Lexical Inference over Multi-Word Predicates: A Distributional Approach

Omri Abend, Shay B. Cohen and Mark Steedman
School of Informatics, University of Edinburgh,
Edinburgh EH8 9AB, United Kingdom
{oabend,scohen,steedman}@inf.ed.ac.uk

Abstract

Representing predicates in terms of their argument distribution is common practice in NLP. Multi-word predicates (MWPs) in this context are often either disregarded or considered as fixed expressions. The latter treatment is unsatisfactory in two ways: (1) identifying MWPs is notoriously difficult, (2) MWPs show varying degrees of compositionality and could benefit from taking into account the identity of their component parts. We propose a novel approach that integrates the distributional representation of multiple sub-sets of the MWP's words. We assume a latent distribution over sub-sets of the MWP, and estimate it relative to a downstream prediction task. Focusing on the supervised identification of lexical inference relations, we compare against state-of-the-art baselines that consider a single sub-set of an MWP, obtaining substantial improvements. To our knowledge, this is the first work to address lexical relations between MWPs of varying degrees of compositionality within distributional semantics.

1 Introduction

Multi-word expressions (MWEs) constitute a large part of the lexicon and account for much of its growth (Jackendoff, 2002; Seaton and Macaulay, 2002). However, despite their importance, MWEs remain difficult to define and model, and consequently pose serious difficulties for NLP applications (Sag et al., 2001). Multi-word Predicates (MWPs; sometimes termed Complex Predicates) form an important and much addressed subclass of MWEs and are the focus of this paper.

MWPs are informally defined as multiple words that constitute a single predicate (Alsina et al., 1997). MWPs encompass a wide range of phenomena, including causatives, light verbs, phrasal verbs, serial verb constructions and many others, and pose considerable challenges to both linguistic theory and NLP applications (see Section 2). Part of the difficulty in treating them stems from their position on the borderline between syntax and the lexicon. It is therefore often unclear whether they should be treated as fixed expressions, as compositional phrases that reflect the properties of their component parts or as both.

This work addresses the modelling of MWPs within the context of distributional semantics (Turney and Pantel, 2010), in which predicates are represented through the distribution of arguments they may take. In order to collect meaningful statistics, the predicate's lexical unit should be sufficiently frequent and semantically unambiguous.

MWPs pose a challenge to such models, as naïvely collecting statistics over all instances of highly ambiguous verbs is likely to result in noisy representations. For instance, the verb "take" may appear in MWPs as varied as "take time", "take effect" and "take to the hills". This heterogeneity of "take" is likely to have a negative effect on downstream systems that use its distributional representation. For instance, while "take" and "accept" are often considered lexically similar, the high frequency in which "take" participates in non-compositional MWPs is likely to push the two verbs' distributional representations apart.

A straightforward approach to this problem is to represent the predicate as a conjunction of multiple words, thereby trading ambiguity for sparsity. For instance, the verb "take" could be conjoined with its object (e.g., "take care", "take a bus"). This approach, however, raises the challenge of identifying the sub-set of the predicate's words that should be taken to represent it (henceforth, its *lexical components* or LCs).

We propose a novel approach that addresses this

challenge in the context of identifying lexical inference relations between predicates (Lin and Pantel, 2001; Schoenmackers et al., 2010; Melamud et al., 2013a, *inter alia*). A (lexical) inference relation $p_L \rightarrow p_R$ is said to hold if the relation denoted by p_R generally holds between a set of arguments whenever the relation p_L does. For instance, an inference relation holds between "annex" and "control" since if a country annexes another, it generally controls it. Most works to this task use distributional similarity, either as their main component (Szpektor and Dagan, 2008; Melamud et al., 2013b), or as part of a more comprehensive system (Berant et al., 2011; Lewis and Steedman, 2013).

For example, consider the verb "take". While the inference relation "*have* → *take*" does not generally hold, it does hold in the case of some light verbs, such as "*have a look* → *take a look*", underscoring the importance of taking more inclusive LCs into account. On the other hand, the predicate "likely to give a green light" is unlikely to appear often even within a very large corpus, and could benefit from taking its lexical sub-units (e.g., "likely" or "give a green light") into account.

We present a novel approach to the task that models the selection and relative weighting of the predicate's LCs using latent variables. This approach allows the classifier that uses the distributional representations to take into account the most relevant LCs in order to make the prediction. By doing so, we avoid the notoriously difficult problem of defining and identifying MWPs and account for predicates of various sizes and degrees of compositionality. To our knowledge, this is the first work to address lexical relations between MWPs of varying degrees of compositionality within distributional semantics.

We conduct experiments on the dataset of Zeichner et al. (2012) and compare our methods with analogous ones that select a fixed LC, using state-of-the-art feature sets. Our method obtains substantial performance gains across all scenarios.

Finally, we note that our approach is cognitively appealing. Significant cognitive findings support the claim that a speaker's lexicon consists of partially overlapping lexical units of various sizes, of which several can be evoked in the interpretation of an utterance (Jackendoff, 2002; Wray, 2008).

2 Background and Related Work

Inference Relations. The detection of inference relations between predicates has become a central task over the past few years (Sekine, 2005; Zanzotto et al., 2006; Schoenmackers et al., 2010; Berant et al., 2011; Melamud et al., 2013a, *inter alia*). Inference rules are used in a wide variety of applications including Question Answering (Ravichandran and Hovy, 2002), Information Extraction (Shinyama and Sekine, 2006), and as a main component in Textual Entailment systems (Dinu and Wang, 2009; Dagan et al., 2013).

Most approaches to the task used distributional similarity as a major component within their system. Lin and Pantel (2001) introduced DIRT, an unsupervised distributional system for detecting inference relations. The system is still considered a state-of-the-art baseline (Melamud et al., 2013a), and is often used as a component within larger systems. Schoenmackers et al. (2010) presented an unsupervised system for learning inference rules directly from open-domain web data. Melamud et al. (2013a) used topic models to combine type-level predicate inference rules with token-level information from their arguments in a specific context. Melamud et al. (2013b) used lexical expansion to improve the representation of infrequent predicates. Lewis and Steedman (2013) combined distributional and symbolic representations, evaluating on a Question Answering task, as well as on a quantification-focused entailment dataset.

Several studies tackled the task using supervised systems. Weisman et al. (2012) used a set of linguistically motivated features, but evaluated their system on a corpus that consists almost entirely of single-word predicates. Mirkin et al. (2006) presented a system for learning inference rules between nouns, using distributional similarity and pattern-based features. Hagiwara et al. (2009) identified synonyms using a supervised approach relying on distributional and syntactic features. Berant et al. (2011) used distributional similarity between predicates to weight the edges of an entailment graph. By imposing global constraints on the structure of the graph, they obtained a more accurate set of inference rules.

Previous work used simple methods to select the predicate's LC. Some filtered out frequent highly ambiguous verbs (Lewis and Steedman, 2013), others selected a single representative word (Melamud et al., 2013a), while yet others used multi-word LCs but treated them as fixed expressions (Lin and Pantel, 2001; Berant et al., 2011).

The goals of the above studies are largely com-

plementary to ours. While previous work focused either on improving the quality of the distributional representations themselves or on their incorporation into more elaborate systems, we focus on the integration of the distributional representation of multiple LCs to improve the identification of inference relations between MWPs.

MWP Extraction and Identification. MWPs have received considerable attention over the years in both theoretical and applicative contexts. Their position on the crossroads of syntax and the lexicon, their varying degrees of compositionality, as well as the wealth of linguistic phenomena they exhibit, made them the object of ongoing linguistic discussion (Alsina et al., 1997; Butt, 2010).

In NLP, the discovery and identification of MWEs in general and MWPs in particular has been the focus of much work over the years (Lin, 1999; Baldwin et al., 2003; Biemann and Giesbrecht, 2011). Despite wide interest, the field has yet to converge to a general and widely agreed-upon method for identifying MWPs. See (Ramisch et al., 2013) for an overview.

Most work on MWEs emphasized idiosyncratic or non-compositional expressions. Other lines of work focused on specific MWP classes such as light verbs (Tu and Roth, 2011; Vincze et al., 2013) and phrasal verbs (McCarthy et al., 2003; Pichotta and DeNero, 2013). Our work proposes a uniform treatment to MWPs of varying degrees of compositionality, and avoids defining MWPs explicitly by modelling their LCs as latent variables.

Compositional Distributional Semantics. Much work in recent years has concentrated on the relation between the distributional representations of composite phrases and the representations of their component sub-parts (Widdows, 2008; Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Coecke et al., 2010). Several works have used compositional distributional semantics (CDS) representations to assess the compositionality of MWEs, such as noun compounds (Reddy et al., 2011) or verb-noun combinations (Kiela and Clark, 2013). Despite significant advances, previous work has mostly been concerned with highly compositional cases and does not address the distributional representation of predicates of varying degrees of compositionality.

3 Our Proposal: A Latent LC Approach

This section details our approach for distributionally representing MWPs by leveraging their component LCs. Section 3.1 describes our general approach, Section 3.2 presents our model and Section 3.3 details the feature set.

3.1 General Approach and Notation

We propose a method for addressing MWPs of varying degrees of compositionality through the integration of the distributional representation of multiple sub-sets of the predicate's words (LCs). We use it to tackle a supervised prediction task that represents predicates distributionally. Our model assumes a latent distribution over the LCs, and estimates its parameters so to best conform to the goals of the target prediction task.

Formally, given a predicate p, we denote the set of words comprising it as $W(p)$. The set of allowable LCs for p is denoted with $H_p \subset 2^{W(p)}$. H_p contains all sub-sets of p that we consider as apriori possible to represent p. For instance, if p is "likely to give a green light", H_p may include LCs such as "likely" or "give light". As our method is aimed at discovering the most relevant LCs, we do not attempt to analyze the MWPs in advance, but rather take an inclusive H_p, allowing the model to estimate the relative weights of the LCs.

The task we use as a testbed for our approach is the lexical inference identification task between predicates. Given a pair of predicates $p = (p_L, p_R)$, the task is to predict whether an inference relation holds between them. For instance, if p_L is "devour" and p_R is "eat greedily", the classifier should use the similarity between "devour" and "eat" in order to correctly predict an inference relation in this case. Selecting the wider LC "eat greedily" might result in sparser statistics. In other examples, however, taking a wider LC is potentially beneficial. For instance, the dissimilarity between "take" and "make" should not prevent the classifier from identifying the inference relation between "take a step" and "make a step".

Our statistical model aims at predicting the correct label by making use of partially overlapping LCs of various sizes, both for the premise left-hand side (LHS) predicate p_L and the hypothesis right-hand side (RHS) predicate p_R. More formally, we take the space of values for our latent LC variables to be $H_{p_L,p_R} = H_{p_L} \times H_{p_R}$.

Our evaluation dataset consists of pairs $p^{(i)} = (p_L^{(i)}, p_R^{(i)})$ for $i \in \{1, \ldots, M\}$, where M is the number of examples available, coupled with their gold-standard labels $y^{(i)} \in \{1, -1\}$. For brevity, we denote $H^{(i)} = H_{p^{(i)}} = H_{p_L^{(i)}, p_R^{(i)}}$. We also as-

sume the existence of a feature function $\Phi(p, y, h)$ which maps a triplet of a predicate pair p, an inference label y, and a latent state $h \in H_p$ to $\mathbb{R}^d$ for some integer d. We denote the training set by $\mathcal{D}$.

3.2 The Model

We address the task with a latent variable log-linear model, representing the LCs of the predicates. We choose this model for its generality, conceptual simplicity, and because it allows to easily incorporate various feature sets and sets of latent variables. We introduce L_2 regularization to avoid over-fitting. We use maximum likelihood estimation, and arrive at the following objective function:

$$L(w|\mathcal{D}) = \frac{1}{M}\sum_{i=1}^{M} \log P(y^{(i)}|p^{(i)}, w) - \frac{\lambda}{2}\|w\|^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\log \sum_{h \in H^{(i)}} \exp\left(w^\top \Phi(p^{(i)}, y^{(i)}, h)\right) - \log Z(w, i)\right) - \frac{\lambda}{2}\|w\|^2$$

where:

$$Z(w, i) = \sum_{y \in \{-1,1\}} \sum_{h \in H^i} \exp(w^\top \Phi(p_i, y, h)).$$

We maximize L using the BFGS algorithm (Nocedal and Wright, 1999). The gradient (with respect to w) is the following:

$$\nabla L = \mathbb{E}_h[\Phi(p_i, y_i, h)] - \mathbb{E}_{h,y}[\Phi(p_i, y, h)] - \lambda \cdot w$$

H_p can be defined to be any sub-set of $2^{W(p)}$ given that taking an expectation over H can be done efficiently. It is therefore possible to use prior linguistic knowledge to consider only sub-sets of p that are likely to be non-compositional (e.g., verb-preposition or verb-noun pairs).

In our experiments we attempt to keep the approach maximally general, and define H_p to be the set of all subsets of size 1 or 2 of content words in W_p[1]. We bound the size of $h \in H_p$ in order to retain computational efficiency and a sufficient frequency of the LCs in H_p. MWPs of length greater than 2 are effectively approximated by their set of subsets of sizes 1 and 2.

Each h can therefore be written as a 4-tuple $(h_L^A, h_L^B, h_R^A, h_R^B)$, where h_L^A (h_R^A) denotes the first word of the LHS (RHS) predicate's LC. h_L^B (h_R^B) denotes the (possibly empty) second word of the predicate. Inference is carried out by maximizing $P(y|p^{(i)})$ over y. As $|H_p| = O(k^4)$, where k is the number of content words in p, and as the number of content words is usually small[2], inference can be carried out by directly summing over $H^{(i)}$.

Initialization. The introduction of latent variables into the log-linear model leads to a non-convex objective function. Consequently, BFGS is not guaranteed to converge to the global optimum, but rather to a stationary point. The result may therefore depend on the parameter initialization. Indeed, preliminary experiments showed that both initializing w to be zero and using a random initializer results in lower performance.

Instead, we initialize our model with a simplified convex model that fixes the LCs to be the pair of left-most content words comprising each of the predicates. This is a common method for selecting the predicate's LC (e.g., Melamud et al., 2013a). Once h has been fixed, the model collapses to a convex log-linear model. The optimal w is then taken as an initialization point for the latent variable model. While this method may still not converge to the global maximum, our experiments show that this initialization technique yields high quality values for w (see Section 6).

3.3 Feature Set

This section lists the features used for our experiments. We intentionally select a feature set that relies on either completely unsupervised or shallow processing tools that are available for a wide variety of languages and domains.

Given a predicate pair $p^{(i)}$, a label $y \in \{1, -1\}$ and a latent state $h \in H^{(i)}$, we define their feature vector as $\Phi(p^{(i)}, y, h) = y \cdot \Phi(p^{(i)}, h)$. The computation of $\Phi(p^{(i)}, h)$ requires a reference corpus $\mathcal{R}$ that contains triplets of the type (p, x, y) where p is a binary predicate and x and y are its arguments. We use the Reverb corpus as $\mathcal{R}$ in our experiments (Fader et al., 2011; see Section 4). We refrain from encoding features that directly reflect the vocabulary of the training set. Such features are not applicable beyond that set's vocabulary, and as available datasets contain no more than a few thousand examples, these features are unlikely to generalize well.

Table 1 presents the set of features we use in our experiments. The features can be divided into two main categories: similarity features between the LHS and the RHS predicates (table's top), and features that reflect the individual properties of each

[1] We use a POS tagger to identify content words. Prepositions are considered content words under this definition.

[2] $|H_p|$ is about 15 on average in our dataset, where less than 5% of the $H^{(i)}$ are of size greater than 50.

Category		Name	Description
	Similarity	COSINE	DIRT cosine similarity between the vectors of h_L and h_R
		$\mathrm{COSINE_A}$	DIRT cosine similarity between the vectors of h_L^A and h_R^A
		BInc	DIRT BInc similarity between the vectors of h_L and h_R
		$\mathrm{BInc_A}$	DIRT BInc similarity between the vectors of h_L^A and h_R^A
	Word A LHS	$\mathrm{POS_L^A}$	The most frequent POS tag for the lemma of h_L^A
		$\mathrm{POS2_L^A}$	The second most frequent POS tag for the word lemma of h_L^A
		$\mathrm{FREQ_L^A}$	The number of occurrences of h_L^A in the reference corpus
		$\mathrm{COMMON_L^A}$	A binary feature indicating whether h_L^A appears in both predicates
		$\mathrm{ORDINAL_L^A}$	The ordinal number of h_L^A among the content words of the LHS predicate
	Pair LHS	$\mathrm{POS_L^{AB}}$	The conjunction of POS_L^A and POS_L^B
		$\mathrm{FREQ_L^{AB}}$	The frequency of h_L^A and h_L^B in the reference corpus
		$\mathrm{PREFAB_L}$	$P(h_L^A \mid h_L^A)$ as estimated from the reference corpus
		$\mathrm{PREFBA_L}$	$P(h_L^B \mid h_L^A)$ as estimated from the reference corpus
		$\mathrm{PMIAB_L}$	The point-wise mutual information of h_L^A and h_L^B
	LDA	$\mathrm{TOPICS_L}$	$P(\mathrm{topic} \mid h_L)$ for each of the induced topics.
		$\mathrm{TOPICENT_L}$	The entropy of the topic distribution $P(\mathrm{topic} \mid h_L)$

Table 1: Our feature set. Features are listed for the LHS predicate (h_L), and for the first word in it (h_L^A). Analogous features are introduced for h_L^B, and for the RHS predicate.

of them. Within the LHS feature set, we distinguish between two sub-types of features: word features that encode the individual properties of h_L^A and h_L^B (table's upper middle part), and pair features that only apply to LCs of size 2 and reflect the relation between h_L^A and h_L^B (table's lower middle part). We further incorporate LDA-based features that reflect the selectional preferences of the predicates (table's bottom).

Distributional Similarity Features. The distributional similarity features are based on the DIRT system (Lin and Pantel, 2001). The score defines for each predicate p and for each argument slot $s \in \{L, R\}$ (corresponding to the arguments to the right and left of that predicate) a vector v_s^p which represents the distribution of arguments appearing in that slot. We take $v_s^p(x)$ to be the number of times that the argument x appeared in the slot s of the predicate p. Given these vectors, the similarity between the predicates p_1 and p_2 is defined as:

$$\mathrm{score}(p_1, p_2) = \sqrt{\mathrm{sim}(v_L^{p_1}, v_L^{p_2}) \cdot \mathrm{sim}(v_R^{p_1}, v_R^{p_2})}$$

where sim is some vector similarity measure.

We use two common similarity measures: the vector cosine metric, and the BInc (Szpektor and Dagan, 2008) similarity measure. These measures give complementary perspectives on the similarity between the predicates, as the cosine similarity is symmetric between the LHS and RHS predicates, while BInc takes into account the directionality of the inference relation. Preliminary experiments with other measures, such as those of Lin (1998) and Weeds and Weir (2003) did not yield additional improvements.

We encode the similarity of all measures for the pair h_L and h_R as well as the pair h_L^A and h_R^A. The latter feature is an approximation to the similarity between the heads of the predicates, as heads in English tend to be to the left of the predicates. These two features coincide for h values of size 1.

Word and Pair Features. These features encode the basic properties of the LC. The motivation behind them is to allow a more accurate leveraging of the similarity features, as well as to better determine the relative weights of $h \in H^{(i)}$.

The feature set is composed of four analogous sets corresponding to h_L^A,h_L^B,h_R^A and h_R^B, as well as two sets of features that capture relations between h_L^A, h_L^B and h_R^A, h_R^B (in cases h is of size 2). The features include the ordinal index of the word within the predicate, the lemma's frequency according to $\mathcal{R}$, and a feature that indicates whether that word's lemma also appears in both predicates of the pair. For instance, when considering the predicates "likely to come" and "likely to leave", "likely" appears in both predicates, while "come" and "leave" appear only in one of them.

In addition, we use POS-based features that encode the most frequent POS tag for the word lemma and the second most frequent POS tag (according to $\mathcal{R}$). Information about the second most frequent POS tag can be important in identifying light verb constructions, such as "take a swim" or "give a smile", where the object is derived from a verb. It can thus be interpreted as a generalization of the feature that indicates whether the object is a deverbal noun, which is used by some light verb identification algorithms (Tu and Roth, 2011).

In cases where h_L is of size 2, we additionally encode features that apply to the conjunction of h_L^A and h_L^B. We encode the conjunction of their POS and the number of times the two lemmas occurred together in $\mathcal{R}$. We also introduce features that capture the statistical correlation between the words of h_L. To do so, we use point-wise mutual information, and the conditional probabilities $P(h_L^A|h_L^B)$ and $P(h_L^B|h_L^A)$. Similar measures have often been used for the unsupervised detection of MWEs (Villavicencio et al., 2007; Fazly and Stevenson, 2006). We also include the analogous set of features for h_R.

LDA-based Features. We further incorporate features based on a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003). Several recent works have underscored the usefulness of using topic models to model a predicate's selectional preferences (Ritter et al., 2010; Dinu and Lapata, 2010; Séaghdha, 2010; Lewis and Steedman, 2013; Melamud et al., 2013a). We adopt the approach of Lewis and Steedman (2013), and define a pseudo-document for each LC in the evaluation corpus. We populate the pseudo-documents of an LC with its arguments according to $\mathcal{R}$. We then train an LDA model with 25 topics over these documents. This yields a probability distribution $P(\text{topic}|h)$ for each LC h, reflecting the types of arguments h may take.

We further include a feature for the entropy of the topic distribution of the predicate, which reflects its heterogeneity. This feature is motivated by the assumption that a heterogeneous predicate is more likely to benefit from selecting a more inclusive LC than a homogeneous one.

Technical Issues. All features used, except the similarity ones and the topic distribution features are binary. Frequency features are binned into 4 bins of equal frequency. We conjoin some of the feature sets by multiplying their values. Specifically, we add the cross product of the features of the category "Similarity" (see Table 1) with the rest of the features. In addition, we conjoin all LHS (RHS) features with an indicator feature that indicates whether h_L (h_R) is of size two. This results in 1605 non-constant features.

We further note that some LCs that appear in the evaluation corpus do not appear at all in $\mathcal{R}$. In our experiments they amounted to 0.2% of the LCs in our evaluation dataset. While previous work often discarded predicates below a certain frequency from the evaluation, we include them in order to facilitate comparison to future work. We assign the similarity features of such examples a 0 value, and assign their other numerical features the mean value of those features.

4 Experimental Setup

Corpora and Preprocessing. As a reference corpus $\mathcal{R}$, we use Reverb (Fader et al., 2011), a web-based corpus consisting of 15M web extractions of binary relations. Each relation is a triplet of a predicate and two arguments, one preceding it and one following it. Relations were extracted using regular expressions over the output of a POS tagger and an NP chunker. Each predicate may consist of a single verb, a verb and a preposition or a sequence of words starting in a verb and ending in a preposition, between which there may nouns, adjectives, adverbs, pronouns, determiners and verbs. The verb may also be a copula. Examples of predicates are "make the most of", "could be exchanged for" and "is happy with".

Reverb is an appealing reference corpus for this task for several reasons. First, it uses fairly shallow preprocessing technology which is available for many domains and languages. Second, Reverb applies considerable noise filtering, which results in extractions of fair quality. Third, our evaluation dataset is based on Reverb extractions.

We evaluate our algorithm on the dataset of Zeichner et al. (2012). This publicly available corpus[3] provides pairs of Reverb binary relations and an indication of whether an inference relation holds between them within the context of a specific pair of argument fillers. The corpus was compiled using distributional methods to detect pairs of relations in Reverb that are likely to have an inference relation between. Annotators, employed through Amazon Mechanical Turk, were then asked to determine whether each pair is meaningful, and if so, to determine whether an inference relation holds. Further measures were taken to monitor the accuracy of the annotation.

For example, the pair of predicates "make the most of" and "take advantage of" appears in the corpus as a pair between which an inference relation holds. The arguments in this case are "students" and "their university experience". An example of a pair between which an inference relation does not hold is "tend to neglect" and "underestimate the importance of", where the arguments are "Robert" and "his family".

[3] http://tinyurl.com/krx2acd

The dataset contains 6,565 instances in total. We use 5,411 pairs of them, discarding instances that were deemed as meaningless by the annotators. We also discard cases where the set of arguments is reversed between the LHS and RHS predicates. In these examples, $p_R(x, y)$ is inferable from $p_L(y, x)$, rather than from $p_L(x, y)$. As there are less than 150 reversed instances in the corpus, experimenting on this sub-set is unlikely to be informative.

The average length of a predicate in the corpus is 2.7 words (including function words). In 87.3% of the predicate pairs, there was more than one LC (i.e., $|H_p| > 1$), underscoring the importance of correctly leveraging the different LCs. We randomly partition the corpus into a training set which contains 4,343 instances ($\sim$80%), and a test set that contains 1,068 instances, maintaining the same positive to negative label ratio in both datasets[4]. Development was carried out using cross-validation on the training data (see below).

We use a Maximum Entropy POS Tagger, trained on the Penn Treebank, and the WordNet lemmatizer, both implemented within the NLTK package (Loper and Bird, 2002). To obtain a coarse-grained set of POS tags, we collapse the tag set to 7 categories: nouns, verbs, adjectives, adverbs, prepositions, the word "to" and a category that includes all other words. A Reverb argument is represented as the conjunction of its content words that appear more than 10 times in the corpus. Function words are defined according to their POS tags and include determiners, possessive pronouns, existential "there", numbers and coordinating conjunctions. Auxiliary verbs and copulas are also considered function words.

To compute the LDA features, we use the online variational Bayes algorithm of Hoffman et al. (2010) as implemented in the *Gensim* software package (Rehurek and Sojka, 2010).

Evaluated Algorithms. The only two previous works on this dataset (Melamud et al., 2013a; Melamud et al., 2013b) are not directly comparable, as they used unsupervised systems and evaluated on sub-sets of the evaluation dataset. Instead, we use several baselines to demonstrate the usefulness of integrating multiple LCs, as well as the relative usefulness of our feature sets.

[4]A script that replicates our train-test partition of the corpus can be found here: `http://homepages.inf.ed.ac.uk/oabend/mwpreds.html`

The simplest baseline is ALLNEG, which predicts the most frequent label in the dataset (in our case: "no inference"). The other evaluated systems are formed by taking various subsets of our feature set. We experiment with 4 feature sets. The smallest set, SIM, includes only the similarity features. This feature set is related to the compositional distributional model of Mitchell and Lapata (2010) (see Section 6). We note that despite recent advances in identifying predicate inference relations, the DIRT system (Lin and Pantel, 2001) remains a strong baseline, and is often used as a component in state-of-the-art systems (Berant et al., 2011), and specifically in the two aforementioned works that used the same evaluation corpus.

The next feature set BASIC includes the features found to be most useful during the development of the model: the most frequent POS tag, the frequency features and the feature Common. More inclusive is the feature set NO-LDA, which includes all features except the LDA features. Experiments with this set were performed in order to isolate the effect of the LDA features. Finally, ALL includes our complete set of features.

The more direct comparison is against partial implementations of our system where the LC h is deterministically selected. Determining h for each predicate yields a regular log-linear binary classification model. We use two variants of this baseline. The first, LEFTMOST, selects the left-most content word for each predicate. Similar selection strategy was carried out by Melamud et al. (2013a). The second, VPREP, selects h to be the verb along with its following preposition. In cases the predicate contains multiple verbs, the one preceding the preposition is selected, and where the predicate does not contain any non-copula verbs, it regresses to LEFTMOST. This LC selection method approximates a baseline that includes subcategorized prepositions. Such cases are highly frequent and account for a large portion of the MWPs in English. Including a verb's preposition in its LC was commonly done in previous work (e.g., Lewis and Steedman, 2013).

We also attempted to identify verb-preposition constructions using a dependency parser. Unfortunately, our evaluation dataset is only available in a lemmatized version, which posed a difficulty for the parser. Due to the low quality of the resulting parses, we implemented VPREP using POS-based regular expressions as defined above.

The full model is denoted with LATENTLC. For each system and feature set, we report results using 10-fold cross-validation on the training set, as well as results on the test set. Both cases use the same set of parameters determined by cross-validation on the training set. As the task at hand is a binary classification problem, we use accuracy scores to rate the performance of our systems.

5 Results

Table 2 presents the results of our experiments. Rows correspond to the evaluated algorithms, while columns correspond to the feature sets used and the evaluation scenarios (i.e., training set cross-validation or test set evaluation). Our experiments make first use of this dataset in its fullest form for the problem of supervised learning of inference relations, and may serve as a starting point for further exploration of this dataset.

For all feature sets and settings, LATENTLC scored highest, often with a considerable margin of up to 3.0% in the cross-validation and up to 4.6% on the test set relative to the LEFTMOST baseline, and 5.1% (cross-validation) and 6.8% (test) margins relative to VPREP.

The best scoring result of our LATENTLC model in the cross-validation scenario is 65.72%, obtained by the feature set All. The best scoring result by any of the baseline models in this scenario is 62.7%, obtained by the same feature set. For the test set scenario, LATENTLC obtained its highest accuracy, 65.73%, when using the feature set Basic. This is a substantial improvement over the highest scoring baseline model in this scenario that obtained 61.6% accuracy, using the feature set All. This performance gap is substantial when taking into consideration that the improvements obtained by the highly competitive DIRT similarity features using the stronger LEFTMOST baseline, result in an improvement of 3.1% and 5.3% over the trivial ALLNEG baseline in the test set and cross-validation scenarios respectively.

Comparing the different feature sets on our proposed model, we find that the Basic feature set gives a consistent and substantial increase over the Sim feature set. Improvements are of 2.8% (test) and 2.2% (cross-validation). Introducing more elaborate features (i.e., the feature sets NoLDA and All) yields some improvements in the cross-validation, but these improvements are not replicated on the test set. This may be due to idiosyncrasies in the test set that are averaged out in the cross-validation scenario.

For a qualitative analysis, we took the best performing model of the data set (i.e., with the Basic feature set), and extracted the set of instances where it made a correct prediction while both baselines made an error. This set contains many verb-preposition pairs, such as "list as → report as" or "submit via → deliver by", underscoring the utility of leveraging multiple LCs rather than considering only a head word (as with LEFTMOST) or the entire phrase (as with VPREP). Other examples in this set contain more complex patterns. These include the positive pairs "talk much about → have much to say about" and "increase with → go up with", and the negative "make prediction about → meet the challenge of" and "enjoy watching → love to play".

6 Discussion

Relation to CDS. Much recent work subsumed under the title *Compositional Distributional Semantics* addressed the distributional representation of multi-word phrases (see Section 2). This line of work focuses on compositional predicates, such as "kick the ball" and not on idiosyncratic predicates such as "kick the bucket".

A variant of the CDS approach can be framed within ours. Assume we wish to compute the similarity of the predicates $p_L = (w_1, ..., w_n)$ and $p_R = (w'_1, ..., w'_m)$. Let us denote the vector space representations of the individual words as $v_1, ..., v_n$ and $v'_1, ..., v'_m$ respectively. A standard approach in CDS is to compose distributional representations by taking their vector sum $v_L = v_1 + v_2 ... + v_n$ and $v_R = v'_1 + ... + v'_m$ (Mitchell and Lapata, 2010). One of the most effective similarity measures is the cosine similarity, which is a normalized dot product. The distributional similarity between p_L and p_R under this model is $\mathrm{sim}(p_L, p_R) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{sim}(w_i, w'_j)$, where $\mathrm{sim}(w_i, w'_j)$ is the dot product between v_i and v'_j.

This similarity score is similar in spirit to a simplified version of our statistical model that restricts the set of allowable LCs H_p to be $\{(\{w_i\}, \{w'_j\}) | i \leq n, j \leq m\}$, i.e., only LCs of size 1. Indeed, taking H_p as above, and cosine similarity as the only feature (i.e., $w \in \mathbb{R}$), yields the distribution

$$P(y|p) \propto \sum_{(w_i, w'_j) \in H_p} \exp\left(w \cdot y \cdot \mathrm{sim}(w_i, w'_j)\right).$$

Algorithm	Test Set				Cross Validation			
	Sim	Basic	NoLDA	All	Sim	Basic	NoLDA	All
LATENTLC	**62.9**	**65.7**	**64.4**	**64.6**	**62.7** ± 1.9	**64.9** ± 1.9	**65.0** ± 1.7	**65.7** ±1.9
LEFTMOST	59.0	61.1	60.0	60.4	61.2 ± 2.1	62.5 ± 2.4	62.4 ±2.2	62.7 ± 2.0
VPREP	56.1	60.9	60.7	61.6*	58.1 ± 1.7	60.8 ± 2.2	60.4 ± 2.6	60.6 ± 2.2
ALLNEG	55.9				55.9			

Table 2: Accuracy results (in percents) for the various systems, followed by standard deviation where applicable. The rows correspond to the various systems as defined in Section 4. LATENTLC is our proposed model. Columns correspond to feature sets, from the least to the most inclusive. SIM includes only similarity features. BASIC adds POS-based and frequency features. NOLDA includes all features except LDA-based features. ALL is the full feature set. ALLNEG invariably predicts the label "no inference". Bold marks best overall accuracy per column, and * marks figures that are not significantly worse (McNemar's test, $p < 0.05$). The positive to negative label ratio was kept the same for the cross validation and test set scenarios. In all cases, LATENTLC substantially outperforms the baseline systems.

This derivation highlights the relation of a simplified version of our approach to the additive CDS model, as both approaches effectively average over the similarities of all pairs of words in p_L and p_R. The derivation also highlights a few advantages of our approach. First, our approach allows to straightforwardly introduce additional features and to weight them in a way most consistent with the task at hand. Second, it allows much more flexibility in defining the set of allowable LCs, H_p. Specifically, H_p may contain LCs of sizes greater than 1. Third, our approach uses standard probabilistic modelling, and therefore has a natural statistical interpretation.

In order to appreciate the effect of these advantages, we perform an experiment that takes H to be the set of all LCs of size 1, and uses a single similarity measure. We run a 10-fold cross-validation on our training data, obtaining 61.3% accuracy using COSINE and 62.2% accuracy using BInc. The performance gap between these results and the accuracy obtained by our full model (65.7%) underscores the latter's effectiveness in integrating multiple features and LCs.

Effectiveness of Optimization Method. Our maximization of the log-likelihood function is not guaranteed to converge to a global optimum. Therefore, the quality of the learned parameters may be sensitive to the initialization point. We hereby describe an experiment that tests the sensitivity of our approach to such variance.

Selecting the highest scoring feature set on our test set (i.e., BASIC), we ran the model with multiple initializers, by randomly perturbing our standard convex initializer (see Section 3). Concretely, given a convex initializer w, we select the starting point to be $w + \eta$, where $\eta_i \sim \mathcal{N}(0, \alpha|w_i|)$. We ran this experiment 400 times with $\alpha = 0.8$.

To combine the resulting weight vectors into a single classifier, we apply two types of standard approaches: a Product of Experts (Hinton, 2002), as well as a voting approach that selects the most frequently predicted label. Neither of these experiments yielded any significant performance gain. This demonstrates the robustness of our optimization method to the initialization point.

7 Conclusion

We have presented a novel approach to the distributional representation of multi-word predicates. Since MWPs demonstrate varying levels of compositionality, a uniform treatment of MWPs either as fixed expressions or through head words is lacking. Instead, our approach integrates multiple lexical units contained in the predicate. The approach takes into account both multi-word LCs that address low compositionality cases, as well as single-word LCs that address compositional cases and are more frequent. It assumes a latent distribution over the LCs of the predicates, and estimates it relative to a target application task.

We addressed the supervised inference identification task, obtaining substantial improvement over state-of-the-art baseline systems. In future work we intend to assess the benefit of this approach in MWP classes that are well-known from the literature. We believe that a permissive approach that integrates multiple analyses would perform better than standard single-analysis methods in a wide range of applications.

Acknowledgements. We would like to thank Mike Lewis, Reshef Meir, Oren Melamud, Michael Roth and Nathan Schneider for their helpful comments. This work was supported by ERC Advanced Fellowship 249520 GRAMPLUS.

References

Alex Alsina, Joan Wanda Bresnan, and Peter Sells. 1997. *Complex predicates*. Center for the Study of Language and Information.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *EMNLP*, pages 1183–1193.

Jonathan Berant, Jacob Goldberger, and Ido Dagan. 2011. Global learning of typed entailment rules. In *ACL*, pages 610–619.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Workshop on Distributional Semantics and Compositionality*, pages 21–28.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Miriam Butt. 2010. The light verb jungle: still hacking away. In *Complex predicates: cross-linguistic perspectives on event structure*, pages 48–78. Cambridge University Press.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. In J. van Bentham, M. Moortgat, and W. Buszkowski, editors, *Linguistic Analysis*, volume 36, pages 435–384.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Georgiana Dinu and Mirella Lapata. 2010. Topic models for meaning similarity in context. In *COLING: Posters*, pages 250–258.

Georgiana Dinu and Rui Wang. 2009. Inference rules and their application to recognizing textual entailment. In *EACL*, pages 211–219.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *EACL*, pages 337–344.

Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2009. Supervised synonym acquisition using distributional features and syntactic patterns. *Information and Media Technologies*, 4(2):558–582.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864.

Ray Jackendoff. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *EMNLP*, pages 1427–1432.

Mike Lewis and Mark Steedman. 2013. Combined distributional and logical semantics. *TACL*, 1:179–192.

Dekang Lin and Patrick Pantel. 2001. DIRT – discovery of inference rules from text. In *SIGKDD 2001*, pages 323–328.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *ACL*, pages 317–324.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *ACL Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pages 63–70.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *ACL workshop on Multiword expressions: analysis, acquisition and treatment*, pages 73–80.

Oren Melamud, Jonathan Berant, Ido Dagan, Jacob Goldberger, and Idan Szpektor. 2013a. A two level model for context sensitive inference rules. In *ACL 2013*, pages 1331–1340.

Oren Melamud, Ido Dagan, Jacob Goldberger, and Idan Szpektor. 2013b. Using lexical expansion to learn inference rules from sparse data. In *ACL: Short Papers*, pages 283–288.

Shachar Mirkin, Ido Dagan, and Maayan Geffet. 2006. Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. In *COLING-ACL: Poster Session*, pages 579–586.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Jorge. Nocedal and Stephen J Wright. 1999. *Numerical optimization*, volume 2. Springer New York.

Karl Pichotta and John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *EMNLP*, pages 636–646.

Carlos Ramisch, Aline Villavicencio, and Valia Kordoni. 2013. Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2):3.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *ACL*, pages 41–47.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *IJCNLP*, pages 210–218.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 46–50.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *ACL*, pages 424–434.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. In *CICLing*, pages 1–15.

Stefan Schoenmackers, Oren Etzioni, Daniel S Weld, and Jesse Davis. 2010. Learning first-order Horn clauses from web text. In *EMNLP*, pages 1088–1098.

Diarmuid Ó. Séaghdha. 2010. Latent variable models of selectional preference. In *ACL 2010*, pages 435–444.

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, 2nd edition.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between NE pairs. In *IWP*, pages 4–6.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *HLT-NAACL*, pages 304–311.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *COLING*, pages 849–856.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: contextual or statistical. In *ACL HLT 2011*, page 31.

Peter D Turney and Patrick Pantel. 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.

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013. Identifying English and Hungarian light verb constructions: A contrastive approach. In *ACL: Short Papers*, pages 255–261.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *EMNLP*, pages 81–88.

Hila Weisman, Jonathan Berant, Idan Szpektor, and Ido Dagan. 2012. Learning verb inference rules from linguistically-motivated evidence. In *EMNLP-CoNLL*, pages 194–204.

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*, volume 26, pages 28–35.

Alison Wray. 2008. *Formulaic language: Pushing the boundaries*. Oxford University Press.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *ACL-COLING*, pages 849–856.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. Crowdsourcing inference-rule evaluation. In *ACL: Short Papers*, pages 156–160.