Evaluating Distributional Models of Semantics for Syntactically Invariant Inference

Jackie CK Cheung and Gerald Penn
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
{jcheung,gpenn}@cs.toronto.edu

Abstract

A major focus of current work in distributional models of semantics is to construct phrase representations compositionally from word representations. However, the syntactic contexts which are modelled are usually severely limited, a fact which is reflected in the lexical-level WSD-like evaluation methods used. In this paper, we broaden the scope of these models to build sentence-level representations, and argue that phrase representations are best evaluated in terms of the inference decisions that they support, invariant to the particular syntactic constructions used to guide composition. We propose two evaluation methods in relation classification and QA which reflect these goals, and apply several recent compositional distributional models to the tasks. We find that the models outperform a simple lemma overlap baseline slightly, demonstrating that distributional approaches can already be useful for tasks requiring deeper inference.

1 Introduction

A number of unsupervised semantic models (Mitchell and Lapata, 2008, for example) have recently been proposed which are inspired at least in part by the distributional hypothesis (Harris, 1954)—that a word's meaning can be characterized by the contexts in which it appears. Such models represent word meaning as one or more high-dimensional vectors which capture the lexical and syntactic contexts of the word's occurrences in a training corpus.

Much of the recent work in this area has, following Mitchell and Lapata (2008), focused on the notion of compositionality as the litmus test of a truly semantic model. Compositionality is a natural way to construct representations of linguistic units larger than a word, and it has a long history in Montagovian semantics for dealing with argument structure and assembling rich semantical expressions of the kind found in predicate logic.

While compositionality may thus provide a convenient recipe for producing representations of propositionally typed phrases, it is not a necessary condition for a semantic representation. Rather, that distinction still belongs to the crucial ability to support inference. It is not the intention of this paper to argue for or against compositionality in semantic representations. Rather, our interest is in evaluating semantic models in order to determine their suitability for inference tasks. In particular, we contend that it is desirable and arguably necessary for a compositional semantic representation to support inference *invariantly*, in the sense that the particular syntactic construction that guided the composition should not matter relative to the representations of syntactically different phrases with the same meanings. For example, we can assert that *John threw the ball* and *The ball was thrown by John* have the same meaning for the purposes of inference, even though they differ syntactically.

An analogy can be drawn to research in image processing, in which it is widely regarded as important for the representations of images to be invariant to rotation and scaling. What we should want is a representation of sentence meaning that is invariant to diathesis, other regular syntactic alternations in the assignment of argument structure, and, ideally, even invariant to other meaning-preserving or near-preserving paraphrases.

Existing evaluations of distributional semantic models fall short of measuring this. One evaluation approach consists of lexical-level word substitution tasks which primarily evaluate a system's ability to disambiguate word senses within a controlled syntactic environment (McCarthy and Navigli, 2009, for example). Another approach is to evaluate parsing accuracy (Socher et al., 2010, for example), which is really a formalism-specific approximation to argument structure analysis. These evaluations may certainly be relevant to specific components of, for example, machine translation or natural language generation systems, but they tell us little about a semantic model's ability to support inference.

In this paper, we propose a general framework for evaluating distributional semantic models that build sentence representations, and suggest two evaluation methods that test the notion of structurally invariant inference directly. Both rely on determining whether sentences express the same semantic relation between entities, a crucial step in solving a wide variety of inference tasks like recognizing textual entailment, information retrieval, question answering, and summarization.

The first evaluation is a relation classification task, where a semantic model is tested on its ability to recognize whether a pair of sentences both contain a particular semantic relation, such as *Company X acquires Company Y*. The second task is a question answering task, the goal of which is to locate the sentence in a document that contains the answer. Here, the semantic model must match the question, which expresses a proposition with a missing argument, to the answer-bearing sentence which contains the full proposition.

We apply these new evaluation protocols to several recent distributional models, extending several of them to build sentence representations. We find that the models outperform a simple lemma overlap model only slightly, but that combining these models with the lemma overlap model can improve performance. This result is likely due to weaknesses in current models' ability to deal with issues such as named entities, coreference, and negation, which are not emphasized by existing evaluation methods, but it does suggest that distributional models of semantics can play a more central role in systems that require deep, precise inference.

2 Compositionality and Distributional Semantics

The idea of compositionality has been central to understanding contemporary natural language semantics from an historiographic perspective. The idea is often credited to Frege, although in fact Frege had very little to say about compositionality that had not already been repeated since the time of Aristotle (Hodges, 2005). Our modern notion of compositionality took shape primarily with the work of Tarski (1956), who was actually arguing that a central difference between formal languages and natural languages is that natural language is not compositional. This in turn was the "the contention that an important theoretical difference exists between formal and natural languages," that Richard Montague so famously rejected (Montague, 1974). Compositionality also features prominently in Fodor and Pylyshyn's (1988) rejection of early connectionist representations of natural language semantics, which seems to have influenced Mitchell and Lapata (2008) as well.

Logic-based forms of compositional semantics have long strived for syntactic invariance in meaning representations, which is known as the doctrine of the canonical form. The traditional justification for canonical forms is that they allow easy access to a knowledge base to retrieve some desired information, which amounts to a form of inference. Our work can be seen as an extension of this notion to distributional semantic models with a more general notion of representational similarity and inference.

There are many regular alternations that semantics models have tried to account for such as passive or dative alternations. There are also many lexical paraphrases which can take drastically different syntactic forms. Take the following example from Poon and Domingos (2009), in which the same semantic relation can be expressed by a transitive verb or an attributive prepositional phrase:

(1) *Utah borders Idaho.*
Utah is next to Idaho.

In distributional semantics, the original sentence similarity test proposed by Kintsch (2001) served as the inspiration for the evaluation performed by Mitchell and Lapata (2008) and most later work in the area. Intransitive verbs are given

in the context of their syntactic subject, and candidate synonyms are ranked for their appropriateness. This method targets the fact that a synonym is appropriate for only some of the verb's senses, and the intended verb sense depends on the surrounding context. For example, *burn* and *beam* are both synonyms of *glow*, but given a particular subject, one of the synonyms (called the High similarity landmark) may be a more appropriate substitution than the other (the Low similarity landmark). So, if *the fire* is the subject, *glowed* is the High similarity landmark, and *beamed* the Low similarity landmark.

Fundamentally, this method was designed as a demonstration that compositionality in computing phrasal semantic representations does not interfere with the ability of a representation to synthesize non-compositional collocation effects that contribute to the disambiguation of homographs. Here, word-sense disambiguation is implicitly viewed as a very restricted, highly lexicalized case of inference for selecting the appropriate disjunct in the representation of a word's meaning.

Kintsch (2001) was interested in sentence similarity, but he only conducted his evaluation on a few hand-selected examples. Mitchell and Lapata (2008) conducted theirs on a much larger scale, but chose to focus only on this single case of syntactic combination, intransitive verbs and their subjects, in order to "factor out inessential degrees of freedom" to compare their various alternative models more equitably. This was not necessary—using the same, sufficiently large, unbiased but syntactically heterogeneous sample of evaluation sentences would have served as an adequate control—and this decision furthermore prevents the evaluation from testing the desired invariance of the semantic representation.

Other lexical evaluations suffer from the same problem. One uses the WordSim-353 dataset (Finkelstein et al., 2002), which contains human word pair similarity judgments that semantic models should reproduce. However, the word pairs are given without context, and homography is unaddressed. Also, it is unclear how reliable the similarity scores are, as different annotators may interpret the integer scale of similarity scores differently. Recent work uses this dataset mostly for parameter tuning. Another is the lexical paraphrase task of McCarthy and Navigli (2009), in which words are given in the context of the surrounding sentence, and the task is to rank a given list of proposed substitutions for that word. The list of substitutions as well as the correct rankings are elicited from annotators. This task was originally conceived as an applied evaluation of WSD systems, not an evaluation of phrase representations.

Parsing accuracy has been used as a preliminary evaluation of semantic models that produce syntactic structure (Socher et al., 2010; Wu and Schuler, 2011). However, syntax does not always reflect semantic content, and we are specifically interested in supporting syntactic invariance when doing semantic inference. Also, this type of evaluation is tied to a particular grammar formalism.

The existing evaluations that are most similar in spirit to what we propose are paraphrase detection tasks that do not assume a restricted syntactic context. Washtell (2011) collected human judgments on the general meaning similarity of candidate phrase pairs. Unfortunately, no additional guidance on the definition of "most similar in meaning" was provided, and it appears likely that subjects conflated lexical, syntactic, and semantic relatedness. Dolan and Brockett (2005) define paraphrase detection as identifying sentences that are in a bidirectional entailment relation. While such sentences do support exactly the same inferences, we are also interested in the inferences that can be made from similar sentences that are not paraphrases according to this strict definition — a situation that is more often encountered in end applications. Thus, we adopt a less restricted notion of paraphrasis.

3 An Evaluation Framework

We now describe a simple, general framework for evaluating semantic models. Our framework consists of the following components: a semantic model to be evaluated, pairs of sentences that are considered to have high similarity, and pairs of sentences that are considered to have low similarity.

In particular, the semantic model is a binary function, $s = \mathcal{M}(x, x')$, which returns a real-valued similarity score, s, given a pair of arbitrary linguistic units (that is, words, phrases, sentences, etc.), x and x'. Note that this formulation of the semantic model is agnostic to whether the models use compositionality to build a phrase represen-

tation from constituent representations, and even to the actual representation used. The model is tested by applying it to each element in the following two sets:

$$H = \{(h, h') | h \text{ and } h' \text{ are linguistic units with high similarity}\} \quad (2)$$

$$L = \{(l, l') | l \text{ and } l' \text{ are linguistic units with low similarity}\} \quad (3)$$

The resulting sets of similarity scores are:

$$\mathcal{S}^H = \{\mathcal{M}(h, h') | (h, h') \in H\} \quad (4)$$

$$\mathcal{S}^L = \{\mathcal{M}(l, l') | (l, l') \in L\} \quad (5)$$

The semantic model is evaluated according to its ability to separate $\mathcal{S}^H$ and $\mathcal{S}^L$. We will define specific measures of separation for the tasks that we propose shortly. While the particular definitions of "high similarity" and "low similarity" depend on the task, at the crux of both our evaluations is that two sentences are similar if they express the same semantic relation between a given entity pair, and dissimilar otherwise. This threshold for similarity is closely tied to the argument structure of the sentence, and allows considerable flexibility in the other semantic content that may be contained in the sentence, unlike the bidirectional paraphrase detection task. Yet it ensures that a consistent and useful distinction for inference is being detected, unlike unconstrained similarity judgments.

Also, compared to word similarity assessments or paraphrase elicitation, determining whether a sentence expresses a semantic relation is a much easier task cognitively for human judges. This binary judgment does not involve interpreting a numerical scale or coming up with an open-ended set of alternative paraphrases. It is thus easier to get reliable annotated data.

Below, we present two tasks that instantiate this evaluation framework and choice of similarity threshold. They differ in that the first is targeted towards recognizing declarative sentences or phrases, while the second is targeted towards a question answering scenario, where one argument in the semantic relation is queried.

3.1 Task 1: Relation Classification

The first task is a relation classification task. Relation extraction and recognition are central to a variety of other tasks, such as information retrieval, ontology construction, recognizing textual entailment and question answering.

In this task, the high and the low similarity sentence pairs are constructed in the following manner. First, a target semantic relation, such as *Company X acquires Company Y* is chosen, and entities are chosen for each slot in the relation, such as *Company X=Pfizer* and *Company Y=Rinat Neuroscience*. Then, sentences containing these entities are extracted and divided into two subsets. In one of them, E, the entities are in the target semantic relation, while in the other, NE, they are not. The evaluation sets H and L are then constructed as follows:

$$H = E \times E \setminus \{(e, e) | e \in E\} \quad (6)$$

$$L = E \times NE \quad (7)$$

In other words, the high similarity sentence pairs are all the pairs where both express the target semantic relation, except the pairs between a sentence and itself, while the low similarity pairs are all the pairs where exactly one of the two sentences expresses the target relation.

Several sentences expressing the relation *Pfizer acquires Rinat Neuroscience* are shown in Examples 8 to 10. These sentences illustrate the amount of syntactic and lexical variation that the semantic model must recognize as expressing the same semantic relation. In particular, besides recognizing synonymy or near-synonymy at the lexical level, models must also account for subcategorization differences, extra arguments or adjuncts, and part-of-speech differences due to nominalization.

(8) *Pfizer buys Rinat Neuroscience to extend neuroscience research and in doing so acquires a product candidate for OA.* (lexical difference)

(9) *A month earlier, Pfizer paid an estimated several hundred million dollars for biotech firm Rinat Neuroscience.* (extra argument, subcategorization)

(10) *Pfizer to Expand Neuroscience Research With Acquisition of Biotech Company Rinat Neuroscience* (nominalization)

Since our interest is to measure the models' ability to separate $\mathcal{S}^H$ and $\mathcal{S}^L$ in an unsupervised setting, standard supervised classification accuracy is not applicable. Instead, we employ

the area under a ROC curve (AUC), which does not depend on choosing an arbitrary classification threshold. A ROC curve is a plot of the true positive versus false positive rate of a binary classifier as the classification threshold is varied. The area under a ROC curve can thus be seen as the performance of linear classifiers over the scores produced by the semantic model. The AUC can also be interpreted as the probability that a randomly chosen positive instance will have a higher similarity score than a randomly chosen negative instance. A random classifier is expected to have an AUC of 0.5.

3.2 Task 2: Restricted QA

The second task that we propose is a restricted form of question answering. In this task, the system is given a question q and a document $\mathcal{D}$ consisting of a list of sentences, in which one of the sentences contains the answer to the question. We define:

$$H = \{(q, d) | d \in \mathcal{D} \text{ and } d \text{ answers } q\} \quad (11)$$

$$L = \{(q, d) | d \in \mathcal{D} \text{ and } d \text{ does not answer } q\} \quad (12)$$

In other words, the sentences are divided into two subsets; those that contain the answer to q should be similar to q, while those that do not should be dissimilar. We also assume that only one sentence in each document contains the answer, so H contains only one sentence.

Unrestricted question answering is a difficult problem that forces a semantic representation to deal sensibly with a number of other semantic issues such as coreference and information aggregation which still seem to be out of reach for contemporary distributional models of meaning. Since our focus in this work is on argument structure semantics, we restrict the question-answer pairs to those that only require dealing with paraphrases of this type.

To do so, we semi-automatically restrict the question-answer pairs by using the output of an unsupervised clustering semantic parser (Poon and Domingos, 2009). The semantic parser clusters semantic sub-expressions derived from a dependency parse of the sentence, so that those subexpressions that express the same semantic relations are clustered. The parser is used to answer questions, and the output of the parser is manually checked. We use only those cases that have thus been determined to be correct question-answer pairs. As a result of this restriction, this task is rather more like Task 1 in how it tests a model's ability to recognize lexical and syntactic paraphrases. This task also involves recognizing voicing alternations, which were automatically extracted by the semantic parser.

An example of a question-answer pair involving a voicing alternation that is used in this task is presented in Example 13.

(13) Q: *What does il-2 activate?*
A: *PI3K*
Sentence: *Phosphatidyl inositol 3-kinase (PI3K) is activated by IL-2.*

Since there is only one element in H and hence $\mathcal{S}^H$ for each question and document, we measure the separation between $\mathcal{S}^H$ and $\mathcal{S}^L$ using the rank of the score of answer-bearing sentence among the scores of all the sentences in the document. We normalize the rank so that it is between 0 (ranked least similar) and 1 (ranked most similar). Where ties occur, the sentence is ranked as if it were in the median position among the tied sentences. If the question-answer pairs are zero-indexed by i, $answer(i)$ is the index of the sentence containing the answer for the ith pair, and $length(i)$ is the number of sentences in the document, then the mean normalized rank score of a system is:

$$\overline{norm\_rank} = \mathop{\mathbf{E}}_{i} \left[1 - \frac{answer(i)}{length(i) - 1} \right] \quad (14)$$

4 Experiments

We drew a number of recent distributional semantic models to compare in this paper. We first describe the models and our reimplementation of them, before describing the tasks and the datasets used in detail and the results.

4.1 Distributional Semantic Models

We tested four recent distributional models and a lemma overlap baseline, which we now describe. We extended several of the models to compositionally construct phrase representations using component-wise vector addition and multiplication, as we note below. Since the focus of this paper is on evaluation methods for such models, we did not experiment with other compositionality

operators. We do note, however, that component-wise operators have been popular in recent literature, and have been applied across unrestricted syntactic contexts (Mitchell and Lapata, 2009), so there is value in evaluating the performance of these operators in itself. The models were trained on the Gigaword corpus (2nd ed., ~2.3B words). All models use cosine similarity to measure the similarity between representations, except for the baseline model.

Lemma Overlap This baseline simply represents a sentence as the counts of each lemma present in the sentence after removing stop words. Let a sentence x consist of lemma-tokens $m_1, \ldots, m_{|x|}$. The similarity between two sentences is then defined as

$$\mathcal{M}(x, x') = \#In(x, x') + \#In(x', x) \quad (15)$$

$$\#In(x, x') = \sum_{i=1}^{|x|} \mathbf{1}_{x'}(m_i) \quad (16)$$

where $\mathbf{1}_{x'}(m_i)$ is an indicator function that returns 1 if $m_i \in x'$, and 0 otherwise. This definition accounts for multiple occurrences of a lemma.

M&L Mitchell and Lapata (2008) propose a framework for compositional distributional semantics using a standard term-context vector space word representation. A phrase is represented as a vector of context-word counts (actually, pmi-scaled values), which is derived compositionally by a function over constituent vectors, such as component-wise addition or multiplication. This model ignores syntactic relations and is insensitive to word-order.

E&P Erk and Padó (2008) introduce a structured vector space model which uses syntactic dependencies to model the selectional preferences of words. The vector representation of a word in context depends on the inverse selectional preferences of its dependents, and the selectional preferences of its head. For example, suppose *catch* occurs with a dependent *ball* in a direct object relation. The vector for *catch* would then be influenced by the inverse direct object preferences of *ball* (e.g. *throw*, *organize*), and the vector for *ball* would be influenced by the selectional preferences of *catch* (e.g. *cold*, *drift*). More formally, given words a and b in a dependency relation r, a distributional representation of a, v_a, the representation of a in context, a', is given by

$$a' = v_a \odot R_b(r^{-1}) \quad (17)$$

$$R_b(r) = \sum_{c: f(c,r,b) > \theta} f(c, r, b) \cdot v_c, \quad (18)$$

where $R_b(r)$ is the vector describing the selectional preference of word b in relation r, $f(c, r, b)$ is the frequency of this dependency triple, θ is a frequency threshold to weed out uncommon dependency triples (10 in our experiments), and $\odot$ is a vector combination operator, here component-wise multiplication. We extend the model to compute sentence representations from the contextualized word vectors using component-wise addition and multiplication.

TFP Thater et al. (2010)'s model is also sensitive to selectional preferences, but to two degrees. For example, the vector for *catch* might contain a dimension labelled (`OBJ,OBJ-1,throw`), which indicates the strength of connection between the two verbs through all of the co-occurring direct objects which they share. Unlike E&P, TFP's model encodes the selectional preferences in a single vector using frequency counts. We extend the model to the sentence level with component-wise addition and multiplication, and word vectors are contextualized by the dependency neighbours. We use a frequency threshold of 10 and a pmi threshold of 2 to prune infrequent word and dependencies.

D&L Dinu and Lapata (2010) (D&L) assume a global set of latent senses for all words, and models each word as a mixture over these latent senses. The vector for a word t_i in the context of a word c_j is modelled by

$$v(t_i, c_j) = P(z_1|t_i, c_j), \ldots P(z_K|t_i, c_j) \quad (19)$$

where $z_{1\ldots K}$ are the latent senses. By making independence assumptions and decomposing probabilities, training becomes a matter of estimating the probability distributions $P(z_k|t_i)$ and $P(c_j|z_k)$ from data. While Dinu and Lapata (2010) describe two methods to do so, based on non-negative matrix factorization and latent Dirichlet allocation, the performances are similar, so we tested only the latent Dirichlet allocation method. Like the two previous models, we extend the model to build sentence representations

	Pfizer/Rinat N.	Yahoo/Inktomi	Besson/Paris	Antoinette/Vienna	Average
Overlap	0.7393	0.6007	0.7395	0.8914	0.7427
Models trained on the entire GigaWord					
M&L add	0.6196	0.5387	0.5259	0.7275	0.6029
M&L mult	0.9036	0.6099	0.6443	0.8467	0.7511
D&L add	0.9214	0.8168	0.6989	0.8932	**0.8326**
D&L mult	0.7732	0.6734	0.6527	0.7659	0.7163
Models trained on the AFP section					
E&P add	0.7536	0.4933	0.2780	0.6408	0.5414
E&P mult	0.5268	0.5328	0.5252	0.8421	0.6067
TFP add	0.4357	0.5325	0.8725	0.7183	0.6398
TFP mult	0.5554	0.5524	0.7283	0.6917	0.6320
M&L add	0.5643	0.5504	0.4594	0.7640	0.5845
M&L mult	0.8679	0.6324	0.4356	0.8258	0.6904
D&L add	0.8143	0.9062	0.6373	0.8664	**0.8061**
D&L mult	0.8429	0.7461	0.645	0.5948	0.7072

Table 1: Task 1 results in AUC scores. The values in bold indicate the best performing model for a particular training corpus. The expected random baseline performance is 0.5.

Entities: $\{X, Y\}$	+	*N*
Relation: acquires		
{Pfizer, Rinat Neuroscience}	41	50
{Yahoo, Inktomi}	115	433
Relation: was born in		
{Luc Besson, Paris}	6	126
{Marie Antoinette, Vienna}	39	105

Table 2: Task 1 dataset characteristics. *N* is the total number of sentences. + is the number of sentences that express the relation.

from the contextualized representations. We set the number of latent senses to 1200, and train for 600 Gibbs sampling iterations.

4.2 Training and Parameter Settings

We reimplemented these four models, following the parameter settings described by previous work where possible, though we also aimed for consistency in parameter settings between models (for example, in the number of context words). For the non-baseline models, we followed previous work and model only the 30000 most frequent lemmata. Context vectors are constructed using a symmetric window of 5 words, and their dimensions represent the 3000 most frequent lemmatized context words excluding stop words. Due to resource limitations, we trained the syntactic models over the AFP subset of Gigaword (~338M words). We also trained the other two models on just the AFP portion for comparison. Note that the AFP portion of Gigaword is three times larger than the BNC corpus (~100M words), on which several previous syntactic models were trained. Because our main goal is to test the general performance of the models and to demonstrate the feasibility of our evaluation methods, we did not further tune the parameter settings to each of the tasks, as doing so would likely only yield minor improvements.

4.3 Task 1

We used the dataset by Bunescu and Mooney (2007), which we selected because it contains multiple realizations of an entity pair in a target semantic relation, unlike similar datasets such as the one by Roth and Yih (2002). Controlling for the target entity pair in this manner makes the task more difficult, because the semantic model cannot make use of distributional information about the entity pair in inference. The dataset is separated into subsets depending on the target binary relation (*Company X acquires Company Y* or *Person X was born in Place Y*) and the entity pair (e.g., *Yahoo* and *Inktomi*) (Table 2).

The dataset was constructed semi-automatically using a Google search for the two entities in order with up to seven content words in between. Then, the extracted sentences were hand-labelled with whether they express the target relation. Because the order of the entities has been fixed, passive alternations do not appear

	Pure models		*Mixed models*	
	All	*Subset*	*All*	*Subset*
Overlap	**0.8770**	**0.7291**	0.8770	0.7291
Models trained on the entire GigaWord				
M&L add	0.7467	0.6106	0.8782	0.7523
M&L mult	0.5331	0.5690	0.8841	0.7678
D&L add	0.6552	0.5716	0.8791	0.7539
D&L mult	0.5488	0.5255	0.8841	0.7466
Models trained on the AFP section				
E&P add	0.4589	0.4516	0.8748	0.7375
E&P mult	0.5201	0.5584	0.8882	0.7719
TFP add	0.6887	0.6443	**0.8940**	**0.7871**
TFP mult	0.5210	0.5199	0.8785	0.7432
M&L add	0.7588	0.6206	0.8710	0.7371
M&L mult	0.5710	0.5540	0.8801	0.7540
D&L add	0.6358	0.5402	0.8713	0.7305
D&L mult	0.5647	0.5461	0.8856	0.7683

Table 3: Task 2 results, in normalized rank scores. *Subset* is the cases where lemma overlap does not achieve a perfect score. The two columns on the right indicate performance using the sum of the scores from the lemma overlap and the semantic model. The expected random baseline performance is 0.5.

in this dataset.

The results for Task 1 indicate that the D&L addition model performs the best (Table 1), though the lemma overlap model presents a surprisingly strong baseline. The syntax-modulated E&P and TFP models perform poorly on this task, even when compared to the other models trained on the AFP subset. The M&L multiplication model outperforms the addition model, a result which corroborates previous findings on the lexical substitution task. The same does not hold in the D&L latent sense space. Overall, some of the datasets (*Yahoo* and *Antoinette*) appear to be easier for the models than others (*Pfizer* and *Besson*), but more entity pairs and relations would be needed to investigate the models' variance across datasets.

4.4 Task 2

We used the question-answer pairs extracted by the Poon and Domingos (2009) semantic parser from the GENIA biomedical corpus that have been manually checked to be correct (295 pairs). Because our models were trained on newspaper text, they required adaptation to this specialized domain. Thus, we also trained the M&L, E&P and TFP models on the GENIA corpus, backing off to word vectors from the GENIA corpus when a word vector could not be found in the Gigaword-trained model. We could not do this for the D&L model, since the global latent senses that are found by latent Dirichlet allocation training do not have any absolute meaning that holds across multiple runs. Instead, we found the 5 words in the Gigaword-trained D&L model that were closest to each novel word in the GENIA corpus according to cosine similarity over the co-occurrence vectors of the words in the GENIA corpus, and took their average latent sense distributions as the vector for that word.

Unlike in Task 1, there is no control for the named entities in a sentence, because one of the entities in the semantic relation is missing. Also, distributional models have problems in dealing with named entities which are common in this corpus, such as the names of genes and proteins. To address these issues, we tested hybrid models where the similarity score from a semantic model is added to the similarity score from the lemma overlap model.

The results are presented in Table 3. Lemma overlap again presents a strong baseline, but the hybridized models are able to outperform simple lemma overlap. Unlike in Task 1, the E&P and TFP models are comparable to the D&L model, and the mixed TFP addition model achieves the best result, likely due to the need to more precisely distinguish syntactic roles in this task. The D&L addition model, which achieved the best performance in Task 1, does not perform as well in this task. This could be due to the domain adaptation procedure for the D&L model, which could not be reasonably trained on such a small, specialized corpus.

5 Related Work

Turney and Pantel (2010) survey various types of vector space models and applications thereof in computational linguistics. We summarize below a number of other word- or phrase-level distributional models.

Several approaches are specialized to deal with homography. The top-down *multi-prototype* approach determines a number of senses for each word, and then clusters the occurrences of the word (Reisinger and Mooney, 2010) into these senses. A prototype vector is created for each of these sense clusters. When a new occurrence

of a word is encountered, it is represented as a combination of the prototype vectors, with the degree of influence from each prototype determined by the similarity of the new context to the existing sense contexts. In contrast, the bottom-up *exemplar*-based approach assumes that each occurrence of a word expresses a different sense of the word. The most similar senses of the word are activated when a new occurrence of it is encountered and combined, for example with a kNN algorithm (Erk and Padó, 2010).

The models we compared and the above work assume each dimension in the feature vector corresponds to a context word. In contrast, Washtell (2011) uses potential paraphrases directly as dimensions in his *expectation vectors*. Unfortunately, this approach does not outperform various context word-based approaches in two phrase similarity tasks.

In terms of the vector composition function, component-wise addition and multiplication are the most popular in recent work, but there exist a number of other operators such as tensor product and convolution product, which are reviewed by Widdows (2008). Instead of vector space representations, one could also use a matrix space representation with its much more expressive matrix operators (Rudolph and Giesbrecht, 2010). So far, however, this has only been applied to specific syntactic contexts (Baroni and Zamparelli, 2010; Guevara, 2010; Grefenstette and Sadrzadeh, 2011), or tasks (Yessenalina and Cardie, 2011).

Neural networks have been used to learn both phrase structure and representations. In Socher et al. (2010), word representations learned by neural network models such as (Bengio et al., 2006; Collobert and Weston, 2008) are fed as input into a recursive neural network whose nodes represent syntactic constituents. Each node models both the probability of the input forming a constituent and the phrase representation resulting from composition.

6 Conclusions

We have proposed an evaluation framework for distributional models of semantics which build phrase- and sentence-level representations, and instantiated two evaluation tasks which test for the crucial ability to recognize whether sentences express the same semantic relation. Our results demonstrate that compositional distributional models of semantics already have some utility in the context of more empirically complex semantic tasks than WSD-like lexical substitution tasks, in which compositional invariance is a requisite property. Simply computing lemma overlap, however, is a very competitive baseline, due to issues in these protocols with named entities and domain adaptivity. The better performance of the mixture models in Task 2 shows that such weaknesses can be addressed by hybrid semantic models. Future work should investigate more refined versions of such hybridization, as well as extend this idea to other semantic phenomena like coreference, negation and modality.

We also observe that no single model or composition operator performs best for all tasks and datasets. The latent sense mixture model of Dinu and Lapata (2010) performs well in recognizing semantic relations in general web text. Because of the difficulty of adapting it to a specialized domain, however, it does less well in biomedical question answering, where the syntax-based model of Thater et al. (2010) performs the best. A more thorough investigation of the factors that can predict the performance and/or invariance of a given composition operator is warranted.

In the future, we would like to evaluate other models of compositional semantics that have been recently proposed. We would also like to collect more comprehensive test data, to increase the external validity of our evaluations.

Acknowledgments

We would like to thank Georgiana Dinu and Stefan Thater for help with reimplementing their models. Saif Mohammad, Peter Turney, and the anonymous reviewers provided valuable comments on drafts of this paper. This project was supported by the Natural Sciences and Engineering Research Council of Canada.

References

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

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain.

2006. Neural probabilistic language models. *Innovations in Machine Learning*, pages 137–186.

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 576–583.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, page 160–167.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pages 9–16.

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.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

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

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37.

Zeller S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Wilfred Hodges. 2005. The interplay of fact and theory in separating syntax from meaning. In *Workshop on Empirical Challenges and Analytical Alternatives to Strict Compositionality*.

Walter Kintsch. 2001. Predication. *Cognitive Science*, 25(2):173–202.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

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

Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 430–439.

Richard Montague. 1974. English as a formal language. *Formal Philosophy*, pages 188–221.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Joseph Reisinger and Raymond J. Mooney. 2010. 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*.

Dan Roth and Wen-tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 835–841.

Sebastian Rudolph and Eugenie Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916.

Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. *Proceedings of the Deep Learning and Unsupervised Feature Learning Workshop of NIPS 2010*, pages 1–9.

Alfred Tarski. 1956. The concept of truth in formalized languages. *Logic, Semantics, Metamathematics*, pages 152–278.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Justin Washtell. 2011. Compositional expectation: A purely distributional model of compositional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 285–294.

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*.

Stephen Wu and William Schuler. 2011. Structured composition of semantic vectors. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 295–304.

Ainur Yessenalina and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 172–182.