Preposition Sense Disambiguation and Representation

Hongyu Gong, Jiaqi Mu, Suma Bhat, Pramod Viswanath
University of Illinois at Urbana-Champaign
{hgong6, jiaqimu2, spbhat2, pramodv}@illinois.edu

Abstract

Prepositions are highly polysemous, and their variegated senses encode significant semantic information. In this paper we match each preposition's left- and right context, and their interplay to the geometry of the word vectors to the left and right of the preposition. Extracting these features from a large corpus and using them with machine learning models makes for an efficient preposition sense disambiguation (PSD) algorithm, which is comparable to and better than state-of-the-art on two benchmark datasets. Our reliance on no linguistic tool allows us to scale the PSD algorithm to a large corpus and learn sense-specific preposition representations. The crucial abstraction of preposition senses as word representations permits their use in downstream applications–phrasal verb paraphrasing and preposition selection–with new state-of-the-art results.

1 Introduction

English prepositions form a closed class showing no inflectional variation and are some of the most frequent words. A computational-linguistic understanding of prepositions remains challenging owing to their highly polysemous nature and frequent participation in idiomatic expressions (Saint-Dizier, 2006). In this paper, we study the problem of sense disambiguation for prepositions.

She blinked **with** *confusion*. (Manner & Mood)
He *combines* professionalism **with** humor. (Accompanier)
He *washed* a small red teacup **with** *water*. (Means)

Table 1: Examples showing polysemous behavior of *with* and the TPP senses.

The highly polysemous nature of prepositions drives several syntactic and semantic processes. For instance, the preposition *with* has 18 senses listed in The Preposition Project (TPP) (Litkowski and Hargraves, 2005), examples of which, are shown in Table 1. We notice that *with* indicates an emotional state in *with confusion* and refers to an accompanier in *combine with*, while it suggests the idea of a tool or means in *wash with water*. Thus, preposition sense disambiguation (PSD) is vital for natural language understanding and a closer look at the function of prepositions in specific contexts is an important computational step.

Previous approaches to PSD (for instance, (Ye and Baldwin, 2007; Hovy et al., 2011)) have relied on linguistic tools and resources (the minimum of which involves dependency parsers and POS taggers) to capture the crucial contextual information of prepositions. We depart from prior art by using *no linguistic resources or tools* other than a set of word representations (trained on a large corpus). We interpret preposition senses as groups of similar contexts, where each instance of the preposition 'sense' is represented as a vector of context-dependent features. We find a simple feature extraction process that creatively harnesses the *geometry* of word representations and contributes to a scalable PSD algorithm. Our algorithm can reach near and even beat state-of-the-art performance on two benchmark datasets (SemEval 2007 and OEC); this is true in both unsupervised and supervised PSD settings.

A PSD algorithm that *efficiently scales to a large corpus* naturally paves the way for distributed representations of the preposition senses: we enrich the corpus with sense-specific information of prepositions using our PSD algorithm. Next, we repurpose an off-the-shelf word representation algorithm (Word2vec (Mikolov et al., 2013)) to relearn word representations with the key aspect that the length of the context surrounding the prepositions is crucially reduced. Sense-specific preposition representations thus learnt are strongly validated by using them in two

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1510–1521
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

applications–phrasal verb paraphrasing and preposition selection–using available datasets. We released our PSD system and paraphrasing dataset [1] available.

We summarize our contributions below:

- **Novel Perspective of Preposition Behavior**: We provide a novel selectional aspect of the context that best represents the sense of a preposition, where we match classical ideas from linguistics with the appropriate geometry of word embeddings. The standard view focuses on the left context (attachment) and the right context (complement) of the preposition; in this paper, we include the *interplay* between these two elements via an appropriate geometric representation.

- **Resource-independent Disambiguation**: We rely only on a set of trained word representations and no other language processing tool, where almost all prior approaches have included at least POS tagging and dependency parsing. Our results are comparable to, or better than, state-of-the-art on standard benchmarks.

- **Preposition Sense Representation Learning**: To the best of our knowledge, this is the first work on preposition sense representation. The power of our sense representation is reflected in the experimental comparisons with strong baseline approaches to phrasal verb paraphrasing and preposition selection, where we demonstrate the superiority of our approach that uses sense representation of prepositions.

2 Related Works

We place our work in the context of related studies in preposition representation and **Preposition Sense Disambiguation**: Preposition disambiguation has been explored on the SemEval dataset via various methods and external resources (part of speech taggers, chunkers, dependency parsers, named entity extractors, WordNet based supersense taggers and semantic role labelers) since 2007 (Yuret, 2007; Ye and Baldwin, 2007; Tratz and Hovy, 2009; Hovy et al., 2011; Popescu et al., 2007; Tratz and Hovy, 2011; Srikumar and Roth, 2013). Recently, (Gonen and Goldberg, 2016) use a multilingual parallel corpus processed using sequence to sequence neural networks for preposition disambiguation and achieve an accuracy within 5% of the state-of-the-art, which includes (Litkowski, 2013; Hovy et al., 2010; Srikumar and Roth, 2013). We note that we achieve the comparable performance as (Gonen and Goldberg, 2016) using *only* word embeddings.

Preposition Representation: Representation learning is fundamental to machine learning models (Wu et al., 2018a,b; Xu et al., 2018). Word embeddings such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been widely recognized for their ability to capture linguistic regularities (including syntactic and semantic relations). On the other hand, no linguistic property of their prepositional embeddings is known; to the best of our knowledge, we propose the *first* sense-specific prepositional embeddings and demonstrate their linguistic regularities. A recent unsupervised approach by Gong et al. learns preposition representations to encode the syntactics and semantics by capturing their attachment and complement properties. Distantly related is (Hashimoto and Tsuruoka, 2015), which learns embeddings of prepositions acting as verb adjuncts by the factorization of a predicate tensor. Similarly, (Belinkov et al., 2014) explores the use of preposition representations optimized for the task of prepositional phrase attachment, but do not analyze their sense-specificity.

Sense-specific Representations: Several prior studies have sought polysemy-aware alternatives to word representations that take into account the context of the target word, including (Erk and Padó, 2008; Mitchell and Lapata, 2008; Reisinger and Mooney, 2010a; Thater et al., 2011; Dinu et al., 2012). More recently polysemy disambiguation for word embeddings have been proposed using external resources such as WordNet (Rothe and Schütze, 2015) or in an unsupervised way (Mu et al., 2017; Song et al., 2016; Arora et al., 2016; Neelakantan et al., 2014) with the latter two limiting the number of senses and validated for only nouns and verbs. The approach of (Neelakantan et al., 2014) is roughly similar to our baseline method using the average context vector. Our unsupervised approach is similar to that in (Reisinger and Mooney, 2010a), but limited to prepositions and uses novel features described next.

[1] https://github.com/HongyuGong/PrepositionSenseDisambiguation.git

3 Preposition Sense Disambiguation

The key intuition behind our sense disambiguation approach is the modern descriptive linguistic view (Huddleston, 1984; DeCarrico, 2000): the sense of a preposition in any sentence is driven by both its *attachment* and its *complement*; classical prescriptive linguistics had focused only on the latter (Beal, 2004), pp. 110, (Cobbett, 1823), pp. 16, (Lowth, 1762), pp. 8, 91.

Referring again to the examples in Table 1 we point out that italicized words determine the sense of “with.” In the first sentence, the word ‘confusion,’ appearing as the right context of the preposition, is the complement of ‘with’, from which we infer that ‘with’ encodes the sense of ‘manner’. In the second sentence, the accompanier sense of ‘with’ is because of its governor (attachment), the verb ‘combine’ appearing in the left context. In the last sentence, the sense of ‘with’ is ‘by means of’ and is determined by *both* the verb in its left context and the argument in its right context. Consider a new sentence with changed right context: ‘He washed a small cup with a handle.’ Here ‘with’ functions as an attribute. Again, changing its left context we get the sentence ‘He asked for a small cup with water’, where ‘with’ serves as an attribute instead of encoding the sense of means.

That the *left* and *right* context and their *interplay* are critical to prepositional sense disambiguation is also well established in the literature (Hovy et al., 2011; Litkowski and Hargraves, 2007). We match these linguistic properties to appropriate *geometric* objects within the space of word embeddings; the word embeddings are borrowed off-the-shelf – this work uses word2vec. We describe this next, focusing first on the left context, next on the right context and then on their interplay.

Left context feature v_ℓ is the average of the vectors of the left k_ℓ words (here k_ℓ is a parameter roughly taking values 1 through 4). This simple geometric operation is motivated by recent works (Faruqui et al., 2015; Kenter et al., 2016; Yu et al., 2014) representing a sentence by the average of its constituent words robustly and successfully in a variety of downstream settings. Although prior work (Hovy et al., 2010) points out that fixed window sizes are insufficient, when compared to using specific syntactic features (e.g., POS tags and dependency as done in prior works), we will see that the semantic information embedded in word vectors largely compensates for this limitation.

Right context feature v_r is the average of the vectors of the right k_r words (here k_r is a parameter roughly taking values 1 through 4). This is identical to the method adopted for the left context.

We model the **Context-interplay feature** v_{inter} to geometrically relate to both the left and the right contexts as follows. We choose it to be the vector *closest* to both the subspace spanned by the left context word vectors and that spanned by the right context word vectors. This geometric representation appears crucial to capture the prepositional-sense when the interplay between the contexts matters decisively, as seen empirically in our extensive experiments.

Let v_i^ℓ and v_j^r be the left- and the right context word vectors respectively. A precise mathematical definition of v_{inter} is below:

$$v_{\text{inter}} = \operatorname*{argmin}_{v:\|v\|_2=1} \Big(\min_{a_1,\ldots,a_{k_\ell}} \|v - \sum_{i=1}^{k_\ell} a_i v_i^\ell\|_2^2 + \min_{b_1,\ldots,b_{k_r}} \|v - \sum_{j=1}^{k_r} b_j v_j^r\|_2^2 \Big) , \quad (1)$$

where $\{a_i\}_{i=1}^{k_\ell}$ and $\{b_j\}_{j=1}^{k_r}$ are scalars. It is easy to find optimal $\{a_i^*\}$ and $\{b_j^*\}$ to solve the inner minimization problem. We have $a_i^* = \frac{v^T v_i^\ell}{\|v_i^\ell\|^2}$, and $b_j^* = \frac{v^T v_j^r}{\|v_j^r\|^2}$.

The minimization problem (1) is a quadratic optimization problem, so we can find a closed form solution to the unit vector v_{inter}. Suppose that we stack context word vectors $\{v_i^\ell\}$ and $\{v_j^r\}$ as a matrix $V_{d\times(k_\ell+k_r)}$, where d is the dimension of word vectors. The optimal $d-$dimension vector v_{inter} is the first principal component of matrix V.

Unsupervised learning of the senses of a given preposition is conducted by *clustering* its instances represented as a concatenation of the three feature vectors, while harnessing the large number of instances of each preposition in the large WikiCorpus (here we fix $k_\ell = k_r = 2$ and use k-means clustering). If the features do capture the prepositional sense efficiently, then the same-sense instances belong to the same cluster. Based on this intuition, we label each cluster with the dominant label of the training instances within this cluster. Given a test instance, we assign it to the nearest cluster (based on its Euclidean distance), and tag it with the cluster label. We note that the preposition senses are not balanced in the training dataset leading to a situation where frequent senses dominate more than one cluster. We address this by setting the number of clusters k equal to twice the number of senses and find that it separates the infrequent senses from frequent ones, about as well as traditional approaches such as those based on information criteria (Sugar and James, 2003) and the elbow method (Ketchen Jr and Shook, 1996).

Supervised learning of the senses using the three feature vectors was conducted based on the training examples provided in the benchmark PSD datasets. We did this using the standard support vector machines (SVM) (Cortes and Vapnik, 1995), multilayer perceptron (MLP) (Glorot and Bengio, 2010) and weighted k-nearest neighbor (k-NN) (Andoni and Indyk, 2006) classifiers. Each of these allows potentially different weighting of the three features in a context dependent way. The parameters were tuned to maximize the disambiguation accuracy on the development set provided in the benchmark PSD datasets. These experiments are discussed in detail next.

4 Experiments on Sense Disambiguation

The PSD algorithms were validated on the general sense disambiguation task using two datasets provided by TPP. We begin by introducing two benchmarks: SemEval and OEC datasets.

The **SemEval Dataset** consists of 34 prepositions instantiated by $24,663$ sentences covering 332 senses. Among them, $16,557$ sentences are used as training instances (**semtrain**) and 8096 sentences are test instances (**semtest**) for the preposition disambiguation task.

The **OEC dataset** consists of $7,650$ sentences collected from the Oxford English Corpus. Since these sentences included more prepositions than those in the SemEval dataset, we chose $3,587$ sentences that included the same 34 prepositions as used in the SemEval task.

Word embeddings. The word embeddings we used in our experiments were trained on the most recent scrape of the English Wikipedia with the Word2Vec CBOW model (Mikolov et al., 2013), with dimension 300. The linear combination of three vectors v_ℓ, v_r and v_{inter} is the feature to k-NN classifier.

Unsupervised PSD was performed by clustering the training instances fom the SemEval dataset using k-means. In the evaluation phase, each test instance was assigned to the closest cluster, and its sense was the dominant training sense within this cluster. In Table 2 we report the disambiguation accuracy on **semtest**, a new state-of-the-art result.

Supervised PSD was conducted by first conducting a 80/20 split of **semtrain** into training and development sets. The disambiguation accuracy calculated on both **semtest** and OEC datasets is reported in Table 3, using standard off-the-shelf classifiers. The sense disambiguation can be regarded as a multi-class classification problem. We used the SVM classifier with a linear kernel and its penalty parameter C as a tunable parameter, the MLP classifier with one hidden layer, and the number of neurons as a tunable parameter, and the k-NN classifier (weighted k-NN), with the number of nearest neighbors and the feature weights as tunable parameters (a linear combination of the three vectors v_ℓ, v_r and v_{inter} was the feature input to the k-NN classifier); all tunable parameters were tuned using the development set. The output dimension of these classifiers was the number of senses of prepositions. Additionally, the context window sizes k_ℓ and k_r were parameters for all the three classifiers, each tuned on the development set.

Baseline. Recent works have shown that the average word embedding serves as a good representation of the compositional sentential semantics (Faruqui et al., 2015; Kenter et al., 2016; Yu et al., 2014), and this single feature – the average of all context word vectors (both to the left and the right) – serves as a natural baseline.

Results. In both the unsupervised and supervised disambiguation settings, the best performance is achieved by using *all three* features, v_ℓ, v_r and v_i.

As summarized in Table 2, our unsupervised method achieves a 2.4% improvement over state-of-art (Hovy et al., 2011). The results in the supervised setting, tabulated in Table 3, reveal that the weighted k-NN classifier performs best. Denoting left, right and interplay features by ℓ, r, i respectively, Table 2 and 3 report our experimental results using only subset combinations of these features on the two disambiguation tasks.

An ablation analysis of the features reveals that the context-interplay feature is most beneficial when testing on the OEC dataset, but on the SemEval dataset, the left context feature appears to be the most beneficial. A likely explanation to this behavior is that several instances in **semtrain** and **semtest** share the governors the prepo-

System	State-of-art	k-means clustering				
	(Hovy et al., 2011)	average	(ℓ, r)	(ℓ, i)	(r, i)	(ℓ, r, i)
Accuracy	0.56	0.555	0.561	0.565	0.534	**0.584**

Table 2: Performance of the unsupervised PSD compared with the state-of-the-art. (ℓ, i), (ℓ, r) and (r, i) correspond to feature ablation results.

	SemEval Dataset					OEC Dataset				
Feature Type	average	(ℓ, r)	(ℓ, i)	(r, i)	(ℓ, r, i)	average	(ℓ, r)	(ℓ, i)	(r, i)	(ℓ, r, i)
SVM	0.712	0.765	0.775	0.700	0.782	0.305	0.330	0.333	0.325	0.351
MLP	0.712	0.758	0.780	0.704	0.777	0.322	0.353	0.353	0.347	0.375
Weighted k-NN	0.731	0.781	0.792	0.733	**0.804**	0.329	0.341	0.380	0.367	**0.400**

Table 3: Supervised disambiguation on SemEval and OEC datasets.

sitions attach to. Hence the left feature (encoding the governor information) helps disambiguation on **semtest**. The governors and complements in OEC instances differ from those in **semtrain**. Therefore, the context-interplay feature provides more general context information than provided by the left and right context features by themselves for sense disambiguation on the OEC dataset.

A side-by-side comparison of the performance of our supervised approach with prior approaches is shown in Table 4. We note that the accuracy of our system is significantly better than that of the best PSD system in SemEval 2007 (11% higher accuracy), and 8% higher on the OEC dataset. It is noteworthy that while (Litkowski, 2013) fared better than our system with the SemEval data, our system outperformed (Litkowski, 2013) on the OEC dataset. Also we achieve performance comparable to the recent work (Gonen and Goldberg, 2016) which had access to a multilingual translation corpus (and other linguistic tools).

4.1 Spatial Expression Disambiguation

Prepositions such as ‘in’ and ‘on’ are used to encode spatial relations between the point of attachment and the complement of the preposition, but their senses show diversity depending on the context. For example, ‘on’ refers to the support from above in the sentence *clothes on the rack*, while it refers to support from below in *clothes on the desk*. These are instances of a phenomenon in natural language in which the mere concatenation of lexical information is not sufficient to derive the meaning of the phrase but the interactions among the meanings of the words is to be considered (termed compositional distributional semantic models (Marelli et al., 2014; Ritter et al., 2015)). We hypothesize that the relative placement of the objects involved in these phrases is achieved by considering the spatial senses of the prepositions involved, which in turn, is done by considering the interaction of its contexts as done by our approach. For this study, we focus on the disambiguation of the spatial senses encoded using the prepositions ‘in’ and ‘on’.

Dataset. (Ritter et al., 2015) studied ways of combining the meanings of the words in context to arrive at the meaning of the phrase which included spatial expressions using the prepositions ‘in’ and ‘on’. Their dataset consists of 420 training examples and 80 test examples, covering 5 types of locative expressions and given a sentence, the task is to arrive at the kind of locative expression encoded by the preposition. As examples, the preposition *in* refers to full containment in the sentence “an apple in the refrigerator”, whereas it refers to partial containment in “finger in the ring”. Similarly, the spatial relations represented by the preposition *on* is classified into three categories: adhesion to vertical surface (e.g., “sign on the building”), support by horizontal surface (e.g., “leaf on the ground”) and support from above (e.g., “bat on the branch”). A key observation here is that the spatial sense of the preposition (equivalently, spatial category) is a function of the two objects connected by the preposition.

Task. Given a sentence, we need to classify each occurrence of the preposition to its spatial category; for our study this is a fine-grained intra-preposition sense disambiguation problem (intra-sense because the sense is one of the spatial senses, albeit different from the TPP senses).

Method. The very small size of the training set in this experiment calls for a reduction in the feature dimension. Accordingly, we use a weighted linear combination of the three features used in our PSD algorithm and the resulting feature is $v = v_\ell + \beta v_r + \gamma v_{\text{inter}}$.

Dataset	System	Resources	Accuracy
SemEval	Our system	English corpus	0.804
	(Litkowski, 2013)	lemmatizer, dependency parser, WordNet	**0.86**
	(Srikumar and Roth, 2013)	dependency parser, WordNet	0.85
	(Gonen and Goldberg, 2016)	multilingual corpus, aligner, dependency parser	0.81
	(Ye and Baldwin, 2007)	chunker, dependency parser, named entity extractor, WordNet	0.69
OEC	Our system	English corpus	**0.40**
	(Litkowski, 2013)	lemmatizer, dependency parser, WordNet	0.32

Table 4: Preposition disambiguation performance comparison on SemEval and OEC datasets.

sense	1	2	3	4
closest words	backwards, reverse, angles, diagonal, between, forward	wearing, dress, hats, dresses, trousers, sleeves, pants, jacket	back, inside, underneath, from, into, where, onto	where, near, from, at, southern, northern, during
example	in all directions, move in, differ in	dress in black, in leather, in size	in the mail	in the UK, in Argentina
TPP sense	Manner_or_Degree	VariableQuality	ThingEntered	ThingEnclosed

Table 5: Example senses of the preposition "in".

Recall that the hyperparameters β and γ of the k-NN classifier on the PSD task were tuned on the (mismatched) SemEval development set. We then generated the weighted features $v = v_\ell + \beta v_r + \gamma v_{\text{inter}}$ using the tuned values of β, γ from the PSD task. Classification within the spatial sense disambiguation is now conducted using a Multi-Layer Perceptron (MLP) with the feature vector v.

Baseline. The state-of-the-art method (Ritter et al., 2015) used the inclusion of the left and the right noun vectors as features, which is equivalent to adding the left and the right context features with a context window size of 1 in our set-up; this serves as our baseline.

Our method achieves an accuracy of **77**%, a significant improvement over the baseline accuracy of 71% in (Ritter et al., 2015). We note that this improvement is achieved even though we tuned the hyperparameters β and γ on the mismatched (but relatively bigger) SemEval dataset. To complete the comparison, we found (via grid search) that the best β, γ values result in only a slightly higher accuracy of 79%. This performance adds credence to the conclusion that the geometric features ($v_\ell, v_r, v_{\text{inter}}$) do indeed represent the *preposition in its context* efficiently and accurately.

5 Preposition Sense Representation

Thus far, we have empirically validated our disambiguation algorithm on standardized, but still stylized, datasets. A more thorough analysis is enabled by conducting preposition sense disambiguation on a very large unlabeled corpus. Such is the goal of this section, where we scale our lightweight PSD algorithm on a large corpus and learn sense-specific prepositional representations. The quality of the representations serves as an "extrinsic" evaluation of our PSD algorithm; this validation is done by repurposing datasets meant for other tasks.

Standard embedding methods do not account for the inherent polysemy in words. This is exacerbated in the case of prepositions. Indeed, to the best of our knowledge, no linguistic properties of the standard embeddings (say, word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014)) are known for preposition vectors. Recent works that learn sense-specific embeddings use the distinct "topics" assumed by the senses of a given word (as in (Rothe and Schütze, 2015) that explicitly uses WordNet senses) and have only been validated with respect to nouns and verbs.

Below, we validate the quality of the resulting sense representations in two tasks, where prepositional senses play an important role: (a) phrasal verb paraphrasing, and (b) preposition selection.

5.1 Phrasal Verb Paraphrasing

Prepositions often act as a connection between verbs and complements, carrying nontrivial semantic information. We used the trained k-NN classifier on TPP senses to label the prepositions in the English Wikipedia corpus, and learned sense-specific embeddings. These sense-specific representations are readily interpretable in terms of the extensive-resources of TPP. Table 5 shows several senses of "in" together with their nearest neighbors in the vector space.

Embedding	Global	Simplex	Sense
Accuracy	0.44	0.44	**0.73**

Table 6: Accuracy on phrasal verbs paraphrasing.

To validate the sense-specific preposition representation, we infer the meaning of verb-particle construction (VPC), such as *climb down* with sense embeddings. This is a lexical paraphrasing task of finding one word that captures the meaning of VPC (e.g., *climb down* = *descend*).

Dataset. Because a dataset for paraphrasing of VPCs was not available, we created one (which is made available in the supplementary material). It consists of 91 phrasal verbs, extracted from the VPC datasets in (Baldwin, 2005), (McCarthy et al., 2003) and the online Oxford dictionary[2].

For each VPC instance, we first disambiguated the preposition sense in the given context using the supervised method described in Section 3. We consider a linear approximation of phrasal embeddings under three settings:

(1) *Sense-specific embedding*, approximating the representation of a VPC as the sum of the vectors of its verb and its preposition with a specific sense. Thus we have $v_{\text{vp}}^{\text{sense}} = v_{\text{verb}} + v_{\text{prep}}^{\text{sense}}$.

(2) *Global embedding baseline*: $v_{\text{vp}}^{\text{global}} = v_{\text{verb}} + v_{\text{prep}}^{\text{global}}$, where $v_{\text{prep}}^{\text{global}}$ is the global preposition embedding disregarding its sense.

(3) *Simplex embedding baseline* approximates the phrasal embedding to be just the verb embedding, i.e., $v_{\text{vp}}^{\text{simplex}} = v_{\text{verb}}$.

For each approximate phrasal embedding ($v_{\text{vp}}^{\text{sense}}$, $v_{\text{vp}}^{\text{global}}$, $v_{\text{vp}}^{\text{simplex}}$), we list the nearest three verbs (excluding the verb in the phrase) as its paraphrase, with the distance measured by the cosine similarity between the word vectors.

Two proficient English speakers set the gold standard for whether the paraphrase was valid or not (for polysemous verbs, we consider the verb to be a valid paraphrase if it conveys the meaning in any of its senses) and reconciled disagreements. We used accuracy as the evaluation metric, which is the percent of phrasal verbs with a valid paraphrase among candidates.

Results. We note that sense representations are able to capture the nuance of polysemous verb phrases. As is shown in Fig. 1, the phrase "fight for" has more than one meaning depending on the sense of its preposition. In the expression "fight for human rights", *for* carries the sense of purpose. Since the expression is semantically equivalent to "defend human rights", "defend" can paraphrase the phrase "fight for". In another context "fight for the prize" where *for* is related with benefits, "fight for" should be paraphrased as "win" and the expression "win the prize" is similar to "fight for the prize".

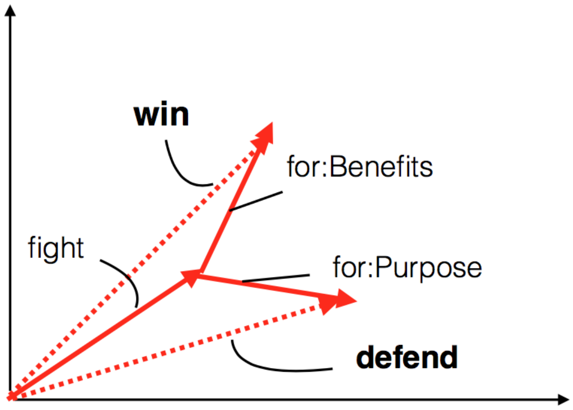

Figure 1: Paraphrasing polysemous verb phrases.

Some examples of phrasal verbs and paraphrases are shown in Table 7, with valid paraphrases highlighted. We report the performance of different embeddings in Table 6, where we notice that paraphrasing with the preposition sense embedding has a much higher accuracy than the two baselines. This validates the sense-specific preposition embedding suggesting its use in automatic paraphrasing of VPCs. Examples of paraphrases are shown in Table 7. A more detailed analysis of the results are in the appendix.

5.2 Preposition Selection

Given the polysemous and idiosyncratic nature of prepositions, choosing a preposition to fit a context can be a particularly challenging task for non-native English learners. Not surprisingly, preposition errors constitute the largest category of grammatical errors made by English learners (Chodorow et al., 2007). In this work, we show how the sense representations adequately capture the prepositional semantics, thus aiding preposition selection. Since TPP senses are fine-grained, we limit the senses available to be concrete or abstract by conflating the TPP senses to one of these

[2] https://en.oxforddictionaries.com

sentence	phrasal verb	paraphrasing		
		sense	global	simplex
The teaching is **carried on** in the form of folklore.	carried on	**conducted**	laid	placed
he **brought in** new ideas in the discussion.	brought in	**introduced**	came	came
She could not **keep from** crying.	keep from	**avoid**	get	maintain
Without a word he leaned forward and switched on the engine.	switched on	**starting**	shifted	reverted
I have certainly been **kicked in** the teeth by those bastards.	kicked in	**knocked**	throw	**knocked**
I have chosen to **block off** the easy track and so turn it into a dead end.	block off	**stopped**	cleared	cleared
The Rishon Le Zion killings **sparked off** a wave of sympathy protests.	sparked off	**ensued**	**spurred**	**ignited**
Stanley **put down** his paper and glared at her.	put down	**laid**	slammed	brought

Table 7: Paraphrasing of phrasal verbs.

two types (Reisinger and Mooney, 2010b). For example, "in a room" stands for the concrete sense of the preposition 'in', while "in her heart" corresponds to the abstract sense. In this part, we consider two senses for each preposition.

Dataset. We used three datasets, which consist of sentences marked with grammatical corrections out of which we only chose those with preposition errors. The Cambridge First Certificate in English (FCE) dataset contains $60,279$ prepositions with 4.8% error, the CoNLL dataset has $3,241$ prepositions with 4.7% errors and the Stack Exchange (SE) dataset has $15,814$ prepositions with 38.2% error (Prokofyev et al., 2014). Owing to its size, the FCE dataset was used for training, and the other two were used for testing. For each sentence with a preposition the task is to replace it with the correct one if it is used incorrectly.

Method. We classify all occurrences of each preposition sense into the two senses (abstract and concrete) by using the unsupervised PSD approach described in Section 4 to cluster the available senses. Then the prepositions in Wikipedia are labeled with one of the two senses, again using the unsupervised PSD approach. We then train sense embeddings on the newly labeled corpus with word2vec.

For a given sentence in the preposition selection task, we first disambiguate the sense of the preposition by checking which cluster it is closest to. The selection task is divided into preposition error detection and error correction. At the detection stage, we decide whether a preposition is used appropriately in the sentence. For this, we use as features the cosine similarity between the preposition sense embedding and the average word embedding in the context (the context size is 3), the rank of the preposition among all preposition choices with respect to the cosine similarity just mentioned, and the probability that the current preposition is replaced estimated from the training corpus. A decision tree classifier is used with these features to identify preposition errors.

At the second stage, we replace the current preposition p with another one if an error was detected at the first stage. Suppose that we consider replacing preposition p with q. We first disambiguate preposition q's sense given the context. We then use preposition q's sense embedding, the left context vector v_ℓ, the right context vector v_r, the interplay vector v_{inter} and the probability that q takes the place of p in the training corpus as input features to a two-layer MLP with 500 and 10 units in each layer. The MLP outputs a scalar to estimate how well the preposition q fits in p's context. The preposition with highest score is selected as the replacement.

Baseline. The state-of-the-art on preposition selection is one of the baselines, which makes use of lexical statistics from a large corpus as well as part-of-speech tags (Prokofyev et al., 2014). Also to evaluate the advantage of preposition sense representation over word representation, we have another baseline which uses the same classifier but the input features are the word embeddings instead of sense embeddings. The word embeddings were trained on Wikipedia English corpus with word2vec CBOW model.

Result. We compare the sense embedding-based approach against baselines in Table 8. As we can see, the use of sense representation achieves comparable performance to the state-of-the-art without using external linguistic tools. It also outperforms the baseline with word representation by a large margin.

6 Discussion

Resource-independence: Previous approaches to PSD relied on a part-of-speech tagger or depen-

Dataset	Method	Precision	Recall	F1 score
CoNLL	State-of-the-art	0.259	0.361	**0.302**
	Word representation	0.156	0.158	0.157
	Sense representation	0.279	0.283	0.281
SE	State-of-the-art	0.270	0.296	0.282
	Word representation	0.245	0.259	0.252
	Sense representation	0.281	0.297	**0.289**

Table 8: Performance in preposition selection.

dency parser to extract words modified by and modifying a preposition. In general, these words occur in the preposition's local context. We have allowed the context window to be a tunable parameter so that the classifier can learn to cover informative words in the context, and thus effectively captures the dependency information in a resource-independent fashion.

Context feature: The context averaging approach, which disregards context word order, suffers in accuracy compared to models that use left and right context words. This indicates that information about the word order relative to the preposition is useful in preposition disambiguation. Additionally, our use of the context-interplay feature combines the information on *both* sides of the preposition to infer its underlying sense.

Suppose that three expressions *a cup of medicine*, *professor of humanity* and *professor of mathematics* are in the training corpus, and the senses of the preposition *of* are 'contents', 'possessor' and 'field'. Given a test instance *professor of medicine*, it would be hard for the method with only the left or the right feature to decide the preposition sense since the test instance has the same word as each of the training instance, and their features in these two baselines are similar. However, the interplay vector in *professor of medicine* is closer to that in "professor of mathematics" than to other two training instances. The interplay feature prompts that *of* refers to a field (or species) instead of contents or possessor.

Data-driven insights into context dependence: Knowing the weights on the context features in the k-NN supervised PSD classifier, we can infer the extent to which prepositions rely on the complement and the attachment. For example, we found that in the case of the prepositions *behind* (occurring in, "shut behind her", "dip behind clouds"), *to* (e.g., "testify to the depth", "mumbling to himself"), and *with* (e.g., "amalgamated with her old school", and "rub with bare hands"), the verbs they attach to strongly influence their sense. For other prepositions such as *during* (e.g., "during the incident") and *on* (e.g., "on his hands"), the complement noun has more influence on the senses than governors.

Sense helps paraphrasing: We observe that sense-specific preposition representation helps improve phrasal verb paraphrasing greatly. Working with the VPC dataset and the simplistic model of compositionality, we interpret the results as positive indicators of the viability of using sense-specific prepositional embeddings to paraphrase verb-particle constructions. In the case of *light verbs*, whose meaning is determined largely by the particles they combine with, (e.g., *come down* $\sim$ *fall*), a valid paraphrase is found in the top 3 candidates when the sense-specific representation is used, and not when the simplex or the global representation is used.

7 Conclusion

This paper studies the preposition sense disambiguation by encoding the attachment and complement properties into context features. The disambiguation method performs well on three standard PSD tasks and readily scales to a large corpus. The resulting sense-specific representations are shown to capture semantics of preposition senses in our quantitative analysis. They are also shown to aid two downstream tasks: phrasal verb paraphrasing and preposition selection.

Acknowledgments

This work is supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network. We thank EMNLP anonymous reviewers for their constructive suggestions.

References

Alexandr Andoni and Piotr Indyk. 2006. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468. IEEE.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Timothy Baldwin. 2005. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414.

Joan C Beal. 2004. *Grammars and grammarians*, volume English in modern times 1700-1945, 89-123. London, Great Britain: Arnold.

Yonatan Belinkov, Tao Lei, Regina Barzilay, and Amir Globerson. 2014. Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics*, 2:561–572.

Martin Chodorow, Joel R Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the fourth ACL-SIGSEM workshop on prepositions*, pages 25–30. Association for Computational Linguistics.

William Cobbett. 1823. *A grammar of the English language*. London: B. Bensley.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jeanette S DeCarrico. 2000. *The structure of English: Studies in form and function for language teaching*, volume 1. University of Michigan Press/ESL.

Nicole Dehé. 2002. *Particle verbs in English: Syntax, information structure and intonation*, volume 59. John Benjamins Publishing.

Georgiana Dinu, Stefan Thater, and Sören Laue. 2012. A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 611–615, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph E Emonds. 1985. A unified theory of syntactic categories. *Studies in generative grammar*, (19):1–356.

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

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256.

Hila Gonen and Yoav Goldberg. 2016. Semi supervised preposition-sense disambiguation using multilingual data. *arXiv preprint arXiv:1611.08813*.

Hongyu Gong, Suma Bhat, and Pramod Viswanath. 2018. Embedding syntax and semantics of prepositions via tensor decomposition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 896–906.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2015. Learning embeddings for transitive verb disambiguation by implicit tensor factorization. *ACL-IJCNLP 2015*, page 1.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition?: dimensions of sense disambiguation for an interesting word class. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 454–462. Association for Computational Linguistics.

Dirk Hovy, Ashish Vaswani, Stephen Tratz, David Chiang, and Eduard Hovy. 2011. Models and training for unsupervised preposition sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 323–328. Association for Computational Linguistics.

Rodney Huddleston. 1984. *Introduction to the Grammar of English*. Cambridge University Press.

Ray Jackendoff. 2002. English particle constructions, the lexicon, and the autonomy of syntax. *Verb-particle explorations*, pages 67–94.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 941951.

David J Ketchen Jr and Christopher L Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, pages 441–458.

Ken Litkowski. 2013. Preposition disambiguation: Still a problem. *CL Research, Damascus, MD*, pages 1–8.

Ken Litkowski and Orin Hargraves. 2007. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 24–29. Association for Computational Linguistics.

Kenneth C Litkowski and Orin Hargraves. 2005. The preposition project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179.

Robert Lowth. 1762. *A short introduction to English grammar*. London: A. Miller & R. & J. Dodsby.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 73–80. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, pages 236–244.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. Geometry of polysemy. *Proceedings of ICLR*.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. *Conference on Empirical Methods in Natural Language Processing*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Octavian Popescu, Sara Tonelli, and Emanuele Pianta. 2007. Irst-bp: Preposition disambiguation based on chain clarifying relationships contexts. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 191–194. Association for Computational Linguistics.

Roman Prokofyev, Ruslan Mavlyutov, Martin Grund, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Correct me if i'm wrong: Fixing grammatical errors by preposition ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 331–340. ACM.

Joseph Reisinger and Raymond J. Mooney. 2010a. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 109–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph Reisinger and Raymond J Mooney. 2010b. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Samuel Ritter, Cotie Long, Denis Paperno, Marco Baroni, Matthew Botvinick, and Adele Goldberg. 2015. Leveraging preposition ambiguity to assess compositional distributional models of semantics. *Lexical and Computational Semantics (* SEM 2015)*, page 199.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Patrick Saint-Dizier. 2006. *Syntax and semantics of prepositions*, volume 29. Springer Science & Business Media.

Linfeng Song, Zhiguo Wang, Haitao Mi, and Daniel Gildea. 2016. Sense embedding learning for word sense induction. *arXiv preprint arXiv:1606.05409*.

Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Catherine A Sugar and Gareth M James. 2003. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, 98(463):750–763.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *IJCNLP*.

Stephen Tratz and Dirk Hovy. 2009. Disambiguation of preposition sense using linguistically motivated features. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 96–100. Association for Computational Linguistics.

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1257–1268. Association for Computational Linguistics.

Aline Villavicencio. 2006. Verb-particle constructions in the world wide web. In *Syntax and Semantics of Prepositions*, pages 115–130. Springer.

Lingfei Wu, Ian En-Hsu Yen, Fangli Xu, Pradeep Ravikuma, and Michael Witbrock. 2018a. D2ke: From distance to kernel and embedding. *arXiv preprint arXiv:1802.04956*.

Lingfei Wu, Ian En-Hsu Yen, Jinfeng Yi, Fangli Xu, Qi Lei, and Michael Witbrock. 2018b. Random warping series: A random features method for time-series embedding. In *International Conference on Artificial Intelligence and Statistics*, pages 793–802.

Kun Xu, Lingfei Wu, Zhiguo Wang, and Vadim Sheinin. 2018. Graph2seq: Graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv:1804.00823*.

Patrick Ye and Timothy Baldwin. 2007. Melb-yb: Preposition sense disambiguation using rich semantic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 241–244. Association for Computational Linguistics.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Deniz Yuret. 2007. Ku: Word sense disambiguation by substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 207–213. Association for Computational Linguistics.