Syntagmatic Word Embeddings for Unsupervised Learning of Selectional Preferences

Renjith P Ravindran, Akshay Badola and **Kavi Narayana Murthy**
School of Computer and Information Sciences
University of Hyderabad
{rpr,badola}@uohyd.ac.in
knmuh@yahoo.com

Abstract

Selectional Preference (SP) captures the tendency of a word to semantically select other words to be in direct syntactic relation with it, and thus informs us about syntactic word configurations that are meaningful. Therefore SP is a valuable resource for Natural Language Processing (NLP) systems and for semanticists. Learning SP has generally been seen as a supervised task, because it requires a parsed corpus as a source of syntactically related word pairs. In this paper we show that simple distributional analysis can learn a good amount of SP without the need for an annotated corpus. We extend the general word embedding technique with directional word context windows giving word representations that better capture syntagmatic relations. We test on the SP-10K dataset and demonstrate that syntagmatic embeddings outperform the paradigmatic embeddings. We also evaluate supervised version of these embeddings and show that unsupervised syntagmatic embeddings can be as good as supervised embeddings. We also make available the source code of our implementation[1].

1 Introduction

Selectional Preference (SP) (Wilks, 1975) encodes the syntagmatic relatedness between two words. Relations between words are either syntagmatic or paradigmatic (de Saussure, 1916). Two words are said to be paradigmatically related if one word can replace the other in a sentence. Words belonging to a narrow semantic class, such as 'cat', 'dog' can often be substituted with each other in a sentence. Syntagmatic relations are between syntactically related co-occurring words in a sentence. Such word relations encode both syntactic and semantic aspects of words. A *noun* may be modified by an *adjective*, but any particular instance of a noun tends to go more with some adjectives than others. For example *black dog* is more likely than *green dog*. SP deals with such semantic preferences between syntactically related word pairs. Common SP relations include 'adjective-noun', 'subject-verb', 'verb-object'. SP finds use in important NLP tasks like sense disambiguation (Resnik, 1997), semantic role classification (Zapirain et al., 2013), co-reference resolution (Hobbs, 1978; Zhang et al., 2019c), etc.

A computational method to induce SP from instances of syntactically related word pairs in a parsed corpus was introduced by Resnik (1996). In order to generalize to unseen data, this method made use of ontological classes obtained from WordNet (Miller, 1995). Rooth et al. (1999) showed that the dependence on external knowledge resources could be removed by learning the classes from the corpus itself using the EM algorithm. Erk (2007) showed that generalization is also possible via co-occurrence similarity between seen and unseen words. SP models are usually evaluated using the Pseudo-word Disambiguation task (Van de Cruys, 2014) which requires the identification of the more probable dependent word, from a less probable (random) word, given the head word and a syntactic relation. The dataset is generally created from the unseen part of a parsed corpus used for learning the model. Therefore this task measures only how well the model fits the corpus, which may be biased, and not how well it learns SP as perceived by humans. Recently, Zhang et al. (2019b) introduced SP-10K, a dataset for SP evaluation across 5 syntactic relations with a total of 10,000 items each with a human-annotated plausibility score. SP-10K measures the correlation between a model's SP score for a given word pair and the average human score. Therefore it is a better test for SP learning.

[1] https://github.com/renjithravindran/spvec

Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 213–222
Bangkok, Thailand (Online), August 6, 2021.

The current state-of-the-art on SP-10K is reported by Multiplex Word Embeddings (MWE) (Zhang et al., 2019a). It is a negative sampling based word embedding model, trained on relation-specific word pairs from a parsed corpus. Compared to unsupervised embedding models such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), MWE provides a substantial boost in SP learning as it has access to syntactic relations. It also improves over D-embeddings (Levy and Goldberg, 2014a) which is a supervised embedding model. However, a dependency-parsed corpus is not readily available in many languages. Therefore the need for an effective unsupervised SP induction technique is palpable in the wider NLP community.

In this work we show that unsupervised word embeddings can easily be extended to get better at learning SP. We do this by taking directional (left/right) word context windows unlike symmetric windows of Word2vec, GloVe, etc. Having directional context windows gives two embeddings per word, one of its left context and other of its right context. This allows us to approximate syntactic relations with directions; all relations that happen to the left of a word are captured by the left embedding and those that happen to the right of a word are captured by the right embedding. Then the cosine similarity between the right embedding of a word and left embedding of another word indicates how likely the two are to be syntagmatically related.

In summary, our contributions are: 1) We provide a simple and effective method to capture selectional preference, called syntagmatic embeddings 2) Demonstrate that syntagmatic embeddings are superior to paradigmatic embeddings 3) We also show that our unsupervised syntagmatic representations can be as good as their supervised counterparts, therefore showing that a good range of SP information can be learned even without a dependency-parsed corpus.

2 Syntagmatic Representation

Symmetric and non-directional context windows in embedding techniques, such as GloVe, relate words that have similar (paradigmatic) contexts. Context words are other words that are in the immediate vicinity of a target word. A symmetric window considers equal number of words on the left and right as context words. Though syntagmatically related words may have similar contexts, a symmetric window tends to encode more of paradigmatic relations. But these paradigmatic embedding spaces do encode syntagmatic properties to a certain degree. For example, we may find that the cosine similarity between 'coffee' and 'cup' is generally greater than 'coffee' and 'car'. These embeddings are considered unsupervised as they are learned from a plain un-annotated corpus. Since their contexts are not dictated by syntactic relations they are generally inferior, at learning SP, compared to an embedding technique that has access to such information (Zhang et al., 2019a). Also, there is no direct way to extract syntagmatically related words. The nearest neighbours of a given word will largely be all paradigmatically related. Though it may include, given a larger context window, associated words ('coffee', 'cup') which have a syntagmatic nature.

2.1 Relations as Directions

Exact learning of SP requires word co-occurrence in a sentence to be defined as a pair of syntactically related words, which is available only in a dependency-parsed corpus. We can obtain a less exact representation for SP by replacing syntactic relations with directions, because in word-ordered languages, word-order or direction plays a major role in assigning syntactic relations. For example in an English sentence, the adjectival modifier of a noun is always found to its left. The nominal subject of a verb is found to its left and direct object to its right. The technique explored here exploits this fact to learn a substantial amount of selectional preference without the need for a large dependency-parsed corpus.

2.2 Unweighted Factorisation Model

Word embeddings are low-rank representations of row/column vectors in a word co-occurrence matrix (Levy and Goldberg, 2014b). Here, we consider unweighted factorisation of a word co-occurrence matrix using Truncated Singular Value Decomposition (SVD) (Kalman, 1996). Let M be the co-occurrence matrix of size $v \times v$, where v is the size of the vocabulary. Instead of a symmetric context window, we use non-symmetric and directional windows, directions being *left* and *right*. Let $M_{i,j}$ be the number of times word i co-occurred to the left of word j within a distance of k throughout the corpus, where k is the size of the co-occurrence window. Consequently, $M_{j,i}$ becomes the number

word	associations
car	**left:** vintage, second-hand, oncoming, luxury, buying, toy, saloon, buy, mercedes...
	right: collided, sped, exploded, maker, skidded, swerved, belonging, makers, roared...
eat	**left:** want, wants, going, wanting, let, tend, ought, let's, allowed, prefer, supposed, able...
	right: salad, beans, soup, cakes, pork, peas, bacon, pasta, fresh, pie, biscuits...
blue	**left:** wore, vivid, dull, wear, luminous, wears, dazzling, plain, dim, dressed, dyed...
	right: scarf, stripe, livery, robe, beret, overalls, blazer, slacks, gloves...
aggressive	**left:** increasingly, extremely, equally, become, very, highly, particularly, becoming...
	right: behaviour, attitude, manner, response, towards, tactics, stance, attack, actions...

Table 1: Examples of word associations from syntagmatic embeddings.

of times word i co-occurred to the right of word j. Thus the row i of matrix M gives the representation of word i using its left context words. And column j gives the representations of word j using its right context words. These two representations are different because our co-occurrence matrix is not symmetric. However, raw co-occurrence representation is very high-dimensional, highly sparse and noisy. A major component of word embedding techniques is dimensionality reduction, by approximating the original co-occurrence matrix with its low-rank representation $\hat{M}$. Dimensionality reduction is found to reduce noise in the data matrix by eliminating the low principle components of the data, thus increasing generalisation. We use Truncated SVD [2] to obtain rank d approximation. Equation 1 gives the factorisation of the matrix M.

$$M \sim \hat{M} = \hat{U}\hat{S}\hat{V}^\intercal \tag{1}$$

Where, $\hat{U}_{v\times d}$, $\hat{S}_{d\times d}$, $\hat{V}_{v\times d}$ are the factor matrices (singular vectors and singular values) obtained in SVD as, $M = USV^\intercal$, but truncated to keep only the top d principle components. $\hat{U}$ and $\hat{V}$ gives the left context and right context representations of words respectively, in terms of the leading d singular vectors. The singular values $\hat{S}$ gives the relative weightage of corresponding singular vectors, which may be used to scale the singular vectors appropriately. Our word representations are obtained by scaling the singular vectors by an exponential factor of their singular values. Thus, the final left embedding is given as $L = \hat{U}\hat{S}^p$ and the right embedding is $R = \hat{S}^p\hat{V}^\intercal$. Caron (2001) showed that the exponential weighting factor p allows for a softer rank selection such that $p > 0$ gives more weightage to the leading components and $p < 0$ gives weightage to the lower components, allowing the fine tuning of embeddings for different tasks. The number of components (dimension), exponential weighting factor, and co-occurrence window size are three important parameters that influence the performance of these embeddings. Our experiments include yet another parameter, the term-weight. So far we have assumed that M contains raw co-occurrence values, or the frequency count of two words to co-occur in the corpus. Various term-weighting schemes can be applied to transform the raw frequencies. We experiment with *log*, *PMI* (Point-wise Mutual Information) and *PPMI* (Positive Point-wise Mutual Information) term-weights along with the raw frequency counts.

[2] *randomized_svd* from *scikit-learn*

2.3 Weighted Factorisation Model

A factorisation model like the one presented in the previous section gives equal weightage to all errors in the low-rank approximation process. It has been shown that weighting errors from each co-occurrence term, by a function of their co-occurrence frequency yields better word embeddings (Levy and Goldberg, 2014b). Neural embedding techniques such as Word2vec do such weighting implicitly (Levy and Goldberg, 2014b), whereas techniques that makes use of co-occurrence matrix, such as GloVe, do this explicitly. For evaluating the performance of weighted factorisation on selectional preference, we minimally modify the GloVe model to get syntagmatic embeddings.

$$\mathbb{L} = \sum_{i,j=1,1}^{V,V} f(M_{i,j})(\mathbf{u}_{w_i}^\intercal \mathbf{v}_{w_j} + b_i + b_j - log\, M_{i,j})^2 \tag{2}$$

Equation 2 gives the loss function $\mathbb{L}$ for approximating the *log* co-occurrence with the dot product

of the left embedding ($\mathbf{u}_w$) and the right embedding ($\mathbf{v}_w$). M here is the co-occurrence matrix and b_i, b_j are bias terms. With symmetric context, the final embeddings in the GloVe model are either just the left embeddings or the sum of left and right embeddings. But with asymmetric context, left and right embeddings are used distinctly. The weighting function (f) is given by equation 3.

$$f(x) = \begin{cases} (x/x_{max})^{\frac{3}{4}}, & \text{if } x < x_{max} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

x_{max} is generally taken as 100. GloVe's weighting function mainly reduces the influence of rarely co-occurring words which tend to be noisy.

2.4 Syntagmatic Association

Let $\overleftarrow{l_i}$ be the left embedding of word i , i.e. i^{th} row of L, and $\overrightarrow{r_j}$ be the right embedding of word j, i.e the j^{th} column of R. Since $\overrightarrow{r_j}$ reflects the right context of word j and $\overleftarrow{l_i}$ reflects the left context of word i, similarity between $\overrightarrow{r_j}$ and $\overleftarrow{l_i}$ would reflect how often word j is found to the left of word i. Thus cosine similarity between $\overrightarrow{r}_j$ and $\overleftarrow{l_i}$ captures the association of word j to the left of word i, and the association of word i to the right of word j.

Table 1 gives few examples of left and right associations from syntagmatic embeddings. These examples have been filtered to remove words that tend to appear as both left and right associates. Let $\mathbf{l}$ and $\mathbf{r}$ be the set of left associates and right associates of a given word in the embedding space, then the examples given here are $\mathbf{l} - \mathbf{r}$ (left) and $\mathbf{r} - \mathbf{l}$ (right). We see that the left associates of a noun (car) tends to have adjectives (vintage) and verbs (buy) that take the noun as its direct object. Right associates of the noun are found to be verbs (collided) that take the noun as its subject. With a verb (eat) we see that its left associates are other verbs (want) to which the given verb is an open clausal component. The right associates are its direct objects (salad). With an adjective (blue) we see that its left associates are other adjectives (vivid) that act as intensifiers and verbs (wore) whose direct objects are modified by the given adjective. The right associates are nouns (scarf) that are modified by the adjective.

3 SP Evaluation

Examples of word association in the previous section gives a qualitative feel about the degree to which syntagmatic embeddings can capture selectional preference. In the next section we follow this up with detailed analysis using quantitative studies.

	head	dependent	human-score
amod	air	fresh	9.7
	number	medium	4.0
	wind	secret	0.7
dobj	eat	meal	10.0
	touch	food	5.5
	eat	mail	0.0
nsubj	sing	singer	10.0
	pray	woman	5.8
	eat	textbook	0.0

Table 2: Samples from SP-10K dataset.

3.1 Dataset

We use the SP-10K (Zhang et al., 2019b) dataset to quantify the correlation of between the SP information learned by our syntagmatic embeddings and that of human judgements. Other datasets with human scores for SP are McRae et al. (1998); Keller and Lapata (2003); Padó et al. (2006). But compared to SP-10K these are much smaller in size. SP-10K has 3 direct relations and 2 indirect relations. For our evaluation we only use the direct relations – **amod**, **nsubj** and **dobj**. In SP-10K there are 2000 evaluation instances under each relation class. Each instance is a triplet (*word1*, *word2*, *human-score*), where *word1* is the head and *word2* is a dependent, and *human-score* gives the plausibility of *word2* being dependent on *word1*, via the given relation, as judged by humans on a 0-10 scale. For **amod** relation, a noun is the head and an adjective is the dependent. For **nsubj** and **dobj** a verb is the head and a noun is the dependent. Table 2 gives some examples from the dataset. The model's capacity for SP is judged by the correlation (Spearman's) between the association score given by the model and the human-score. The model-score for a given head-dependent pair is the cosine similarity between the head and the dependent in the embedding space.

Since the syntagmatic embeddings relegate relations to left and right directions, the cosine similarity for each of the relations are computed as: **amod**: $\overrightarrow{r_d} \cdot \overleftarrow{l_h}$, **nsubj**: $\overrightarrow{r_d} \cdot \overleftarrow{l_h}$, **dobj**: $\overrightarrow{r_h} \cdot \overleftarrow{l_d}$, where subscript h and d denotes head and dependent words respectively, and symbol '$\cdot$' denotes cosine similarity.

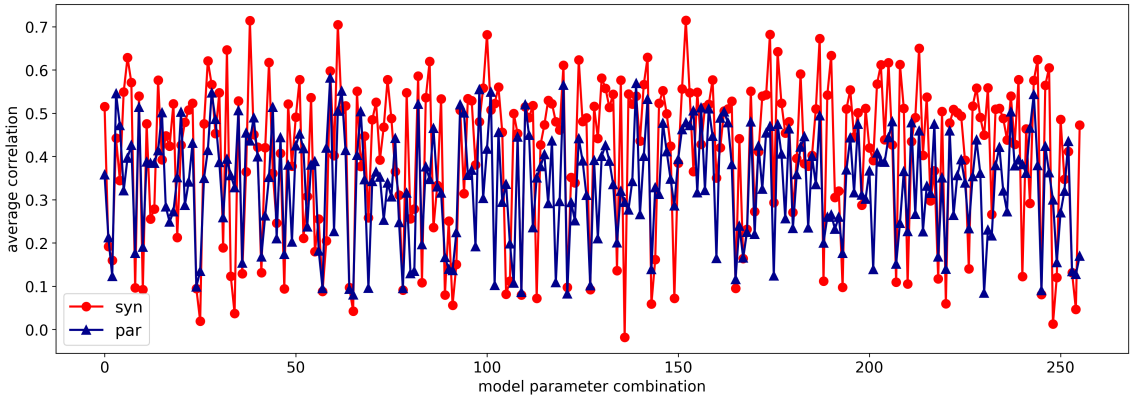

Figure 1: Average correlation of syntagmatic and paradigmatic models over various parameter combinations.

3.2 Baseline Models

We compare our syntagmatic model with 3 paradigmatic models: Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and DSG (Song et al., 2018). Both Word2vec (w2v) and GloVe (glove) are typical paradigmatic embeddings. DSG (Directional Skip-Gram) is a variant of Word2vec that claims to encode directional information by predicting the co-occurring words and also their directions. However, unlike syntagmatic embeddings DSG gives only one embedding per word. The best reported supervised model on SP-10K is Multiplex Word Embeddings (MWE). However, we could not use [3] the available implementation[4] for our experiments. Older supervised models for SP, that are not based on embeddings, have been previously evaluated on SP-10K (Zhang et al., 2019a), therefore we do not include those here.

3.3 Corpus

We use the British National Corpus (BNC-Consortium, 2007) as the source for word co-occurrences for the embeddings. Since BNC is sentence segmented, our co-occurrence counting never jumps across a sentence. The word casing is normalized to *small*, punctuations are removed, and the vocabulary is limited to words occurring at least 100 times in the corpus.

4 Experiments

In the following experiments, we compare our syntagmatic embeddings with its paradigmatic counterpart, identify its best parameters, distinguish weighted from unweighted factorisation, evaluate against baseline embeddings and test how our unsupervised SP learning method compares with a supervised model. The parameters involved in the factorisation of the word co-occurrence matrix are: 1) size of the co-occurrence window (**ws**), 2) term-weight or the co-occurrence weighting function (**tw**), 3) dimensionality of the embedding space or the number of principle components (**dim**), and 4) the exponential weight on singular values (**p**).

We experiment with the following parameter values: **ws**=[1, 2, 3, 4], **dim**=[20, 50, 100, 300], **p**=[-0.5, 0, 0.5, 1], **tw**=[*raw*, *log*, *pmi*, *ppmi*]. In term-weights *raw* denotes the co-occurrence frequency of the word as it is , *log* is the log_2 of the raw co-occurrence frequency, *pmi* is the point-wise mutual information given by equation 4 where subscript '*' stands for a summation across a particular axis, and *ppmi* is the positive-only variant of *pmi* as given by equation 5.

$$PMI_{i,j} = log \frac{M_{i,j}\, M_{*,*}}{M_{i,*}\, M_{*,j}} \quad (4)$$

$$PPMI_{i,j} = max(0, PMI_{i,j}) \quad (5)$$

4.1 Syntagmatic Vs Paradigmatic

In our first experiment we compare syntagmatic representation to paradigmatic representation. Here we consider only the unweighted factorisation model. The paradigmatic model is similar to the syntagmatic model described in section 2.2, but has a context window that is symmetric and non-directional. To get a more realistic picture of these methods, we compare a cohort of syntagmatic and paradigmatic models that have different parameter values. Each of the 4 parameters have 4 chosen parameter values. Since each parameter value combination gives us a different model, we get a total of 256 syntagmatic and 256 paradigmatic models.

[3] it runs only on a given prepackaged corpus, we found it difficult to replicate their packaging for our corpus

[4] https://github.com/HKUST-KnowComp/MWE

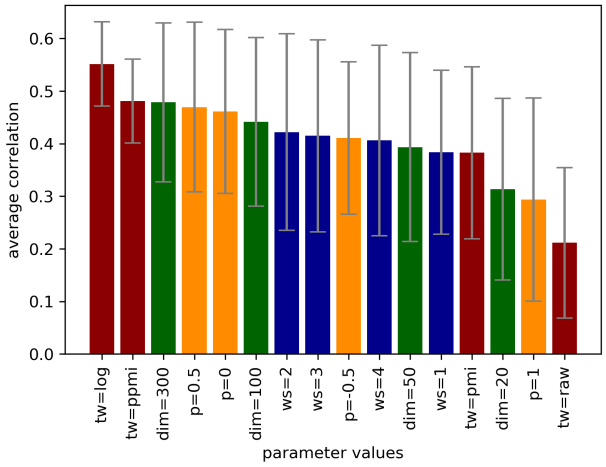

Figure 2: Average correlation (with standard deviation) in syntagmatic models that have the same parameter-value.

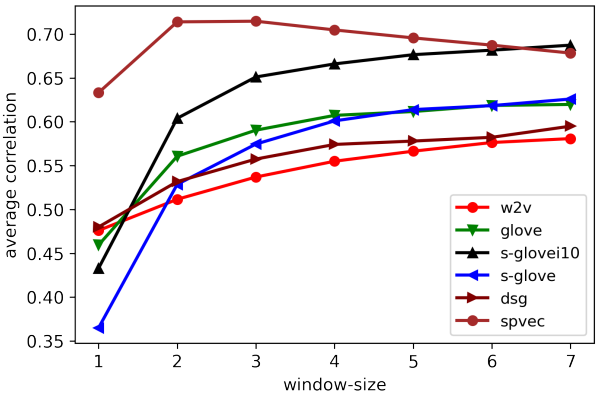

Figure 3: Average correlation of weighted and unweighted models with varying window sizes.

For each model (parameter-value combination) we compute the average correlation over the 3 SP relations. We see that in 69% of the total parameter instances the syntagmatic model is better than paradigmatic model. In those instances, on average the syntagmatic model improves the correlation by 0.14 points, which is an improvement of 54%. The maximum correlation obtained by a syntagmatic model is 0.71 and by the paradigmatic model is 0.58.

Figure 1 shows two line plots for the average correlation values of syntagmatic and paradigmatic embeddings. Each particular parameter-value combination is a value on the x-axis, for which the there are two correlation values on the y-axis; one of the syntagmatic model and the other of the paradigmatic model. Apart from showing that syntagmatic models are generally better than paradigmatic models, it shows that certain parameter combinations give syntagmatic models a much greater advantage. On the downside we see that for a good number of poorly performing paradigmatic models their, syntagmatic counterpart performed even worse. There are also certain pathological parameter combinations that substantially pull down syntagmatic representations compared to corresponding paradigmatic representation. But overall, this experiment shows that syntagmatic embeddings are substantially better at capturing SP.

4.2 Parameter Impact

In our second experiment we try to understand the relative importance of each parameter-value. For this we look at all 256 syntagmatic models and compute the mean and standard deviation of the correlation score among those models that have a particular parameter-value. For example we take the parameter-value **tw**=*log* and look at all syntagmatic models with that particular parameter-value, and compute the mean and standard deviation of their correlation score. We do the same with all 16 parameter-values.

Figure 2 gives the results of this experiment. We see that term-weight is the most important parameter, and **tw**=*log* the most significant parameter-value. No matter what the other parameters values are, using *log* as the term-weight gives on average a correlation score of 0.55 ± 0.07. Further, we see that the dimensionality of the embedding space is the next most significant parameter. Here we see that higher values are better, but this is only because we didn't consider even higher[5] values in this experiment (>300). It is well understood that there is an optimal dimension which is task and corpus dependent, below which a model does not have enough capacity, and above which the model tends to pick up noise (Yin and Shen, 2018). A more interesting aspect is the significance of the exponential weighting factor p. The SVD factorizes the co-occurrence matrix as $M = USV^\intercal$, which can be factored into left and right components as $M = [US^{\frac{1}{2}}][S^{\frac{1}{2}}V^\intercal]$. We see that **p**=*0.5* is indeed the right[5] value for the exponential weight factor.

4.3 Influence of Weighted Factorisation

To understand the influence of weighted factorisation on syntagmatic embeddings, we compare the syntagmatic GloVe (**s-glove**) model, introduced in section 2.3 to our SVD based unweighted fac-

[5]See figure 5 in appendix

torisation model. We choose our best performing SVD based syntagmatic model (**tw**=*log*, **dim**=*300*, **p**=*0.5*) naming it **spvec**. We also test the SkipGram Word2vec (**w2v**) and GloVe (**glove**), for providing a comparison with popular paradigmatic models, and DSG to compare against a model with directional information. Embedding sizes in all models are 300, and window-sizes 1 to 7 are evaluated. Other parameters of **dsg**, **s-glove**, **glove**, **w2v** are kept to the default values in their respective implementations.

Figure 3 shows the results of the experiment. We find that our SVD based unweighted syntagmatic model outperforms all other models, including the weighted syntagmatic model based on GloVe. The **s-glove** model performed slightly worse than the paradigmatic glove (**glove**) model under low window-sizes. We tried increasing the number of iterations in the training process, from the default 5 to 10. The resulting model (**s-glovei10**) performed much better than than paradigmatic GloVe model. It is interesting to note that all weighted models behave similarly to increasing window-sizes. They perform better as window-sizes increase. Whereas, our SVD based unweighted model (**spvec**) gives a better performance at window-size 2 and 3 and gradually decreases in performance as window-size is further increased. The directional variant of Word2vec (**dsg**) performs better than Word2vec, but performs poorly compared to **spvec**. Comparing **s-glovei10** and **spvec**, we see that even at much higher window-size of 15 (not shown in figure 3), **s-glovei10** barely reaches an average correlation of 0.69. **spvec** on the other hand gets an average correlation 0.71 at a much smaller window-sizes (2 and 3).

4.4 Comparison to Supervised Models

Our syntagmatic word embedding model aims to provide an effective method to approach selectional preference in the absence of a parsed corpus. In this experiment we assess how deficient our unsupervised model is when compared to supervised models. Since we were not able to use the available implementation of MWE, we simply compare our unsupervised syntagmatic model (**spvec**) with supervised versions of itself. The supervised version of syntagmatic embeddings is obtained by defining word co-occurrence as a pair of words related by a dependency relation. For this we parse our corpus (BNC) using the Stanford dependency parser (Qi et al., 2020). In order to remain compatible with a syntagmatic model, we maintain word ordering of the co-occurrences. For example, the sentence 'big cat ate rat' gives three co-occurrences where the head and the dependent are ordered as they are found in the sentence: 'big cat', 'cat ate' and 'ate rat'. We test two supervised models 1) **spvec-s**: which uses all dependency related word pairs 2) **spvec-sr**: which uses only related word pairs in a particular dependency relation. **spvec-sr** thus has 3 distinct embedding pairs (left/right) per word, an embedding pair for each of the tested dependency relation: *amod*, *nsubj*, *dobj*. For comparison we also show the results of unsupervised paradigmatic models.

model	amod	nsubj	dobj	AVG
w2v	0.582	0.489	0.539	0.536
glove	0.694	0.489	0.587	0.590
dsg	0.625	0.490	0.556	0.557
s-glovei10	0.738	0.565	0.649	0.650
spvec	**0.750**	**0.654**	**0.738**	**0.714**
spvec-s	0.761	0.637	0.740	0.712
spvec-sr	0.757	0.653	0.741	0.717

Table 3: Spearman's correlation for supervised and unsupervised models on the SP-10K dataset.

Table 3 gives the results of this experiment. Surprisingly we see that our unsupervised model (**spvec**) is as good as its supervised counterparts (**spvec-s** and **spvec-sr**). The model trained on all dependency related word pairs scores lower than the fully unsupervised model. The model with relation specific embeddings improves on the fully unsupervised model only by a meager 0.4%. We clearly see that unsupervised syntagmatic embeddings are not deficient but may be as good as supervised models.

5 Related Work

There have been previous studies that explored Syntagmatic representations. Rapp (2002); Sahlgren (2006) viewed syntagmatic representations as first-order word co-occurrence statistics, and paradigmatic representations as second-order statistics. First-order models represent words using text units in which they appear. Text units are generally documents or large regions of text, like paragraphs. Thus, first order statistics come from a word-document co-occurrence matrix, whereas paradig-

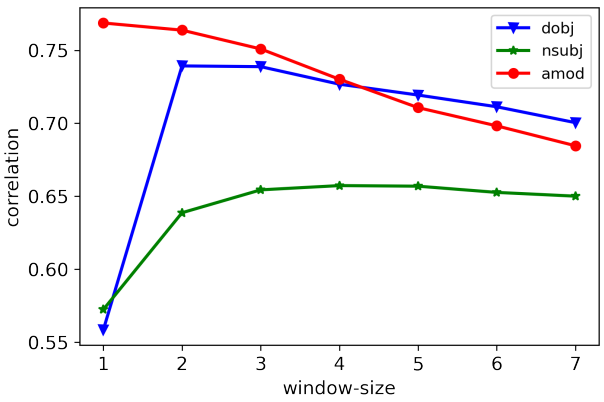

Figure 4: Window-size preferences of **spvec** for different relations.

matic representations come from word-word co-occurrence matrix and hence called second order. While their evaluation of paradigmatic representation as second-order statistics was appropriate, their claim of syntagmatic representation as first-order statistics is not well justified. This is because the evaluation datasets they used for first-order models were a mix of (mostly) paradigmatic and syntagmatic relations, and not purely syntagmatic. A large-scale study by Lapesa et al. (2014) showed that fine-tuned second-order statistics can capture both syntagmatic and paradigmatic relations. Different parametrisations, mainly window size and dimensionality reduction, were shown to adapt the second-order statistics to either relations accordingly.

The notion of syntagmatic representation explored in our work is adapted from Schütze and Pedersen (1993), in which the syntagmatic representation is introduced qualitatively without resorting to any quantitative studies. Our study on the other hand applies syntagmatic representation to the task of selectional preference, exploring various model parametrisations.

6 Discussion

Our experiments have shown that a weakly structured model can be as good as a strongly structured model. The **spvec** model, though unsupervised, incorporates a simple linguistically motivated bias/structure – directionality or word order. Such a weakly biased model, when coupled with low-rank embedding process, seems to pickup appropriate linguistic structure by effectively getting rid of noise. But why did the supervised models not have a bigger advantage when compared to the unsupervised model? We can hypothesize that words that are not directly related by a dependency relation but are in the vicinity of a target word make substantial contribution to the semantics of the word which may not be captured by a dependency-parsed model. It can also be because the low-rank embedding process is as good at removing noise as a dependency parse. A closer look at the results reveal that *amod* and *dobj* relations do benefit from supervision, although it is minor. The effect of window-size on each of the dependency relation, may help us to better understand this (figure 4). In the unsupervised model, *amod* relation is maximized with a window-size of 1, but the results reported in table 3 are of window-size 3. Certainly, the excess window-size will result in noise which may be mitigated by a dependency parse, as seen in the results of supervised models. Similarly, *dobj* relation which is maximized in the unsupervised model at window-size of 4 also benefits from the dependency parse. However, the case of *nsubj* relation does not fit this reasoning. *nsubj* is maximized in the unsupervised model at a window-size of 2, but even at window-size 3 it improves over the supervised model. Here we may have to consider the possibility that, words that are not directly related may contribute to the semantics, which is lost in a dependency-parsed model. We would also like to point out that parsing a large corpus can be resource intensive. Parsing the BNC consumed about 24 GPU[6] hours. However, our experiments show that the gains derived do not substantiate the compute incurred. The unsupervised **spvec** model performs the factorisation in less than 5 minutes on a 20-core CPU.

Weighted factorisation of word co-occurrences is generally found to produce high quality word embeddings. Previously such embeddings showed improvements in tasks such as word similarity and solving word analogies. But we have shown that, when it comes to selectional preference and syntagmatic embeddings, weighted factorisation may be detrimental.

We also observe that appropriate co-occurrence term-weights are crucial for the performance. *PPMI* has been shown to work well for tasks that test paradigmatic nature such as word similarity (Bullinaria and Levy, 2007). Pennington et al. (2014) remarked that *log* is better for solving word

[6]Nvidia RTX 2080 GPU

analogies than *PPMI*. Our experiments show that *log* is also valuable for learning selectional preference.

Here we have tested our syntagmatic embeddings only on English, but it should be directly applicable to other word-ordered languages also.

7 Conclusion

In this paper, we have introduced syntagmatic word embeddings, a simple and effective method, for learning selectional preference (SP). Our model is simple because it captures SP by direct factorisation of a word co-occurrence matrix. We have showed that by incorporating a weak linguistic bias of directionality as a proxy for syntactic relations, our model can be made as effective as a model with access to syntactic relations. This is important because SP has always been seen as a task that requires a dependency-parsed corpus, our work shows that it need not be the case.

We hope that syntagmatic embeddings will be a valuable source of selectional preference information for resource-poor as well as resource-rich languages. We also hope that the structural bias of directionality will be further explored in simple models for other NLP tasks, instead of relying on models that are complex and opaque to interpretation.

Acknowledgement

Renjith P Ravindran is funded by Department of Science and Technology (DST), Government of India, under the Inspire Fellowship Programme.

References

BNC-Consortium. 2007. The british national corpus, version 3 (bnc xml edition). Bodleian Libraries, University of Oxford. Http://www.natcorp.ox.ac.uk/.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, pages 510–526.

John Caron. 2001. *Experiments with LSA Scoring: Optimal Rank and Basis*, page 157–169. Society for Industrial and Applied Mathematics, USA.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26–35, Doha, Qatar. Association for Computational Linguistics.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic. Association for Computational Linguistics.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Dan Kalman. 1996. A singularly valuable decomposition: The svd of a matrix. *The College Mathematics Journal*, 27(1):2–23.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Comput. Linguist.*, 29(3):459–484.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 160–170, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA. MIT Press.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Ulrike Padó, Frank Keller, and Matthew W Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of the 28th CogSci*, pages 657–662.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Reinhard Rapp. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, page 104–111, USA. Association for Computational Linguistics.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Institutionen för lingvistik, Stockholm University.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.

Hinrich Schütze and Jan Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. In *Making Sense of Words - Ninth Annual Conference of the UW Centre for the New OED and Text Re- search*, pages 104–113.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.

Y. Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artif. Intell.*, 6:53–74.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 887–898. Curran Associates, Inc.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Hongming Zhang, Jiaxin Bai, Yan Song, Kun Xu, Changlong Yu, Yangqiu Song, Wilfred Ng, and Dong Yu. 2019a. Multiplex word embeddings for selectional preference acquisition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5247–5256, Hong Kong, China. Association for Computational Linguistics.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019b. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019c. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 872–881, Minneapolis, Minnesota. Association for Computational Linguistics.

A Detailed Parameter Study

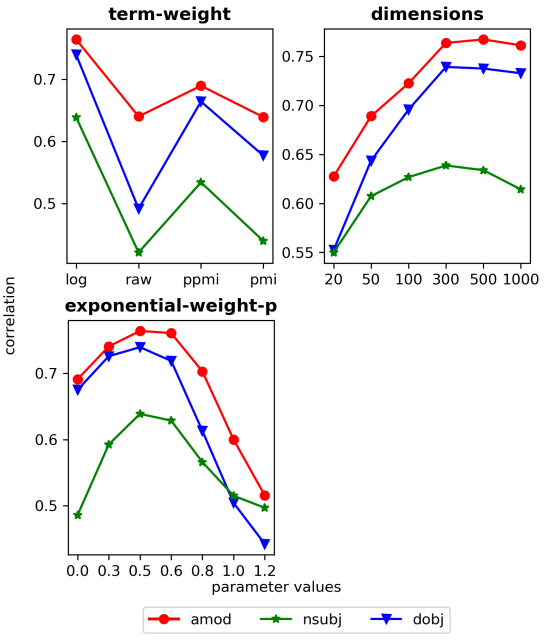

Figure 5: Variations in SP correlation of **spvec** on each relation with variations in parameter-values of term-weight, dimensions, and exponential-weight-p. Variations in window-size are shown in figure 4.