Vector space semantics with frequency-driven motifs

Shashank Srivastava
Carnegie Mellon University
Pittsburgh, PA 15217
ssrivastava@cmu.edu

Eduard Hovy
Carnegie Mellon University
Pittsburgh, PA 15217
hovy@cmu.edu

Abstract

Traditional models of distributional semantics suffer from computational issues such as data sparsity for individual lexemes and complexities of modeling semantic composition when dealing with structures larger than single lexical items. In this work, we present a frequency-driven paradigm for robust distributional semantics in terms of semantically cohesive lineal constituents, or *motifs*. The framework subsumes issues such as differential compositional as well as non-compositional behavior of phrasal consituents, and circumvents some problems of data sparsity by design. We design a segmentation model to optimally partition a sentence into lineal constituents, which can be used to define distributional contexts that are less noisy, semantically more interpretable, and linguistically disambiguated. Hellinger PCA embeddings learnt using the framework show competitive results on empirical tasks.

1 Introduction

Meaning in language is a confluence of experientially acquired semantics of words or multi-word phrases, and their semantic composition to create new meanings. For instance, successfully interpreting a sentence such as

The old senator kicked the bucket.

requires the knowledge that the semantic connotations of 'kicking the bucket' as a unit are the same as those for 'dying'. Short of explicit supervision, such semantic mappings must be inferred by a new language speaker through inductive mechanisms operating on observed linguistic usage. This perspective of acquired meaning aligns with the 'meaning is usage' adage, consonant with Wittgenstein's view of semantics. At the same time, the ability to adaptively communicate elaborate meanings can only be conciled through Frege's principle of compositionality, i.e., meanings of larger linguistic constructs can be derived from the meanings of individual components, modulated by their syntactic interrelations. Indeed, most linguistic usage appears compositional. This is supported by the fact even with very limited vocabulary, children and non-native speakers can often communicate surprisingly effectively.

It can be argued that to be sustainable, inductive aspects of meaning must be recurrent enough to be learnable by new users. That is, a non-compositional phrase such as 'kick the bucket' is likely to persist in common parlance only if it is frequently used with its associated semantic mapping. If a usage-driven meaning of a motif is not recurrent enough, learning this mapping is inefficient in two ways. First, the sparseness of observations would severely limit accurate inductive acquisition by new observers. Second, the value of learning a very infrequent semantic mapping is likely marginal. This motivates the need for a frequency-driven view of lexical semantics. In particular, such a perspective can be especially advantageous for distributional semantics for reasons we outline below.

Distributional semantic models (DSMs) that represent words as distributions over neighbouring contexts have been particularly effective in capturing fine-grained lexical semantics (Turney et al., 2010). Such models have engendered improvements in diverse applications such as selectional preference modeling (Erk, 2007), word-sense discrimination (McCarthy and Carroll, 2003), automatic dictionary building (Curran, 2003), and information retrieval (Manning et al., 2008). However, while conventional DSMs consider colloca-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 634–643,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

	With the bad press in wake of the financial crisis, businesses are leaving our shores .
crisis:	<bad, businesses, financial, leaving, press, shores, wake>
financial_crisis:	<bad press, businesses, in wake of, leaving our shores>

Table 1: Meaning representation by conventional DSMs vs notional ideal

tion strengths (through counts and PMI scores) of word neighbourhoods, they disregard much of the regularity in human language. Most significantly, word tokens that act as latent dimensions are often derived from arbitrary tokenization. The example given in Table 1 succinctly describes this. The first row in the table shows a representation of the meaning of the token ‘crisis’ that a conventional DSM might extract from the given sentence after stopword removal. While helpful, the representation seems unsatisfying since words such as ‘press’, ‘wake’ and ‘shores’ seem to have little to do with a crisis. From a semantic perspective, a representation similar to the second is more valuable: not only does it represent a semantic mapping for a more specific meaning, but the latent dimensions of the representation have are less noisy (e.g., while ‘wake’ is semantically ambiguous, its surrounding context in ‘in wake of’ disambiguates it) and more intuitive in regards of semantic interepretability. This is the overarching theme of this work: we present a frequency driven paradigm for extending distributional semantics to phrasal and sentential levels in terms of such semantically cohesive, recurrent lexical units or *motifs*.

We propose to identify such semantically cohesive motifs in terms of features inspired from frequency-characteristics, linguistic idiosyncrasies, and shallow syntactic analysis; and explore both supervised and semi-supervised models to optimally segment a sentence into such motifs. Through exploiting regularities in language usage, the framework can efficiently account for both compositional and non-compositional word usage, while avoiding the issue of data-sparsity by design. Our principal contributions in this paper are:

- We present a framework for extending distributional semantics to learn semantic representations of both words and phrases in terms of recurrent motifs, rather than arbitrary word tokens
- We present a simple model to segment a sentence into such motifs using a feature-set drawing from frequency statistics, information theory, linguistic theories and shallow syntactic analysis
- Word and phrasal representations learnt through the approach outperform conventional DSM representations on empirical tasks

This paper is organized as follows: In Section 2, we briefly review related work in the domain of compositional distributional semantics, and motivate our formulation. Section 3 describes our methodology, which consists of a frequency-driven segmentation model to partition text into semantically meaningful recurring lineal-subunits, a representation learning framework for learning new semantic embeddings based on this segmentation, and an approach to use such embeddings in downstream applications. We present experiments and empirical evaluations for our method in Section 4. Finally, we conclude in Section 5 with a summary of our principal findings, and a discussion of possible directions for future work.

2 Related Work

While DSMs have been valuable in representing semantics of single words, approaches to extend them to represent the semantics of phrases and sentences has met with only marginal success. While there is considerable variety in approaches and formulations, existing approaches for phrasal level and sentential semantics can broadly be partitioned into two categories.

2.1 Compositional approaches

These have aimed at using semantic representations for individual words to learn semantic representations for larger linguistic structures. These methods implicitly make an assumption of compositionality, and often include explicit computational models of compositionality. Notable among such models are the additive and multiplicative models of composition by Mitchell and Lapata (2008), Grefenstette et al. (2010), Baroni and

Zamparelli's (2010) model that differentially models content and function words for semantic composition, and Goyal et al.'s SDSM model (2013) that incorporates syntactic roles to model semantic composition. Notable among the most effective distributional representations are the recent deep-learning approaches by Socher et al. (2012), that model vector composition through non-linear transformations. While word embeddings and language models from such methods have been useful for tasks such as relation classification, polarity detection, event coreference and parsing; much of existing literature on composition is based on abstract linguistic theory and conjecture, and there is little evidence to support that learnt representations for larger linguistic units correspond to their semantic meanings. While works such as the SDSM model suffer from the *problem of sparsity* in composing structures beyond bigrams and trigrams, methods such as Mitchell and Lapata (2008)and (Socher et al., 2012) and Grefenstette and Sadrzadeh (2011) are restricted by significant *model biases* in representing semantic composition by generic algebraic operations. Finally, the assumption that semantic meanings for sentences could have representations similar to those for smaller individual tokens is in some sense unintuitive, and not supported by linguistic or semantic theories.

2.2 Tree kernels

Tree Kernel methods have gained popularity in the last decade for capturing syntactic information in the structure of parse trees (Collins and Duffy, 2002; Moschitti, 2006). Instead of procuring explicit representations, the kernel paradigm directly focuses on the larger goal of quantifying semantic similarity of larger linguistic units. Structural kernels for NLP are based on matching substructures within two parse trees , consisting of word-nodes with similar labels. These methods have been useful for eclectic tasks such as parsing, NER, semantic role labeling, and sentiment analysis. Recent approaches such as by Croce et al. (2011) and Srivastava et al. (2013) have attempted to provide formulations to incorporate semantics into tree kernels through the use of distributional word vectors at the individual word-nodes. While this framework is attractive in the lack of assumptions on representation that it makes, the use of distributional embeddings for individual tokens means that it suffers from the same shortcomings as described for the example in Table 1, and hence these methods model semantic relations between word-nodes very weakly. Figure 1 shows an example of the shortcomings of this general approach.

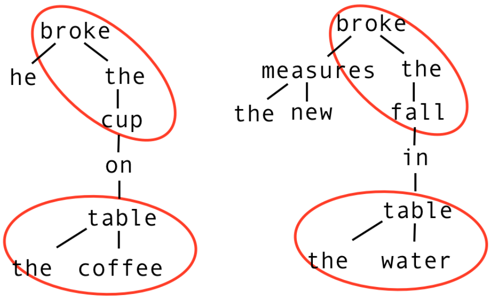

Figure 1: Tokenwise syntactic and semantic similarities don't imply sentential semantic similarity

While the two sentences in consideration have near-identical syntax and could be argued to have semantically aligned words in similar positions, the semantics of the complete sentences are widely divergent. Specifically, the 'bag of words' assumption in tree kernels doesn't suffice for these lexemes, and a stronger semantic model is needed to capture phrasal semantics as well as diverging inter-word relations such as in 'coffee table' and 'water table'. Our hypothesis is that a model that can even weakly identify recurrent motifs such as 'water table' or 'breaking a fall' would be helpful in building more effective semantic representations. A significant advantage of a frequency driven view is that it makes the concern of compositionality of recurrent phrases immaterial. If a motif occurs frequently enough in common parlance, its semantics could be captured with distributional models irrespective of whether its associated semantics are compositional or acquired.

2.3 Identifying multi-word expressions

Several approaches have focused on supervised identification of multi-word expressions (MWEs) through statistical (Pecina, 2008; Villavicencio et al., 2007) and linguistically motivated (Piao et al., 2005) techniques. More recently, hybrid methods based on both statistical as well as linguistic features have been popular (Tsvetkov and Wintner, 2011). Ramisch et al. (2008) demonstrate that adding part-of-speech tags to frequency counts substantially improves performance. Other methods have attempted to exploit morphological, syntactic and semantic characteristics of MWEs. In particular, approaches such as Bannard (2007) use syntactic rigidity to characterize MWEs. While existing work has focused on the classification task of categorizing a phrasal constituent as a MWE or a non-MWE, the general ideas of most of these works are in line with our current framework, and the feature-set for our motif segmentation model is designed to subsume most of these ideas. It is worthwhile to point out that the task of motif segmentation is slightly different from MWE identification. Specifically, the onus on recurrent occurrences means that non-decomposibility is not an essential consideration for a word to be considered a motif. In line with the proposed paradigm, typical MWEs such as 'shoot the breeze', 'sour note' and 'hot dog' would be considered valid lineal motifs. [1] In addition, even decomposable recurrent lineal phrases such as 'love story', 'federal government', and 'millions of people' are marked as meaningful recurrent motifs. Finally, and least interestingly, we include common named entities such as 'United States' and 'Java Virtual Machine' within the ambit of motifs.

3 Method

In this section, we define our frequency-driven framework for distributional semantics in detail. As just described above, our definition for motifs is less specific than MWEs. With such a working definition, contiguous motifs are likely to make distributional representations less noisy and also assist in disambiguating context. Also, the lack of specificity ensures that such motifs are common enough to meaningfully influence distributional representation beyond single tokens. A method towards frequency-driven distributional semantics could involve the following principal components:

3.1 Linear segmentation model

The segmentation model forms the core of the framework. Ideally, it fragments a given sentence into non-overlapping, semantically meaningful, empirically frequent contiguous sub-units or motifs. The model accounts for possible segmentations of a sentence into potential motifs, and prefers recurrent and cohesive motifs through features that capture frequency-based and statistical features, as well as linguistic idiosyncracies. This is accomplished using a very simple linear chain model and a rich feature set consisting of a combination of frequency-driven, information theoretic and linguistically motivated features.

Let an observed sentence be denoted by $\mathbf{x}$, with the individual tokens x_i denoting the i'th token in the sentence. The segmentation model is a chain LVM (latent variable model) that aims to maximize a linear objective defined by:

$$J = \sum_i w_i f_i(y_k, y_{k-1}, \mathbf{x})$$

where f_i are arbitrary Markov features that can depend on segments (potential motifs) of the observed sentence $\mathbf{x}$, and contiguous latent states. The features are chosen so as to best represent frequency-based, statistical as well as linguistic considerations for treating a segment as an agglutinative unit, or a motif. In specific, these features could encode characteristics such as frequency statistics, collocation strengths and syntactic distinctness, or inflectional rigidity of the considered segments; described in detail in Section 3.2. The model is an instantiation of a simple featurized HMM, and the weighted sum of features corresponding to a segment is cognate with an affinity score for the 'stickiness' of the segment, i.e., the affinity for the segment to be treated as holistic unit or a single motif.

We also associate a penalizing cost for each non unary-motif to avoid aggressive agglutination of tokens. In particular, for an ngram occurrence to be considered a motif, the marginal contribution due to the affinity of the prospective motif should at minimum exceed this penalty. The weights for the affinity functions as well as these penalties are learnt from data using full as well as partial annotations. The latent state-variables y_k denotes the membership of the token $\mathbf{x_k}$ to a unary or a larger motif; and the state-sequence collectively gives the segmentation of the sentence. An individual state-variable y_k encodes a pairing of the size of the encompassing ngram motif, and the position of the word x_k within it. For instance, $y_k = T_3$ denotes that the token $\mathbf{x_k}$ is the final position in a trigram motif.

3.1.1 Inference of optimal segmentation

If the optimal weights w_i are known, inference for the best motif segmentation can be performed

[1] We note that since we take motifs as lineal units, the current method doesn't subsume several common non-contiguous MWEs such as 'let off' in 'let him off'.

in linear time (in the number of tokens) following the generalized Viterbi algorithm. A slightly modified version of Viterbi could also be used to find segmentations that are constrained to agree with some given motif boundaries, but can segment other parts of the sentence optimally under these constraints. This is necessary for the scenario of semi-supervised learning of weights with partially annotated sentences, as described later.

3.2 Learning motif affinities and penalties

We briefly discuss data-driven learning of weights for features that define the motif affinity scores and penalties. We describe learning of the model parameters with fully annotated training data, as well as an approach for learning motif segmentation that requires only partial supervision.

Supervised learning: In the supervised case, optimal state sequences $\mathbf{y}^{(\mathbf{k})}$ are fully observed for the training set. For this purpose, we created a dataset of 1000 sentences from the Simple English Wikipedia and the Gigaword Corpus, and manually annotated it with motif boundaries using BRAT (Stenetorp et al., 2012). In this case, learning can follow the online structured perceptron learning procedure by Collins (2002), where weights updates for the k'th training example $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ are given as:

$$w_i \leftarrow w_i + \alpha(f_i(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - f_i(\mathbf{x}^{(k)}, \mathbf{y}'))$$

Here $\mathbf{y}' = Decode(\mathbf{x}^{(k)}, \mathbf{w})$ is the optimal Viterbi decoding using the current estimates of the weights. Updates are run for a large number of iterations until the change in objective drops below a threshold, and the learning rate α is adaptively modified as described in Collins et al. Implicitly, the weight learning algorithm can be seen as a gradient descent procedure minimizing the difference between the scores of highest scoring (Viterbi) state sequences, and the label state sequences.

Semi-supervised learning: In the semi-supervised case, the labels $y_i^{(k)}$ are known only for some of the tokens in $\mathbf{x}^{(\mathbf{k})}$. This is a commonplace scenario, where a part of a sentence has clear motif-boundaries, whereas the rest of the sentence is not annotated. For accumulating such data, we looked for occurrences of 2500 expressions from the WikiMWE dataset in sentences from the combined Simple English Wikipedia and Gigaword corpora. The query expressions in the retrieved sentences were marked with motif boundaries, while the remaining tokens in the sentences were left unannotated.

While the Viterbi algorithm can be used for tagging optimal state-sequences given the weights, the structured perceptron can learn optimal model weights given gold-standard sequence labels. Hence, in this case, we use a variation of the hard EM algorithm for learning. The algorithm proceeds as follows: in the E-step, we use the current values of weights to compute hard-expectations, i.e., the best scoring Viterbi sequences among those consistent with the observed state labels. In the M-step, we take the decoded state-sequences in the E-step as observed, and run perceptron learning to update feature weights w_i. Pseudocode of the learning algorithm for the partially labeled case is given in Algorithm 1.

Algorithm 1

```
Input: Partially labeled data $D = \{(x, y)_i\}$
Output: Weights $w$
Initialization: Set $w_i$ randomly, $\forall i$
for $i$ : 1 to $maxIter$ do
    Decode $D$ with current $w$ to find optimal Viterbi paths that agree with (partial) ground truths.
    Run Structured Perceptron algorithm with decoded tag-sequences to update weights $w$
end for
return $w$
```

The semi-supervised approach enables incorporation of significantly more training data. In particular, this method could be used in conjunction with a supervised approach. This would involve initializing the weights prior to the semi-supervised procedure with the weights from the supervised learning model, so as to seed the semi-supervised approach with reasonable model, and use the partially annotated data to fine-tune the supervised model. The sequential approach, akin to annealing weights, can efficiently utilize both full and partial annotations.

3.2.1 Feature engineering

In this section, we describe the principal features used in the segmentation model
Transitional features and penalties:

- Transitional features $f_{trans}(y_{i-1}, y_i) =$

I_{y_{i-1},y_i} [2] describing the transitional affinities of state pairs. Since our state definitions preclude certain transitions (such as from state T_2 to T_1), these weights are initialized to $-\infty$ to expedite training.

- N-gram penalties: f_{ngram} We define a penalty for tagging each non-unary motif as described before. For a motif to be tagged, the improvement in objective score should at least exceed the corresponding penalty. e.g., $f_{qgram}(y_i) = I_{y_i=Q_4}$ denotes the penalty for tagging a tetragram. [3]

Frequency-based, information theoretic, and POS features:

- Absolute and log-normalized motif frequencies $f_{ngram}(x_{i-n+1}, ...x_{i-1}, x_i, y_i)$. This feature is associated with a particular token-sequence and ngram-tag, and takes the value of the motif-frequency if the motif token-sequence matches the feature token-sequence, and is marked as with a matching tag. e.g., $f_{bgram}(x_{i-1} = love, x_i = \text{story}, y_i = B_2)$.

- Absolute and log-normalized motif frequencies for a particular POS-sequence. This feature is associated with a particular POS-tag sequence and ngram-tag, and takes the value of the motif-frequency if the motif token-sequence gets a matching tag, and is marked as with a matching ngram tag. e.g., $f_{bgram}(p_{i-1} = VB, p_i = NN, y_i = B_2)$.

- Medians and maxima of pairwise collocation statistics for tokens for a particular size of ngram motifs: we use the following statistics: pointwise mutual information, Chi-square statistic, and conditional probability. We also used POS sensitive versions of these, which performed much better than plain versions in our evaluations.

- Histogram counts of inflectional forms of token sequence for the corresponding ngram motif and POS sequence: this features takes the value of the count of inflectional forms of an ngram that account for 90% of occurrences of all inflectional forms.

- Entropies of histogram distributions of inflectional variants (described above).

- Features encoding syntactic rigidity: ratios and log-ratios of frequencies of an ngram motif and variations by replacing a token using near synonyms from its synset.

Additionally, a few feature for the segmentations model contained minor orthographic features based on word shape (length and capitalization patterns). Also, all numbers, URLs, and currency symbols were normalized to the special NUMERIC, URL, and CURRENCY tokens respectively. Finally, a gazetteer feature checked for occurrences of motifs in a gazetteer of named entities.

3.3 Representation learning

With the segmentation model described in the previous section, we process text from the English Gigaword corpus and the Simple English Wikipedia to partition sentences into motifs. Since the segmentation model accounts for the contexts of the entire sentence in determining motifs, different instances of the same token could evoke different meaning representations. Consider the following sentences tagged by the segmentation model, that would correspond to different representations of the token 'remains': once as a standalone motif, and once as part of an encompassing bigram motif ('remains classified').

Hog prices have declined sharply , while the cost of corn remains relatively high.

Even with the release of such documents, questions are not answered, since only the agency knows what remains classified

Given constituent motifs of each sentence in the data, we can now define neighbourhood distributions for unary or phrasal motifs in terms of other motifs (as envisioned in Table 1). In our experiments, we use a window-length of 5 adjoining motifs on either side to define the neighbourhood of a constituent. Naturally, in the presence of multi-word motifs, the neighbourhood boundary could be more extended than in a conventional DSM.

With such neighbourhood contexts, the distributional paradigm posits that semantic similarity between a pair of motifs can be given by a sense of 'distance' between the two distributions. Most popularly, traditional measures of vector distance

[2] Here, I denotes the indicator function

[3] It is straightforward to preclude partial n-gram annotations near sentence boundaries with prohibitive penalties.

such as the cosine similarity, Euclidean distance and City-block distance have been used in several distributional approaches. Additionally, several distance measures between discrete distributions exist in statistical literature, most famously the Kullback Leibler divergence, Bhattacharyya distance and the Hellinger distance. Recent work (Lebret and Lebret, 2013) has shown that the Hellinger distance is an especially effective measure in learning distributional embeddings, with Hellinger PCA being much more computationally inexpensive than neural language modeling approaches, while performing much better than standard PCA, and competitive with the state-of-the-art in downstream evaluations. Hence, we use the Hellinger measure between neighbourhood motif distributions in learning representations.

The Hellinger distance between two categorical distributions $P = (p_1 ... p_k)$ and $Q = (q_1 ... q_k)$ is defined as:

$$
\begin{aligned}
H(P,Q) &= \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \\
&= \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2
\end{aligned}
$$

The Hellinger measure has intuitively desirable properties: specifically, it can be seen as the Euclidean distance between the square-roots transformed distributions, where both vectors $\sqrt{P}$ and $\sqrt{Q}$ are length-normalized under the same(Euclidean) norm. Finally, we perform SVD on the motif similarity matrix (with size of the order of the total vocabulary in the corpus), and retain the first k principal eigenvectors to obtain low-dimensional vector representations that are more convenient to work with. In our preliminary experiments, we found that $k = 300$ gave quantitatively good results, with marginal change with added dimensionality. We use this setting for all our experiments.

4 Experiments

In this section, we describe some experimental evaluations and findings for our approach. We first quantitatively and qualitatively analyze the performance of the segmentation model, and then evaluate the distributional motif representations learnt by the model through two downstream applications.

4.1 Motif segmentation

In an evaluation of the motif segmentations model within the perspective of our framework, we believe that exact correspondence to human judgment is unrealistic, since guiding principles for defining motifs, such as semantic cohesion, are hard to define and only serve as working principles. However, for purposes of relative comparison, we quantitatively evaluate the performance of the motif segmentation models on the fully annotated dataset. For this experiment, the gold-annotated corpus was split into a training and test sets in a 9:1 proportion. A small fraction of the training split was set apart for development and validation. For this evaluation, we considered a motif boundary as correct only for an exact match, i.e., when both its boundaries (left and right) were correctly predicted. Also, since a majority of motifs are unary tokens, including them into consideration artificially boosts the accuracy, whereas we are more interested in the prediction of larger n-gram tokens. Hence we report results on the performance on only non-unary motifs.

	P	**R**	**F**
Rule-based baseline	0.85	0.10	0.18
Supervised	0.62	0.28	0.39
Semi-supervised	0.30	0.17	0.22
Supervised + annealing	0.69	0.38	0.49

Table 2: Results for motif segmentations

Table 2 shows the performance of the segmentation model with the three proposed learning approaches described earlier. For a baseline, we consider a rule-based model that simply learns all ngram segmentations seen in the training data, and marks any occurrence of a matching token sequence as a motif; without taking neighbouring context into account. We observe that this model has a very high precision (since many token sequences marked as motifs would recur in similar contexts, and would thus have the same motif boundaries). However, the rule-based method has a very row recall due to lack of generalization capabilities. We see that while all three learning algorithms perform better than the baseline, the performance of the purely unsupervised system is inferior to supervised approaches. This is not unexpected: the supervision provided to the model is very weak due to a lack of negative examples (which leads to spurious motif taggings,

While men often (openly or privately) sympathized with Prince Charles when the princess went public about her rotten marriage , women cheered her on.
The healthcare initiative has become a White elephant for the federal government.
Chirac and Juppe have made a bad situation worse by seeking to meet Maastricht criteria not by cutting spending, but by raising taxes still further.
Now , say Vatican observers , Pope John Paul II wants to show the world that many church members did resist the Third Reich and paid the price.

Table 3: Examples of output from sentence segmentation model

leading to a low precision), as well as no examples of transitions between adjacent motifs (to learn transitional weights and penalties). The supervised model expectedly outperforms both the rule-based and the semi-supervised systems. However, the supervised learning model with subsequent annealing outperforms the supervised model in terms of both precision and recall; showing the utility of the semi-supervised method when seeded with a good initial model, and the additive value of partially labeled data.

Qualitative analysis of motif-segmented sentences shows that our designed feature-set is effective and helpful in identifying semantically cohesive ngrams. Table 3 provides four examples. The first example correctly identifies ‘went public’, while missing out on the potential motif ‘cheered her on’. In general, these examples illustrate that the model can identify idiomatic and idiosyncratic themes as well as commonly recurrent ngrams (in the second example, the model picks out ‘has become’ which is highly recurrent, but doesn’t have the semantic cohesiveness of some of the other motifs). In particular, consider the second example, where the model picks ‘white elephant’ as a motif. In such cases, the disambiguating influence of context incorporated by the motif is apparent.

Elephant	White elephant
tusks	expensive
trunk	spend
african	biggest
white	the project
indian	very high
baby	multibillion dollar

The above table shows some of the top results for the unary token ‘elephant’ by frequency, and frequent unary and non-unary motifs for the motif ‘white elephant’ retrieved by the segmentation model.

4.2 Distributional representations

For evaluating distributional representations for motifs (in terms of other motifs) learnt by the framework, we test these representations in two downstream tasks: sentence polarity classification and metaphor detection. For sentence polarity, we consider the Cornell Sentence Polarity corpus by Pang and Lee (2005), where the task is to classify the polarity of a sentence as positive or negative. The data consists of 10662 sentences from movie reviews that have been annotated as either positive or negative. For composing the motifs representations to get judgments on semantic similarity of sentences, we use our recent Vector Tree Kernel approach The VTK approach defines a convolutional kernel over graphs defined by the dependency parses of sentences, using a vector representation at each graph node that representing a single lexical token. For our purposes, we modify the approach to merge the nodes of all tokens that constitute a motif occurrence, and use the motif representation as the vector associated with the node. Table 4 shows results for the sentence polarity task.

	P	**R**	**F1**
DSM	0.56	0.50	0.53
AVM	0.55	0.53	0.54
MVM	0.55	0.49	0.52
VTK	0.65	0.58	0.62
VTK + MotifDSM	**0.66**	**0.60**	**0.63**

Table 4: Results for Sentence Polarity detection

For this task, the motif based distributional embeddings vastly outperform a conventional distributional model (DSM) based on token distributions, as well as additive (AVM) and multiplicative (MVM) models of vector compositionality, as proposed by Lapata et al. The model is competitive with the state-of-the-art VTK (Srivastava et al., 2013) that uses the SENNA neural embeddings by Collobert et al. (2011).

	P	**R**	**F1**
CRF	**0.74**	0.50	0.59
SVM+DSM	0.63	0.80	0.71
VTK+ SENNA	0.67	**0.87**	0.76
VTK+ MotifDSM	0.70	**0.87**	**0.78**

Table 5: Results for Metaphor identification

On the metaphor detection task, we use the Metaphor dataset (Hovy et al., 2013). The data consists of sentences with defined phrases, and the task consists of identifying the linguistic use in these phrases as metaphorical or literal. For this task, the motif based model is expected to perform well as common metaphorical usage is generally through idiosyncratic MWEs, which the motif based models is specially geared to capture through the features of the segmentation model. For this task, we again use the VTK formalism for combining vector representations of the individual motifs. Table 5 shows that the motif-based DSM does better than discriminative models such as CRFs and SVMs, and also slightly improves on the VTK kernel with distributional embeddings.

5 Conclusion

We have presented a new frequency-driven framework for distributional semantics of not only lexical items but also longer cohesive motifs. The theme of this work is a general paradigm of seeking motifs that are recurrent in common parlance, are semantically coherent, and are possibly non-compositional. Such a framework for distributional models avoids the issue of data sparsity in learning of representations for larger linguistic structures. The approach depends on drawing features from frequency statistics, statistical correlations, and linguistic theories; and this work provides a computational framework to jointly model recurrence and semantic cohesiveness of motifs through compositional penalties and affinity scores in a data driven way.

While being deliberately vague in our working definition of motifs, we have presented simple efficient formulations to extract such motifs that uses both annotated as well as partially unannotated data. The qualitative and quantitative analyis of results from our preliminary motif segmentation model indicate that such motifs can help to disambiguate contexts of single tokens, and provide cleaner, more interpretable representations. Finally, we obtain motif representations in form of low-dimensional vector-space embeddings, and our experimental findings indicate value of the learnt representations in downstream applications. We believe that the approach has considerable theoretical as well as practical merits, and provides a simple and clean formulation for modeling phrasal and sentential semantics.

In particular, we believe that ours is the first method that can invoke different meaning representations for a token depending on textual context of the sentence. The flexibility of having separate representations to model different semantic senses has considerable valuable, as compared with extant approaches that assign a single representation to each token, and are hence constrained to conflate several semantic senses into a common representation. The approach also elegantly deals with the problematic issue of differential compositional and non-compositional usage of words. Future work can focus on a more thorough quantitative evaluation of the paradigm, as well as extension to model non-contiguous motifs.

References

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 263–270. Association for Computational Linguistics.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

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

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046. Association for Computational Linguistics.

James Richard Curran. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *ACL*.

Kartik Goyal, Sujay Kumar Jauhar, Huiying Li, Mrinmaya Sachan, Shashank Srivastava, and Eduard Hovy. 2013. A structured distributional semantic model: Integrating structure with semantics. *ACL 2013*, page 20.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2010. Concrete sentence spaces for compositional distributional models of meaning. *arXiv preprint arXiv:1101.0309*.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. *Meta4NLP 2013*, page 52.

Rémi Lebret and Ronan Lebret. 2013. Word emdeddings through hellinger pca. *arXiv preprint arXiv:1312.5542*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning: ECML 2006*, pages 318–329. Springer.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC MWE 2008 Workshop*, pages 54–57. Citeseer.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech & Language*, 19(4):378–397.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53.

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*, pages 1201–1211. Association for Computational Linguistics.

Shashank Srivastava, Dirk Hovy, and Eduard H. Hovy. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In *EMNLP*, pages 1411–1416.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 836–845. Association for Computational Linguistics.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *EMNLP-CoNLL*, pages 1034–1043.