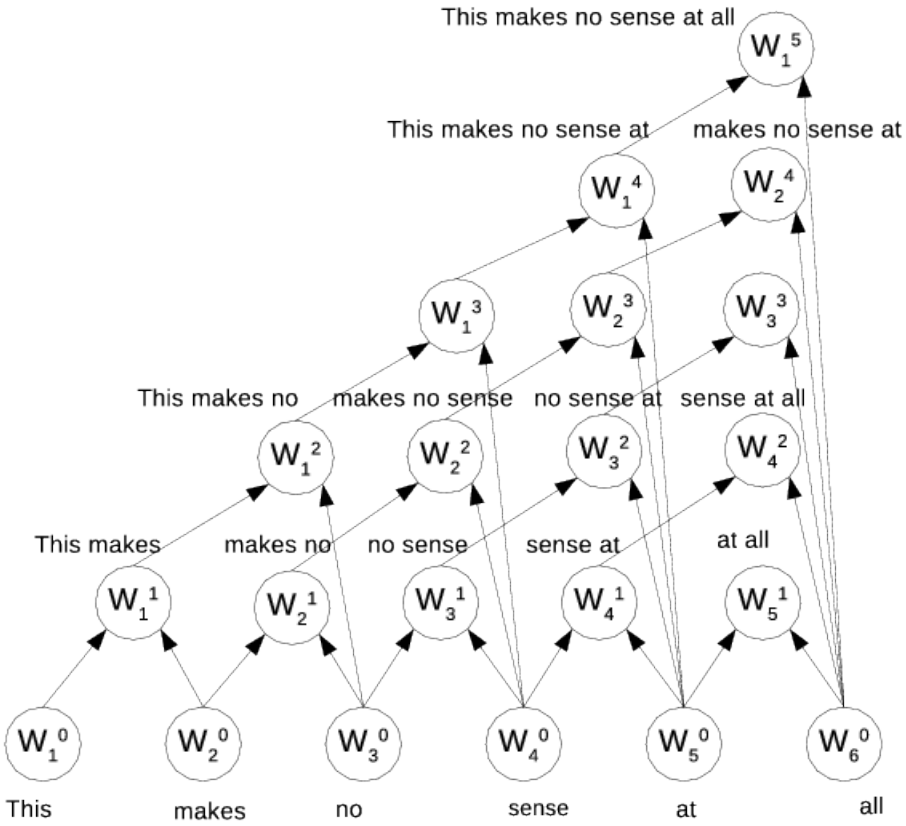

Figure 1: Composition network of a sentence

Let N be the maximum number of words a sentence model can handle. Let w_i^l, x_i^l be the ith phrase of length l and its embedding. The length of lth layer is $N - l$ nodes. Each layer is parameterized by a matrix W_{dX2d}^l, where d is the dimension of the embedding space. Composition function for the phrase x_i^l is as

$$x_i^l = g(x_i^{l-1}, x_{i+l}^1; W^l) \tag{1}$$

The unnormalized probability of x_i^l can be is impacted by the prior probability of $x_i^l$and x_{i+l}^1 and how these two come together to form x_i^l. The composition factor is captured by the parameter W_{dXd}^{prob}. Unnormalized probability is calculated in the following way

$$\tilde{p}_i^l = f(x_l^l, x_{i+l}^1, p_i^{l-1}, p_{i+l}^1; W^{prob}) \tag{2}$$

Proposed function for $g(\cdot)$ and $f(\cdot)$ are *tanh* and *sigmoid*. All the embeddings are initialized using normal distribution $\mathcal{N}(0, 1)$. Probability for words in layer 0 is the calculated from the corpus and is kept fixed. Rewriting 1 and 2

$$x_i^l = tanh(W^l \begin{bmatrix} x_i^{l-1} \\ x_{i+l}^1 \end{bmatrix}) \tag{3}$$

$$\tilde{p}_i^l = sigmoid(x_i^{l-1} \cdot W_{prob} \cdot x_{i+l}^1 * p_i^{l-1} * p_{i+l}^1) \tag{4}$$

The training is done in a layer-wise fashion by minimizing the cross entropy between the original distribution of phrases of length l and probability distribution of the composed phrases. The original distribution is calculated using n-grams from the corpus.

Objective Function

The objective function is composed in such a way that all three characteristics we defined for ideal language model are explicitly expressed. In this section, composition of the objective function is discussed step by step incorporating each of these properties.

Let S be the training corpus, w^l_{ji} be the ith word of the jth sentence of the corpus. Let W^l be set of the all phrases of length l and X^lbe the set of embedding of phrases of length l. Let $P(w^l_{ji})$ be the original probability of w^l_{ji}. The loss function for cross entropy minimization function for layer l is defined as

$$\mathcal{L}^l_1(\theta) = \sum_{j=1}^{|S|} \sum_{i=1}^{N-l} H(P(w^l_{ji})||p^l_{ji};\theta) \tag{5}$$

As the model is learns the unnormalized probability, an additional normalization constraint $\sum_{x^l \in X^l.} \tilde{p_{x^l}} = 1$ is enforced. So the objective function now becomes

$$\mathcal{L}^l_1(\theta) = \sum_{j=1}^{|S|} \sum_{i=1}^{N-l} H(P(w^l_{ji})||p^l_{ji};\theta) + \sum_{w^l \in W^l.} \tilde{p_{w^l}} = 1 \tag{6}$$

A dense embedding space would lead to a better generalization. To achieve this, this model minimize the spread of lth layer by reducing the distance of individual point from the centroid of the layer. Let $\tilde{x^l}$be the centroid of the embedding of length l phrases.

$$\mathcal{L}^l_2(\theta) = \sum_{j=1}^{|S|} \sum_{i=1}^{N-l} ||x^l_{ji} - \tilde{x}^l;\theta||^2 \tag{7}$$

One of the properties of the ideal embedding would be meaningful regularities as perceived by us. For example, we perceive words and phrases used in same context to be similar or close. Taking cue from that, this model implicitly learns syntactic and semantic relations like synonyms, plural and singular forms,and possessive and non possessive forms using an explicit clustering objective. This is done by clustering all the words used in conjugation with a phrase and all phrases used in conjugation with a word in embedding space. Let $m_{x^l/x^{l-1}}$be the mean of embeddings of all the words(length 1) used with phrase x^{l-1} to create phrase in X^l and m_{x^l/x^1} be mean of embeddings of all the phrases used with the word x^1 to form phrase in X^l.

$$\mathcal{L}_3^l(\theta) = \sum_{x^{l-1} \in X^{l-1}} \sum_{\substack{x^1 \in X^1 \\ g(x^{l-1}, x^1) \in X^l}} ||x^1 - m_{x^l / x^{l-1}}; \theta||^2 + \sum_{x^1 \in X^1} \sum_{\substack{x^{l-1} \in X^{l-1} \\ g(x^{l-1}, x^1) \in X^l}} ||x^{l-1} - m_{x^l / x^1}; \theta||^2 \tag{8}$$

Finally, bringing all of it together, we get

$$\mathcal{L}^l(\theta) = \mathcal{L}_1^l(\theta) + \lambda_1 \mathcal{L}_2^l(\theta) + \lambda_2 \mathcal{L}_3^l(\theta) \tag{9}$$

The parameter θ we train on are

$$\theta^l = [W^{prob}, W^l, X^{l-1}, X^1]$$

We can layer-wise train for each layer l and use the minimizing negative log likelihood of training sentences as the stopping criteria.

Evaluation Criteria

We evaluate our model on two type of tasks. The first type is the standard task of evaluating the perplexity of the unseen test corpus. This tasks test the compositionality and generalization abilities of our model. We will report the use perplexity as the measure in this type of evaluation as done by [1, 7, 8, 10] The second type is to evaluate semantic and syntactic ability of words and phrase embeddings. WordSim-353[3] is one of most used dataset for the evaluating similarities among the words. Other similarity tasks/datasets that can be evaluated are SCWS[4] and RW[6]. We can also use dataset provided by [9].

References

[1] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[2] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[3] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001.

[4] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 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, 2012.

[5] RichardSocher JeffreyPennington and ChristopherD Manning. Glove: Global vectors for word representation.

[6] Minh-Thang Luong, Richard Socher, and Christopher D Manning. Better word representations with recursive neural networks for morphology. *CoNLL-2013*, 104, 2013.

[7] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[9] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013.

[10] Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 723–730. Association for Computational Linguistics, 2006.

[11] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 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, 2012.

[12] Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pages 1–9, 2010.

[13] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

[14] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.