Paraphrase-Supervised Models of Compositionality

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Compositional vector space models of meaning promise new solutions to stubborn language understanding problems. This paper makes two contributions toward this end: (i) it uses automatically-extracted paraphrase examples as a source of supervision for training compositional models (replacing previous work which relied on human-annotated examples) and (ii) develops a context-aware model for scoring phrasal compositionality. Experimental results indicate that these multiple sources of information can be used to learn partial semantic supervision that matches previous techniques in intrinsic evaluation tasks. Our approaches are also evaluated for their impact on a machine translation system where we show improvements in translation quality, demonstrating that compositionality in interpretation correlates with compositionality in translation.

1 Introduction

Numerous lexical semantic properties are captured by representations encoding distributional properties of words, as has been demonstrated in a variety of tasks (Turian et al., 2010; Turney and Pantel, 2010; Mikolov et al., 2013). However, this distributional account of meaning does not scale to larger units like phrases and sentences (Sahlgren, 2006; Collobert et al., 2011, *inter alia*),[1] motivating research into compositional models that *combine* word representations to produce representations of the semantics of longer units (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Socher et al., 2013). Previous work has learned these models using autoencoder formulations (Socher et al., 2011) or limited human supervision (Mitchell and Lapata, 2010). In contrast, we hypothesize that the equivalent knowledge can be obtained through monolingual paraphrases that have been extracted using word alignments and an intermediate language (Ganitkevitch et al., 2013). Confirming this hypothesis would allow the rapid development of compositional models in a large number of languages.

As their name suggests, these models also impose the assumption that longer units like phrases are **compositional**, i.e., a phrase's meaning can be understood from the literal meaning of its parts. However, countless examples that run contrary to the assumption exist, and handling these **non-compositional** phrases has been problematic and of long-standing interest in the community (Lin, 1999; Sag et al., 2002). (Non-) Compositionality detection as a problem is particularly relevant for downstream tasks like machine translation (MT) or information retrieval, where a literal interpretation of a non-compositional phrase could have comical (or disastrous) consequences. Scoring this phenomenon can provide vital information to other language processing systems on whether a multiword unit should be treated semantically as a single entity or not, and we posit that this behavior is strongly dependent on the context in which the expression occurs.

The goal of this work is to learn relatively accurate context-sensitive compositional models that are also directly applicable in real-world, noisy-data scenarios. This objective necessitates certain design decisions, and we propose a robust, scalable framework that learns compositional functions and scores relative phrasal compositionality. We make three contributions: first, a novel way to learn compositional functions for part-of-speech pairs that uses supervision from an

[1] There are simply far more distinct phrasal units whose representations have to be learned from the same amount of data.

automatically-extracted list of paraphrases (§3.1). Second, a context-dependent scoring model that scores the relative compositionality of a phrase (McCarthy et al., 2003) by computing the likelihood of its context given its paraphrase-learned representation (§4.2). And third, an evaluation of the impact of compositionality knowledge in an end-to-end MT setup. Our experiments (§5) reveal that using supervision from automatically extracted paraphrases produces compositional functions with equivalent performance to previous approaches that have relied on hand-annotated training data. Furthermore, compositionality features consistently improve the translations produced by an English–Spanish translation system.

2 Parametric Composition Functions

Our goal is to learn a function $\mathbf{f}(\mathbf{x}, \mathbf{y})$ that maps N-dimensional vector representations of phrase constituents $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N \times 1}$ to an N-dimensional vector representation of the phrase[2], i.e., the *composed* representation. A phrase is defined as any contiguous sequence of words of length 2 or greater, and does not have to adhere to constituents in a phrase structure grammar. This definition is in line with our MT application and ignores "gappy" phrases like many verb-object relations, as they are not treated as phrasal units in standard phrase-based MT (Koehn et al., 2003). We assume the existence of word-level vector representations for every word in our vocabulary of size V. Compositionality is modeled as a bilinear map, and two classes of linear models with different levels of parametrization are proposed. Unlike previous work (Baroni and Zamparelli, 2010; Socher et al., 2013; Grefenstette et al., 2013, *inter alia*) where the functions are word-specific, our compositional functions operate on part-of-speech (POS) tag pairs, which facilitates learning by drastically reducing the number of parameters, and only requires a shallow syntactic parse of the input.

2.1 Concatenation Models

Our first class of models is a generalization of the additive models introduced in Mitchell and Lapata (2008):

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{W}[\mathbf{x}; \mathbf{y}] \tag{1}$$

where the notation $[\mathbf{x}; \mathbf{y}]$ represents a vertical (row-wise) concatenation of two vectors; namely, the concatenation that results in a $2N \times 1$-sized vector. In addition to the $N \times V$ parameters for the word vector representations that are provided *a priori*, this model introduces $N \times 2N \times T$ parameters, where T is the number of POS-tag pairs we consider.

Mitchell and Lapata (2008) significantly simplify parameter estimation by assuming a certain structure for the parameter matrix $\mathbf{W}$, which is necessary given the limited human-annotated data they use. For example, by assuming a block-diagonal structure, we get a scaled element-wise addition model $f_i(x_i, y_i) = \alpha_i x_i + \beta_i y_i$. While not strictly in this category due to the non-linearities involved, neural network-based compositional models (Socher et al., 2013; Hermann and Blunsom, 2013) can be viewed as concatenation models, although the order of concatenation and matrix multiplication is switched. However, these models introduce more than $V \times N^2$ parameters.

2.2 Tensor Models

The second class of models leverages pairwise multiplicative interactions between the components of the two word vectors:

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = (\mathbf{W} \times_3 \mathbf{y}) \times_2 \mathbf{x} \tag{2}$$

where $\times_n$ corresponds to a tensor contraction along the n^{th} mode of the tensor $\mathbf{W}$. In this case, we first compute a contraction (tensor-vector product) between $\mathbf{W}$ and $\mathbf{y}$ along $\mathbf{W}$'s third mode, corresponding to interactions with the second word vector of a two-word phrase and resulting in a matrix, which is then multiplied along its second mode (corresponding to traditional matrix multiplication on the right) by $\mathbf{x}$. The final result is an $N \times 1$ vector. This model introduces $N \times N \times N \times T$ parameters.

Tensor models are a generalization of the element-wise multiplicative model (Mitchell and Lapata, 2008), which permits non-zero values only on the tensor diagonal. Operating at the vocabulary level, the model of Baroni and Zamparelli (2010) has interesting parallels to our tensor model. They focus on adjective–noun relationships and learn a specific matrix for every adjective in their dataset; in our case, the specific matrix for each adjective has a particular form, namely that it can be factorized into the product of a tensor and a vector; the tensor corresponds to the actual adjective–noun combiner function, and the vector corresponds to specific lexical information

[2] We discuss handling phrases longer than 2 words in §2.3.

that the adjective carries. This concept generalizes to other POS pairs: for example, multiplying the tensor that represents determiner-noun combinations along the second mode with the vector for "the" results in a matrix that represents the semantic operation of definiteness. Learning these parameters jointly is statistically more efficient than separately learning versions for each word.

2.3 Longer Phrases

The proposed models operate on pairs of words at a time. To handle phrases of length greater than two, we greedily construct a left-branching tree of the phrase constituents that eventually dictates the application of the learned bilinear maps.[3] For each internal tree node, we consider the POS tags of its children: if the right child is a noun, and the left child is either a noun, adjective, or determiner, then the internal node is marked as a noun, otherwise we mark it with a generic *other* tag. At the end of the procedure, unattached nodes (words) are attached at the highest point in the tree.

After the tree is constructed, we can compute the overall phrasal representation in a bottom-up manner, guided by the labels of leaf and internal nodes. We note that the emphasis of this work is not to compute sentence-level representations. This goal has been explored in recent research (Le and Mikolov, 2014; Kalchbrenner et al., 2014), and combining our models with methods presented therein for sentence-level representations is relatively straightforward.

3 Learning

The models described above rely on parameters $\mathbf{W}$ that must be learned. In this section, we argue that automatically constructed databases of paraphrases should provide adequate supervision for learning notions of compositionality.

3.1 Supervision from Automatic Paraphrases

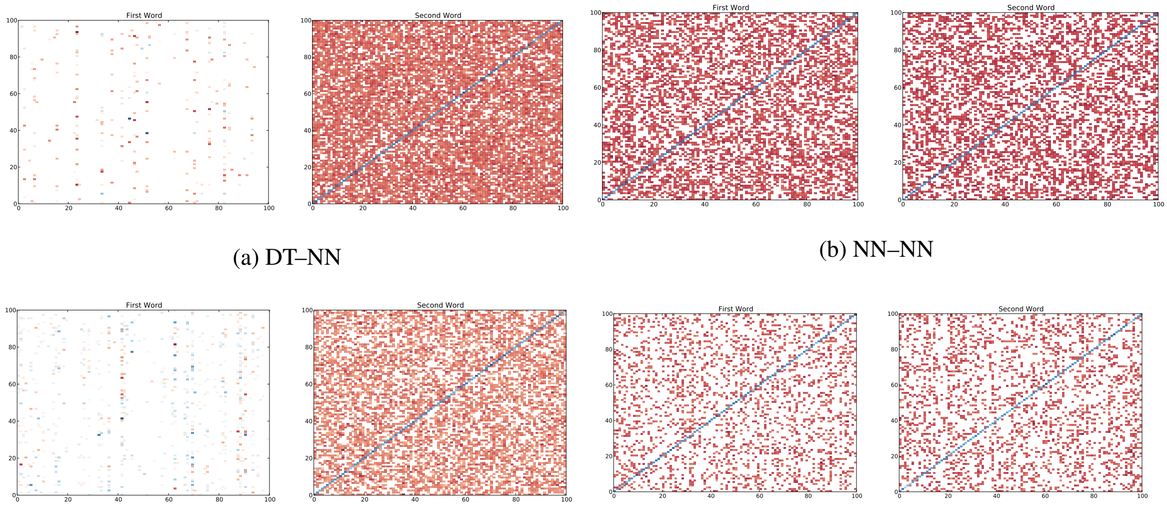
The Paraphrase Database (Ganitkevitch et al., 2013, PPDB) is a collection of ranked monolingual paraphrases that have been extracted from word-aligned parallel corpora using the bilingual pivot method (Bannard and Callison-Burch, 2005). The underlying assumption is that if two strings in the same language align to the same string in another language, then the strings in the original language share the same meaning. Paraphrases are ranked by their word alignment scores, and in this work we use the preselected SMALL portion of PPDB as our training data. Although we can directly extract phrasal representations of a pre-specified list of phrases from the corpus used to compute word representations (Baroni and Zamparelli, 2010), this approach is both computationally and statistically inefficient: the number of phrases increases exponentially in the length of the phrase, and correspondingly the occurrence of any individual phrase decreases exponentially. We can thus circumvent these computational and statistical issues by using monolingual paraphrases.

The training data is filtered to provide only two-to-one word paraphrase mappings, and the multiword portion of the paraphrase is subsequently POS-tagged. Table 1 provides a breakdown of such paraphrases by their POS pair type. Given the lack of context when tagging, it is likely that the POS tagger yields the most probable tag for words and not the most probable tag given the (limited) context. Furthermore, even the higher quality portions of PPDB yield paraphrases of ranging quality, ranging from non-trivial mappings such as *young people* → *youth*, to redundant ones like *the ceasefire* → *ceasefire*. However, PPDB-like resources are more easily available than human-annotated resources (in multiple languages too: Ganitkevitch and Callison-Burch (2014)), so it is imperative that methods which learn compositional functions from such sources handle noisy supervision adequately.

POS Pair	Size
DT–NN	10,982
NN–NN	4781
JJ–NN	3924
VB–VB	2021
RB–JJ	1640
other	8548

Table 1: Number of paraphrase examples per POS pair type out of the two-to-one word paraphrases in the SMALL version of PPDB (using the Penn Treebank tag-set). We distinguish between the five most common POS pair types, and group the remaining pairs into the generic *other* category.

[3] We also tried constructing right-branching trees, but found that performance was never as good as the left-branching ones.



(a) DT–NN

(b) NN–NN

(c) VB–VB

(d) *other*

Figure 1: Parameter heat-maps for specific POS pair compositional functions. Positive values are blue, negative values red, and zero values are white. Certain phrasal relationships (e.g., DT-NN and VB-VB) exhibit headedness.

3.2 Parameter Estimation

The parameters **W** in Eq. 1 and 2 can be estimated through standard linear regression techniques in conjunction with the data presented in §3.1. These methods provide a natural way to regularize **W** via ℓ_2 (ridge) or ℓ_1 (LASSO) regularization, which also helps handle noisy paraphrases. Parameters for the ℓ_1-regularized concatenation model for select POS pairs are displayed in Fig. 1.[4] The heat-maps display the relative magnitude of parameters, with positive values colored blue, negative values colored red, and white cells indicating zero values. It is evident that the parameters learned from PPDB indicate a notion of linguistic headedness, namely that for particular POS pairs, the semantic information is primarily contained in the right word, but for others such as the noun–noun combination, each constituent's contribution is relatively more equal.

4 Measuring of Compositionality

The concatenation and tensor models compute an N-dimensional vector representation for a multi-word phrase by assuming the meaning of the phrase can be expressed in terms of the meaning of its constituents. This assumption holds true to varying degrees; while it clearly holds for "large amount" and breaks down for "cloud nine", it is partially valid for phrases such as "zebra crossing" or "crash course". In line with previous work, we assume a **compositionality continuum** (McCarthy et al., 2003), but further conjecture that a phrase's level of compositionality is dependent on the specific context in which it occurs, motivating a context-based approach (§4.2) which scores compositionality by computing the likelihoods of surrounding context words given a phrase representation. The effect of context is directly measured through a comparison with context-independent methods from prior work (Bannard et al., 2003; Reddy et al., 2011)

It is important to note that most prior work on compositionality scoring assumes access to *both* word and phrase vector representations (for select phrases that will be evaluated) *a priori*. The latter are distinct from representations that are computed from learned compositional functions as they are extracted directly from the corpus, which is an expensive procedure. Our aim is to develop compositional models that are applicable in downstream tasks, and thus assuming pre-existing phrase vectors is unreasonable.[5] Hence for phrases, we only rely on representations computed from our learned compositional functions.

[4] Parameters learned with ℓ_2 regularization yield too many non-zero values, making visualization less informative.

[5] If these phrase representations were easy to extract from corpora, that would obviate the need to learn compositional functions.

4.1 At the Type Level

Given vector representations for the constituent words in a phrase and the phrase itself, the idea behind the type-based model is to compute similarities between the constituent word representations and the phrasal representation, and average the similarities across the constituents. If the contexts in which a constituent word occurs, as dictated by its vector representation, are very different from the contexts of the composed phrase, as indicated by the cosine similarity between the word and phrase representations, then the phrase is likely to be non-compositional. Assuming unit-normalized word vectors $\mathbf{x}, \mathbf{y}$ and phrase vector $\mathbf{z} = \mathbf{f}(\mathbf{x}, \mathbf{y})$ computed from one of the learned models in §2:

$$g(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \alpha(\mathbf{x} \cdot \mathbf{z}) + (1 - \alpha)(\mathbf{y} \cdot \mathbf{z}) \qquad (3)$$

where α is a hyperparameter that controls the contribution of individual constituents. This model leverages the average statistics computed over the training corpora (as encapsulated in the word and phrase vectors) to detect compositionality, and is the primary way compositionality has been evaluated previously (Reddy et al., 2011; Kiela and Clark, 2013).

4.2 At the Token Level

Eq. 3 scores phrases for compositionality regardless of the context that these phrases occur in. However, phrases such as “big fish” or “heavy metal” may occur in both compositional and non-compositional situations, depending on the nature and topic of the texts they occur in.[6] Here, we propose a context-driven model for compositionality detection, inspired by the skip-gram model for learning word representations (Mikolov et al., 2013). The intuition is simple: if a phrase is compositional, it should be sufficiently predictive of the context words around it; otherwise, it is acting in a non-compositional manner. Thus, we would like to compute the likelihood of the context ($\boldsymbol{c}$) given a phrasal representation ($\mathbf{z} = \mathbf{f}(\mathbf{x}, \mathbf{y})$) and normalization constant Z:

$$P(\boldsymbol{c} \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{|\boldsymbol{c}|} \frac{\exp \mathbf{f}(\mathbf{x}, \mathbf{y}) \cdot \boldsymbol{v}_{c_i}}{Z(\mathbf{x}, \mathbf{y})}. \qquad (4)$$

As explained in Goldberg and Levy (2014), the context representations are distinct from the word representations. In practice, we compute the log-likelihood averaged over the context words or the perplexity instead of the actual likelihood.

5 Evaluation

Our experiments had three aims: first, demonstrate that the compositional functions learned using paraphrase supervision compute semantically meaningful results for compositional phrases by evaluating on a phrase similarity task (§5.1); second, verify the hypothesis that compositionality is context-dependent by comparing a type-based and token-based approach on a compound noun evaluation task (§5.2); and third, determine if the compositionality-scoring models based on learned representations improve the translations produced by a state-of-the-art phrase-based MT system (§5.3).

The word vectors used in all of our experiments were produced by `word2vec`[7] using the skip-gram model with 20 negative samples, a context window size of 10, a minimum token count of 3, and sub-sampling of frequent words with a parameter of 10^{-5}. We extracted corpus statistics for `word2vec` using the AFP portion of the English Gigaword[8], which consists of 887.5 million tokens. The code used to generate the results is available at `http://www.github.com/xyz`, and the evaluation datasets are publicly available.

5.1 Phrasal Similarity

For the phrase similarity task we first compare our concatenation and tensor models learned using ℓ_1 and ℓ_2 regularization to three baselines:

- ADD: $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{x} + \mathbf{y}$
- MULT1: $f_i(x_i, y_i) = x_i y_i$
- MULT2: $f_i(x_i, y_i) = \alpha_i x_i y_i$

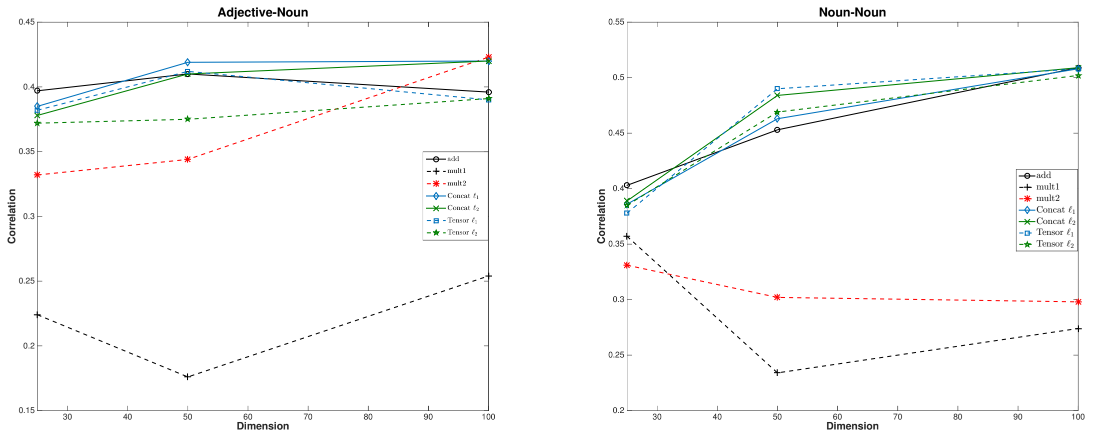
Other additive models from previous work (Mitchell and Lapata, 2010; Zanzotto et al., 2010; Blacoe and Lapata, 2012) that impose varying amounts of structural assumptions on the semantic interactions between word representations e.g., $f_i(x_i, y_i) = \alpha_i x_i + \beta_i y_i$ or $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \alpha \mathbf{x} + \beta \mathbf{y}$ are subsumed by our concatenation model. The regularization strength hyperparameter for ℓ_1 and ℓ_2 regularization was selected using 5-fold cross-validation on the PPDB training data.

We evaluated the phrase compositionality models on the adjective–noun and noun–noun phrase

[6] In fact, human annotators have access to such context when making compositionality judgments.

[7] `http://code.google.com/p/word2vec`

[8] LDC2011T07



(a) JJ–NN

(b) NN–NN

Figure 2: Spearman's ρ correlation with respect to human judgments for the adjective–noun and noun–noun phrase similarity tasks. Dashed lines correspond to tensor models (and baselines), and solid lines are concatenation models (and additive baseline).

similarity tasks compiled by Mitchell and Lapata (2010), using the same evaluation scheme as in the original work.[9] Spearman's ρ between phrasal similarities derived from our compositional functions and the human annotators (computed individually per annotator and then averaged across all annotators) was the evaluation measure.

Figure 2 presents the correlation results for the two POS pair types as a function of the dimensionality N of the representations for the concatenation models (and additive baseline) and tensor models (and multiplicative baselines). The concatenation models seem more effective than the tensor models in the adjective–noun case and give roughly the same performance on the noun–noun dataset, which is consistent with previous work that uses dense, low-dimensional representations (Guevara, 2011; Hermann and Blunsom, 2013; Hashimoto et al., 2014).[10] Since the concatenation model involve fewer parameters, we use it as the compositional model of choice for subsequent experiments. The absolute results are also consistent with state-of-the-art results on this dataset[11] (Blacoe and Lapata, 2012; Hashimoto et al., 2014), indicating that paraphrases are an excellent source of information for learning compositional functions and a reasonable alternative to human-annotated training sets. For reference, the inter-annotator agreements are 0.52 for the adjective–noun evaluation and 0.51 for the noun–noun one. The unweighted additive baseline is surprisingly very strong on the noun–noun set, so we also compare against it in subsequent experiments.

5.2 Compositionality

To evaluate the compositionality-scoring models, we used the compound noun compositionality dataset introduced in Reddy et al. (2011). This dataset consists of 2670 annotations of 90 compound-noun phrases exhibiting varying levels compositionality, with scores ranging from 0 to 5 provided by 30 annotators. It also contains three to five example sentences of these phrases that were shown to the annotators, which we make use of in our context-dependent model. Consistent with the original work, Spearman's ρ is computed on the averaged compositionality score for a phrase across all the annotators that scored that phrase (which varies per phrase). For computing the compositional functions, we evaluate three of the best performing setups from §5.1: the ℓ_1 and ℓ_2-regularized concatenation models, and the simple additive baseline.

For the context-independent model, we select the hyperparameter α in Eq. 3 from the values $\{0.25, 0.5, 0.75\}$. For the context-dependent

[9] The evaluation set also consists of verb-object phrases constructed from dependency relations and their similarity, but such phrases generally do not fall into our phrasal definition since the words are not contiguous.

[10] Multiplicative or tensor-based models seem to do better on sparse, high-dimensional representations (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010) since multiplication represents a conjunction of co-occurrence features.

[11] There are differences in the corpora and experimental setup which explains the small discrepancies.

		ρ	
Setup	**ADD**	**Concat. ℓ_1**	**Concat. ℓ_2**
Type $\alpha = 0.25$	|	0.43	0.46
Type $\alpha = 0.5$	0.41	0.42	0.47
Type $\alpha = 0.75$	|	0.41	0.43
Token $l = 4$	**0.55**	**0.58**	0.58
Token $l = 6$	0.54	0.57	**0.59**
Token $l = 8$	0.53	0.58	0.59

Table 2: Correlation between model judgments on phrase compositionality and human judgments, measured by Spearman's ρ. Context-dependent (token-based) and concatenation models do better.

model, we vary the context window size $|\boldsymbol{c}|$ by selecting from the values $\{4, 6, 8\}$. Table 2 presents Spearman's ρ for these setups. Note that for the simple additive model with unit-normalized word vectors, the cosine similarity score is independent of the hyperparameter α. In all cases, the context-dependent models outperform the context-independent ones, and using a relatively simple token-based model we can match the performance of the Bayesian model proposed by Hermann et al. (2012). The concatenation models are also consistently better than the additive compositional model, indicating the benefit of learning the compositional parameters via PPDB.

5.3 Machine Translation

While any truly successful model of semantics must match human intuitions, understanding the applications of our models is likewise important. To this end, we consider the problem of machine translation, operating under the hypothesis that sentences which express their meaning non-compositionally should also translate non-compositionally.

Modern phrase-based translation systems are faced with a large number of possible segmentations of a source-language sentence during decoding, and all segmentations are considered equally likely (Koehn et al., 2003). Thus, it would be helpful to provide guidance on more likely segmentations, as dictated by the compositionality scores of the phrases extracted from a sentence, to the decoder. A low compositionality score would ideally force the decoder to consider the entire phrase as a translation unit, due to its unique semantic characteristics. Correspondingly, a high score informs the decoder that it is safe to rely on word-level translations of the phrasal constituents. Thus, if we reveal to the translation system that a phrase is non-compositional, it should be able to learn that translation decisions which translate it as a unit are to be favored, leading to better translations.

To test this hypothesis, we built an English-Spanish MT system using the CDEC decoder (Dyer et al., 2010) for the entire training pipeline (word alignments, phrase extraction, feature weight tuning, and decoding). Corpora from the WMT 2011 evaluation[12] was used to build the translation and language models, and for tuning (on `news-test2010`) and evaluation (on `news-test2011`), with scoring done using BLEU (Papineni et al., 2002). The baseline is a hierarchical phrase-based system (Chiang, 2007) with a 4-gram language model, with feature weights tuned using MIRA (Chiang, 2012). For features, each translation rule is decorated with two lexical and phrasal features corresponding to the forward $(e|f)$ and backward $(f|e)$ conditional log frequencies, along with the log joint frequency (e, f), the log frequency of the source phrase (f), and whether the phrase pair or the source phrase is a singleton. Weights for the language model, glue rule, and word penalty are also tuned. This setup (Baseline) achieves scores *en par* with the published WMT results.

We added the compositionality score as an additional feature, and also added two binary-valued features: the first indicates if the given translation rule has not been decorated with a compositionality score (either because it consists of non-terminals only or the lexical items in the translation rule are unigrams), and correspondingly the second feature indicates if the translation rule has been scored. Therefore, an appropriate additional baseline would be to mark translation rules with these indicator functions but without the scores, akin to identifying rules with phrases in them (Baseline + SegOn).

Table 3 presents the results of the MT evaluation, comparing the baselines to the best-performing context-independent and dependent scoring models from §5.2. The scores have been averaged over three tuning runs with standard deviation in parentheses; bold results on the test set are statistically significant ($p < 0.05$) with respect to the baseline. While knowledge of relative compositionality consistently helps, the improvements using the context-dependent scoring models, especially with the ℓ_2 concatenation model, are noticeably better.

[12] `http://www.statmt.org/wmt11/`

	BLEU	
Setup	Dev	Test
Baseline	25.23 (0.05)	26.89 (0.13)
Baseline + SegOn	25.15 (0.21)	26.87 (0.19)
ℓ_2 CosSim $\alpha = 0.5$	25.08 (0.03)	**26.99** (0.04)
ADD $l = 4$	24.85 (0.12)	26.82 (0.05)
ℓ_1 $l = 4$	25.08 (0.08)	**27.03** (0.10)
ℓ_2 $l = 6$	25.12 (0.22)	**27.26** (0.21)

Table 3: MT results. Bold results are statistically significant, and our best context-dependent setup is 0.4 BLEU points better than the baseline.

6 Related Work

There has been a large amount of work on compositional models that operate on vector representations of words. With some exceptions (Mitchell and Lapata, 2008; Mitchell and Lapata, 2010), all of these approaches are lexicalized i.e., parameters (generally in the form of vectors, matrices, or tensors) for *specific* words are learned, which works well for frequently occurring words but fails when dealing with compositions of arbitrary word sequences containing infrequent words. The functions are either learned with a neural network architecture (Socher et al., 2013, *inter alia*) or as a linear regression (Baroni and Zamparelli, 2010); the latter require phrase representations extracted directly from the corpus for supervision, which can be computationally expensive and statistically inefficient. In contrast, we obtain this information through many-to-one PPDB mappings. Most of these models also require additional syntactic (Socher et al., 2012) or semantic (Hermann and Blunsom, 2013; Grefenstette et al., 2013) resources; on the other hand, our proposed approach only requires a shallow syntactic parse (POS tags). Recent efforts to make these models more practical (Paperno et al., 2014) attempt to reduce their statistically complex and overly-parametrized nature, but with the exception of Zanzotto et al. (2010), who propose a way to extract compositional function training examples from a dictionary, these models generally require human-annotated data to work.

Most models that score the relative (non-) compositionality of phrases do so in a context-independent manner. A central idea is to replace phrase constituents with semantically-related words and compute the similarity of the new phrase to the original (Kiela and Clark, 2013; Salehi et al., 2014) or make use of a variety of lexical association measures (Lin, 1999; Pecina and Schlesinger, 2006). Sporleder and Li (2009) however, do make use of context in a token-based approach, where the context in which a phrase occurs as well as the phrase itself is modeled as a lexical chain, and the cohesion of the chain is measured as an indicator of a phrase's compositionality. Cohesion is computed using a web search engine-based measure, whereas we use a probabilistic model of context given a phrase representation. Hermann et al. (2012) propose a Bayesian generative model that is also context-based, but learning and inference is done through a relatively expensive Gibbs sampling scheme.

In the context of MT, Zhang et al. (2008) present a Bayesian model that learns non-compositional phrases from a synchronous parse tree of a sentence pair. However, the primary aim of their work is phrase extraction for MT, and the non-compositional constraints are only applied to make the space of phrase pairs more tractable when bootstrapping their phrasal parser from their word-based parser. In contrast, we score every phrase that is extracted with the standard phrase extraction heuristics (Chiang, 2007), allowing the decoder to make the final decision on the impact of compositionality scores in translation. Thus, our work is more similar to Xiong et al. (2010), who propose maximum entropy classifiers that mark positions between words in a sentence as being a phrase boundary or not, and integrate these scores as additional features in an MT system.

7 Conclusion

In this work, we presented two new sources of information for compositionality modeling and scoring, paraphrase information and context. For modeling, we showed that the paraphrase-learned compositional representations performs as well on a phrase similarity task as the average human annotator. For scoring, the importance of context was shown through the comparison of context-independent and dependent models. Improvements by the context-dependent model on an extrinsic machine translation task corroborate the utility of these additional knowledge sources. We hope that this work encourages further research in making compositional semantic approaches applicable in downstream tasks.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of EMNLP-CoNLL*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*.

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

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL*.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *Proceedings of LREC*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR: abs/1402.3722*.

Edward Grefenstette, Georgiana Dinu, Yao-Zhang Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2013. Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS*.

Emiliano Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of IWCS*.

Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2014. Jointly learning word representations and composition functions using predicate-argument structures. In *Proceedings of EMNLP*.

Karl Moritz Hermann and Phil Blunsom. 2013. The role of syntax in vector space models of compositional semantics. In *Proceedings of ACL*.

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of *SEM*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of ACL*.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of EMNLP*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR: abs/1310.4546*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-HLT*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*.

Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of COLING-ACL*.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceddings of CICLing*.

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, Department of Linguistics, Stockholm University.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of EACL*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of EMNLP*.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings EMNLP*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of EACL*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Deyi Xiong, Min Zhang, and Haizhou Li. 2010. Learning translation boundaries for phrase-based decoding. In *Proceedings of NAACL-HLT*.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*.

Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proceedings of ACL-HLT*.