Probabilistic models of similarity in syntactic context

Diarmuid Ó Séaghdha
Computer Laboratory
University of Cambridge
United Kingdom
do242@cl.cam.ac.uk

Anna Korhonen
Computer Laboratory
University of Cambridge
United Kingdom
Anna.Korhonen@cl.cam.ac.uk

Abstract

This paper investigates novel methods for incorporating syntactic information in probabilistic latent variable models of lexical choice and contextual similarity. The resulting models capture the effects of context on the interpretation of a word and in particular its effect on the appropriateness of replacing that word with a potentially related one. Evaluating our techniques on two datasets, we report performance above the prior state of the art for estimating sentence similarity and ranking lexical substitutes.

1 Introduction

Distributional models of lexical semantics, which assume that aspects of a word's meaning can be related to the contexts in which that word is typically used, have a long history in Natural Language Processing (Spärck Jones, 1964; Harper, 1965). Such models still constitute one of the most popular approaches to lexical semantics, with many proven applications. Much work in distributional semantics treats words as non-contextualised units; the models that are constructed can answer questions such as "how similar are the words *body* and *corpse*?" but do not capture the way the syntactic context in which a word appears can affect its interpretation. Recent developments (Mitchell and Lapata, 2008; Erk and Padó, 2008; Thater et al., 2010; Grefenstette et al., 2011) have aimed to address compositionality of meaning in terms of distributional semantics, leading to new kinds of questions such as "how similar are the usages of the words *body* and *corpse* in the phrase *the body/corpse deliberated the motion*. . . ?" and "how similar are the phrases *the body deliberated the motion* and *the corpse rotted*?". In this paper we focus on answering questions of the former type and investigate models that describe the effect of syntactic context on the meaning of a single word.

The work described in this paper uses probabilistic latent variable models to describe patterns of syntactic interaction, building on the selectional preference models of Ó Séaghdha (2010) and Ritter et al. (2010) and the lexical substitution models of Dinu and Lapata (2010). We propose novel methods for incorporating information about syntactic context in models of lexical choice, yielding a probabilistic analogue to dependency-based models of contextual similarity. Our models attain state-of-the-art performance on two evaluation datasets: a set of sentence similarity judgements collected by Mitchell and Lapata (2008) and the dataset of the English Lexical Substitution Task (McCarthy and Navigli, 2009). In view of the well-established effectiveness of dependency-based distributional semantics and of probabilistic frameworks for semantic inference, we expect that our approach will prove to be of value in a wide range of application settings.

2 Related work

The literature on distributional semantics is vast; in this section we focus on outlining the research that is most directly related to capturing effects of context and compositionality.[1] Mitchell and Lapata (2008)

[1] The interested reader is referred to Padó and Lapata (2007) and Turney and Pantel (2010) for a general overview.

follow Kintsch (2001) in observing that most distributional approaches to meaning at the phrase or sentence level assume that the contribution of syntactic structure can be ignored and the meaning of a phrase is simply the commutative sum of the meanings of its constituent words. As Mitchell and Lapata argue, this assumption clearly leads to an impoverished model of semantics. Mitchell and Lapata investigate a number of simple methods for combining distributional word vectors, concluding that pointwise multiplication best corresponds to the effects of syntactic interaction.

Erk and Padó (2008) introduce the concept of a *structured vector space* in which each word is associated with a set of selectional preference vectors corresponding to different syntactic dependencies. Thater et al. (2010) develop this geometric approach further using a space of *second-order* distributional vectors that represent the words typically co-occurring with the contexts in which a word typically appears. The primary concern of these authors is to model the effect of context on word meaning; the work we present in this paper uses similar intuitions in a probabilistic modelling framework.

A parallel strand of research seeks to represent the meaning of larger compositional structures using matrix and tensor algebra (Smolensky, 1990; Rudolph and Giesbrecht, 2010; Baroni and Zamparelli, 2010; Grefenstette et al., 2011). This nascent approach holds the promise of providing a much richer notion of context than is currently exploited in semantic applications.

Probabilistic latent variable frameworks for generalising about contextual behaviour (in the form of verb-noun selectional preferences) were proposed by Pereira et al. (1993) and Rooth et al. (1999). Latent variable models are also conceptually similar to non-probabilistic dimensionality reduction techniques such as Latent Semantic Analysis (Landauer and Dumais, 1997). More recently, Ó Séaghdha (2010) and Ritter et al. (2010) reformulated Rooth et al.'s approach in a Bayesian framework using models related to Latent Dirichlet Allocation (Blei et al., 2003), demonstrating that this "topic modelling" architecture is a very good fit for capturing selectional preferences. Reisinger and Mooney (2010) investigate nonparametric Bayesian models for teasing apart the context distributions of polysemous words. As described in Section 3 below, Dinu and Lapata (2010) propose an LDA-based model for lexical substitution; the techniques presented in this paper can be viewed as a generalisation of theirs. Topic models have also been applied to other classes of semantic task, for example word sense disambiguation (Li et al., 2010), word sense induction (Brody and Lapata, 2009) and modelling human judgements of semantic association (Griffiths et al., 2007).

3 Models

3.1 Latent variable context models

In this paper we consider generative models of lexical choice that assign a probability to a particular word appearing in a given linguistic context. In particular, we follow recent work (Dinu and Lapata, 2010; Ó Séaghdha, 2010; Ritter et al., 2010) in assuming a latent variable model that associates contexts with distributions over a shared set of variables and associates each variable with a distribution over the vocabulary of word types:

$$P(w|c) = \sum_{z \in Z} P(w|z)P(z|c) \qquad (1)$$

The set of latent variables Z is typically much smaller than the vocabulary size; this induces a (soft) clustering of the vocabulary. Latent Dirichlet Allocation (Blei et al., 2003) is a powerful method for learning such models from a text corpus in an unsupervised way; LDA was originally applied to document modelling, but it has recently been shown to be very effective at inducing models for a variety of semantic tasks (see Section 2).

Given the latent variable framework in (1) we can develop a generative model of paraphrasing a word o with another word n in a particular context c:

$$P_{C \to T}(n|o,c) = \sum_{z} P(n|z)P(z|o,c) \qquad (2)$$

$$P(z|o,c) = \frac{P(o|z)P(z|c)}{\sum_{z'} P(o|z')P(z'|c)} \qquad (3)$$

In words, the probability $P(n|o,c)$ is the probability that n would be generated given the latent variable distribution associated with seeing o in context c; this latter distribution $P(z|o,c)$ can be derived using Bayes' rule and the assumption $P(o|z,c) = P(o|z)$. Given a set of contexts C in which an instance o

appears (e.g., it may be both the subject of a verb and modified by an adjective), (2) and (3) become:

$$P_{C\to T}(n|o,C) = \sum_z P(n|z)P(z|o,C) \quad (4)$$

$$P(z|o,C) = \frac{P(o|z)P(z|C)}{\sum_{z'} P(o|z')P(z'|C)} \quad (5)$$

$$P(z|C) = \frac{\prod_{c\in C} P(z|c)}{\sum_{z'} \prod_{c\in C} P(z'|c)} \quad (6)$$

Equation (6) can be viewed as defining a "product of experts" model (Hinton, 2002). Dinu and Lapata (2010) also use a similar formulation to (5), except that $P(z|o,C)$ is factorised over $P(z|o,C)$ rather than just $P(z|C)$:

$$P_{DL10}(z|o,C) = \prod_{c\in C} \frac{P(o|z)P(z|c)}{\sum_{z'} P(o|z')P(z'|c)} \quad (7)$$

In Section 5 below, we find that using (5) rather than (7) gives better results.

The model described above (henceforth $C \to T$) models the dependence of a target word on its context. An alternative perspective is to model the dependence of a set of contexts on a target word, i.e., we induce a model

$$P(c|w) = \sum_z P(c|z)P(z|w) \quad (8)$$

Making certain assumptions, a formula for $P(n|o,c)$ can be derived from (8):

$$P_{T\to C}(n|o,c) = \frac{P(c|o,n)P(n|o)}{P(c|o)} \quad (9)$$

$$P(c|o,n) = \sum_z P(c|z)P(z|o,n)$$

$$P(z|o,n) = \frac{P(z|o)P(z|n)}{\sum_{z'} P(z'|o)P(z'|n)} \quad (10)$$

$$P(c|o) = \sum_z P(c|z)P(z|o) \quad (11)$$

$$P(n|o) = 1/V \quad (12)$$

The assumption of a uniform prior $P(n|o)$ on the choice of a paraphrase n for o is clearly not appropriate from a language modelling perspective (one could imagine an alternative $P(n)$ based on corpus frequency), but in the context of measuring semantic similarity it serves well. The $T \to C$ model for a set of contexts C is:

$$P_{T\to C}(n|o,C) = \frac{P(C|o,n)P(n|o)}{P(C|o)} \quad (13)$$

$$P(C|o,n) = \sum_z P(z|o,n) \prod_{c\in C} P(c|z) \quad (14)$$

$$P(C|z) = \prod_{c\in C} P(c|z) \quad (15)$$

$$P(z|o,C) = \frac{P(z|o)P(C|z)}{\sum_{z'} P(z'|o)P(C|z')} \quad (16)$$

With appropriate priors chosen for the distributions over words and latent variables, $P(n|o,C)$ is a fully generative model of lexical substitution. A non-generative alternative is one that estimates the similarity of the latent variable distributions associated with seeing n and o in context C. The principle that similarity between topic distributions corresponds to semantic similarity is well-known in document modelling and was proposed in the context of lexical substitution by Dinu and Lapata (2010). In terms of the equations presented above, we could compare the distributions $P(\mathbf{z}|o,C)$ with $P(\mathbf{z}|n,C)$ using equations (5) or (16). However, Thater et al. (2010) and Dinu and Lapata (2010) both observe that contextualising both o and n can degrade performance; in view of this we actually compare $P(\mathbf{z}|o,C)$ with $P(\mathbf{z}|n)$ and make the further simplifying assumption that $P(z|n) \propto P(n|z)$. The similarity measure we adopt is the Bhattacharyya coefficient, which is a natural measure of similarity between probability distributions and is closely related to the Hellinger distance used in previous work on topic modelling (Blei and Lafferty, 2007):

$$sim_{bhatt}(P_x(\mathbf{z}), P_y(\mathbf{z})) = \sum_z \sqrt{P_x(z)P_y(z)} \quad (17)$$

This measure takes values between 0 and 1.

In this paper we train LDA models of $P(w|c)$ and $P(c|w)$. In the former case, the analogy to document modelling is that each context type plays the role of a "document" consisting of all the words observed in that context in a corpus; for $P(c|w)$ the roles are reversed. The models are trained by Gibbs sampling using the efficient procedure of Yao et al. (2009). The empirical estimates for distributions over words and latent variables are derived from the assignment of topics over the training corpus in a single sampling state. For example, to model $P(w|c)$ we calculate:

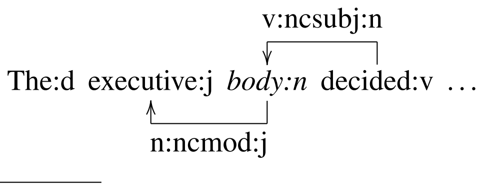
$$P(w|z) = \frac{f_{zw} + \beta}{f_{z\cdot} + N\beta} \tag{18}$$

$$P(z|c) = \frac{f_{zc} + \alpha_z}{f_{\cdot c} + \sum_{z'} \alpha_{z'}} \tag{19}$$

where f_{zw} is the number of words of type w assigned topic z, f_{zc} is the number of times z is associated with context c, $f_{z\cdot}$ and $f_{\cdot c}$ are the marginal topic and context counts respectively, N is the number of word types and α and β parameterise the Dirichlet prior distributions over $P(z|c)$ and $P(w|z)$. Following the recommendations of Wallach et al. (2009) we use asymmetric α and symmetric β; rather than using fixed values for these hyperparameters we estimate them from data in the course of LDA training using an EM-like method.[2] We use standard settings for the number of training iterations (1000), the length of the burnin period before hyperparameter estimation begins (200 iterations) and the frequency of hyperparameter estimation (50 iterations).

3.2 Context types

We have not yet defined what the contexts c look like. In vector space models of semantics it is common to distinguish between window-based and dependency-based models (Padó and Lapata, 2007); one can make the same distinction for probabilistic context models. A broad generalisation is that window-based models capture semantic association (e.g. *referee* is associated with *football*), while dependency models capture a finer-grained notion of similarity (*referee* is similar to *umpire* but not to *football*). Dinu and Lapata (2010) propose a window-based model of lexical substitution; the set of contexts in which a word appears is the set of surrounding words within a prespecified "window size". In this paper we also investigate dependency-based context sets derived from syntactic structure. Given a sentence such as



the set C of dependency contexts for the noun *body* is {*executive:j:ncmod*$^{-1}$*:n*, *decide:v:ncsubj:n*}, where $ncmod^{-1}$ denotes that *body* stands in an inverse non-clausal modifier relation to *executive* (we assume that nouns are the heads of their adjectival modifiers).

4 Experiment 1: Similarity in context

4.1 Data

Mitchell and Lapata (2008) collected human judgements of semantic similarity for pairs of short sentences, where the sentences in a pair share the same subject but different verbs. For example, *the sales slumped* and *the sales declined* should be judged as very similar while *the shoulders slumped* and *the shoulders declined* should be judged as less similar. The resulting dataset (henceforth ML08) consists of 120 such pairs using 15 verbs, balanced across high and low expected similarity. 60 subjects rated the data using a scale of 1–7; Mitchell and Lapata calculate average interannotator correlation to be 0.40 (using Spearman's ρ). Both Mitchell and Lapata and Erk and Padó (2008) split the data into a development portion and a test portion, the development portion consisting of the judgements of six annotators; in order to compare our results with previous research we use the same data split. To evaluate performance, the predictions made by a model are compared to the judgements of each annotator in turn (using ρ) and the resulting per-annotator ρ values are averaged.

4.2 Models

All models were trained on the written section of the British National Corpus (around 90 million words), parsed with RASP (Briscoe et al., 2006). The BNC was also used by Mitchell and Lapata (2008) and Erk and Padó (2008); as the ML08 dataset was compiled using words appearing more than 50 times in the BNC, there are no coverage problems caused by data sparsity. We trained LDA models for the grammatical relations *v:ncsubj:n* and *n:ncsubj*$^{-1}$*:v* and used these to create predictors of type $C \rightarrow T$ and $T \rightarrow C$, respectively. For each predictor, we trained five runs with 100 topics for 1000 iterations and averaged the predictions produced from their final states. We investigate both the generative paraphrasing model (PARA) and the method of comparing topic distributions (SIM). For both PARA and SIM we

[2] We use the estimation methods provided by the MALLET toolkit, available from `http://mallet.cs.umass.edu/`.

	Model	PARA	SIM
No optimisation	C → T	0.24	0.34
	T → C	0.36	**0.39**
	T ↔ C	0.33	0.39
Optimised on dev	C → T	0.24	0.35
	T → C	**0.41**	**0.41**
	T ↔ C	0.37	**0.41**
Erk and Padó (2008)	Mult	0.24	
	SVS	0.27	

Table 1: Performance (average ρ) on the ML08 test set

present results using each predictor type on its own as well as a combination of both types ($T \leftrightarrow C$); for PARA the contributions of the types are multiplied and for SIM they are averaged.[3] One potential complication is that the PARA model is trained to predict $P(n|c,o)$, which might not be comparable across different combinations of subject c and verb o. Using $P(n|c,o)$ as a proxy for the desired joint distribution $P(n,c,o)$ is tantamount to assuming a uniform distribution $P(c,o)$, which can be defended on the basis that the choice of subject noun and reference verb is not directly relevant to the task. As shown by the results below, this assumption seems to work reasonably well in practice.

As well as reporting correlations for straightforward averages of each set of five runs, we also investigate whether the development data can be used to select an optimal subset of runs. This is done by simply evaluating every possible subset of 1–5 runs on the development data and picking the best-scoring subset.

4.3 Results

Table 1 presents the results of the PARA and SIM predictors on the ML08 dataset. The best results previously reported for this dataset were given by Erk and Padó (2008), who measured average ρ values of 0.24 for a vector multiplication method and 0.27 for their *structured vector space* (SVS) syntactic disambiguation method. Even without using the development set to select models, performance is well above the previous state of the art for all predictors except $\text{PARA}_{C \rightarrow T}$. Model selection on the development data brings average ρ up to 0.41, which is comparable to the human "ceiling" of 0.40 measured by Mitchell and Lapata. In all cases the $T \rightarrow C$ predictors outperform $C \rightarrow T$: models that associate target words with distributions over context clusters are superior to those that associate contexts with distributions over target words.

Figure 1 plots the beneficial effect of averaging over multiple runs; as the number of runs n is increased, the average performance over all combinations of n predictors chosen from the set of five $T \rightarrow C$ and five $C \rightarrow T$ runs is observed to increase monotonically. Figure 1 also shows that the model selection procedure is very effective at selecting the optimal combination of models; development set performance is a reliable indicator of test set performance.

5 Experiment 2: Lexical substitution

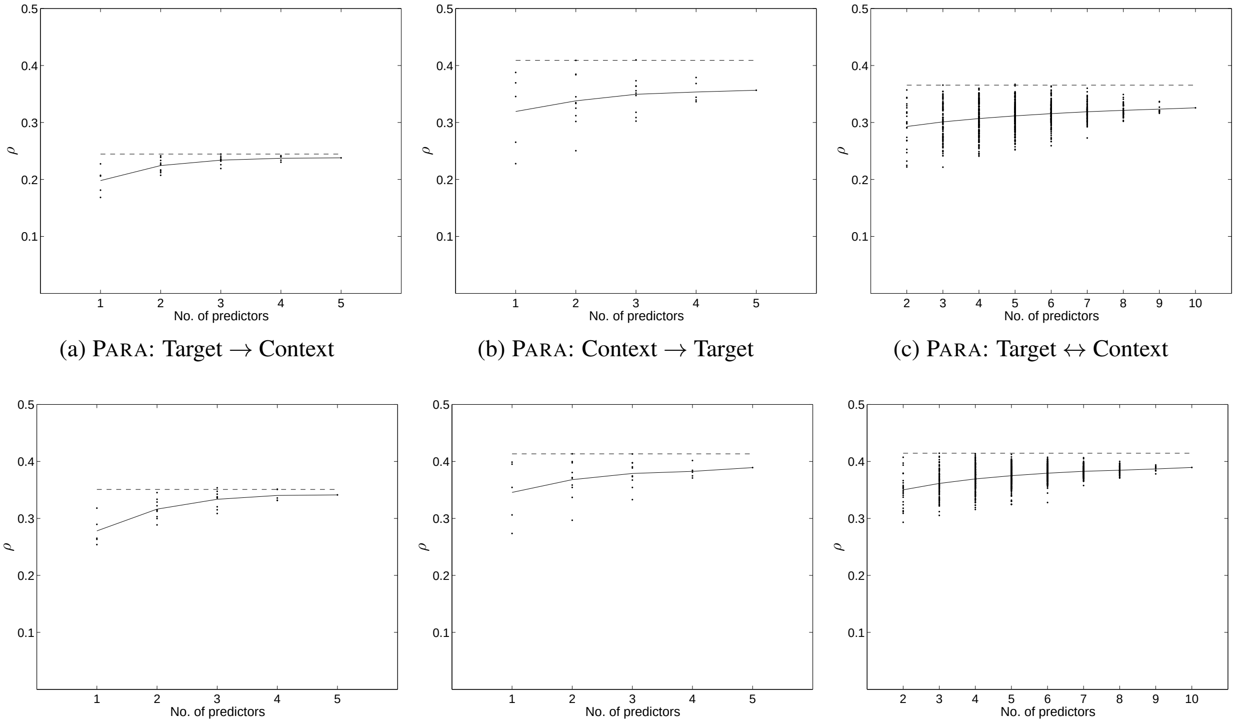
5.1 Data

The English Lexical Substitution task, run as part of the SemEval-1 competition, required participants to propose good substitutes for a set of target words in various sentential contexts (McCarthy and Navigli, 2009). Table 2 shows two example sentences and the substitutes appearing in the gold standard, ranked by the number of human annotators who proposed each substitute. The dataset contains a total of 2,010 annotated sentences with 205 distinct target words across four parts of speech (noun, verb, adjective, adverb). In line with previous work on contextual disambiguation, we focus here on the subtask of ranking attested substitutes rather than proposing them from an unrestricted vocabulary. To this end, a candidate set is constructed for each target word from all the substitutes proposed for that word in all sentences in the dataset.

The data contains a number of multiword paraphrases such as *rush at*; as our models (like most current models of distributional semantics) do not represent multiword expressions, we remove such paraphrases and discard the 17 sentences which have only multiword substitutes in the gold standard.[4] There are also 7 sentences for which the gold stan-

[3] This configuration seems the most intuitive; averaging PARA predictors and multiplying SIM also give good results.

[4] Thater et al. (2010) and Dinu and Lapata (2010) similarly remove multiword paraphrases (Georgiana Dinu, p.c.).



(a) PARA: Target → Context

(b) PARA: Context → Target

(c) PARA: Target ↔ Context

(d) SIM: Target → Context

(e) SIM: Context → Target

(f) SIM: Target ↔ Context

Figure 1: Performance on the ML08 test set with different predictor types and different numbers of LDA runs per predictor type; the solid line tracks the average performance, the dashed line shows the performance of the predictor combination that scores best on the development set.

Realizing immediately that strangers have come, the animals *charge* them and the horses began to fight.	attack (5), rush at (1)
Commission is the amount *charged* to execute a trade.	levy (2), impose (1), take (1), demand (1)

Table 2: Examples for the verb *charge* from the English Lexical Substitution Task

dard contains no substitutes. This leaves a total of 1986 sentences. These sentences were lemmatised and parsed with RASP.

Previous authors have partitioned the dataset in various ways. Erk and Padó (2008) use only a subset of the data where the target is a noun headed by a verb or a verb heading a noun. Thater et al. (2010) discard sentences which their parser cannot parse and paraphrases absent from their training corpus and then optimise the parameters of their model through four-fold cross-validation. Here we aim for complete coverage on the dataset and do not perform any parameter tuning. We use two measures to evaluate performance: Generalised Averaged Precision (Kishida, 2005) and Kendall's τ_b rank correlation coefficient, which were used for this task by Thater et al. (2010) and Dinu and Lapata (2010), respectively. Generalised Averaged Precision (GAP) is a precision-like measure for evaluating ranked predictions against a gold standard. τ_b is a variant of Kendall's τ that is appropriate for data containing tied ranks. We do not use the "precision out of ten" measure that was used in the original Lexical Substitution Task; this measure assigns credit for the proportion of the first 10 proposed paraphrases that are present in the gold standard and in the context of ranking attested substitutes it is unclear how to obtain non-trivial results for target words with 10 or fewer possible substitutes. We calculate statistical significance of performance

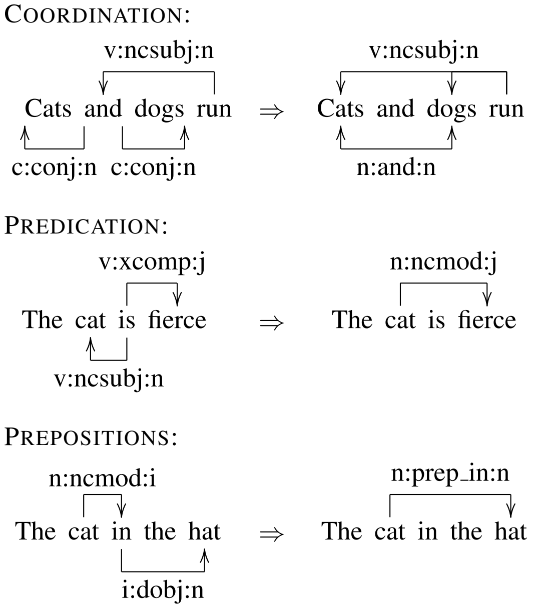


Table 3: Dependency graph preprocessing

differences using stratified shuffling (Yeh, 2000).[5]

5.2 Models

We apply the models developed in Section 3.1 to the Lexical Substitution Task dataset using dependency- and window-based context information. Here we only use the SIM predictor type. PARA did not give satisfactory results; in particular, it tended to rank common words highly in most contexts.[6]

As before we compiled training data by extracting target-context cooccurrences from a text corpus. In addition to the parsed BNC described above we used a corpus of Wikipedia text consisting of over 45 million sentences (almost 1 billion words) parsed using the fast Combinatory Categorial Grammar (CCG) parser described by Clark et al. (2009). The dependency representation produced by this parser is interoperable with the RASP dependency format. In order to focus our models on semantically discriminative information and make inference more tractable we ignored all parts of speech other than nouns, verbs, adjectives, prepositions and adverbs. Stopwords and words of fewer than three characters were removed. We also removed the very frequent but semantically weak lemmas *be* and *have*.

We compare two classes of context models: models learned from window-based contexts and models learned from syntactic dependency contexts. For the syntactic models we extracted all dependencies and inverse dependencies between lemmas of the aforementioned POS types; in order to maximise the extraction yield, the dependency graph for each sentence was preprocessed using the transformations shown in Table 3. For the window-based context model we follow Dinu and Lapata (2010) in treating each word within five words of a target as a member of its context set.

It proved necessary to subsample the corpora in order to make LDA training tractable, especially for the window-based model where the training set of context-target counts is extremely dense (each instance of a word in the corpus contributes up to 10 context instances). For the window-based data, we divided each context-target count by a factor of 5 and a factor of 70 for the BNC and Wikipedia corpora respectively, rounding fractional counts to the closest integer. The choice of 70 for scaling Wikipedia counts is adopted from Dinu and Lapata (2010), who used the same factor for the comparably sized English Gigaword corpus. As the dependency data is an order of magnitude smaller we downsampled the Wikipedia counts by 5 and left the BNC counts untouched. Finally, we created a larger corpus by combining the counts from the BNC and Wikipedia datasets. Type and token counts for the BNC and combined corpora are given in Table 4.

We trained three LDA predictors for each corpus: a window-based predictor (W5), a Context → Target predictor ($C \rightarrow T$) and a Target → Context predictor ($T \rightarrow C$). For W5 the sets of types and contexts should be symmetrical (in practice there is some discrepancy due to preprocessing artefacts). For $C \rightarrow T$, individual models were trained for each of the four target parts of speech; in each case the set of types is the vocabulary for that part of speech and the set of contexts is the set of dependencies taking those types as dependents. For $T \rightarrow C$ we again train four models; the sets of types and contexts are reversed. For the both corpora we trained models with $Z = \{600, 800, 1000, 1200\}$ topics; for each setting of Z we ran five estimation runs. Each individual prediction of similarity between $P(z|C, o)$

[5] We use the software package available at `http://www.nlpado.de/~sebastian/sigf.html`.

[6] Favouring more general words may indeed make sense in some paraphrasing tasks (Nulty and Costello, 2010).

	BNC			BNC+Wikipedia		
	Tokens	Types	Contexts	Tokens	Types	Contexts
Nouns	18723082	122999	316237	54145216	106448	514257
Verbs	7893462	18494	57528	20082658	16673	82580
Adjectives	4385788	73684	37163	11536424	88488	57531
Adverbs	1976837	7124	14867	3017936	4056	18510
Window5	28329238	88265	102792	42828094	139640	143443

Table 4: Type and token counts for the BNC and downsampled BNC+Wikipedia corpora

	BNC			BNC + Wikipedia		
	GAP	τ_b	Coverage	GAP	τ_b	Coverage
W5	44.5	0.17	100.0	44.8	0.17	100.0
$C \to T$	43.2	0.16	86.4	48.7	0.21	86.5
$T \to C$	47.2	0.21	86.4	49.3	0.22	86.5
$T \leftrightarrow C$	45.7	0.20	86.4	49.1	0.23	86.5
W5 + $C \to T$	46.0	0.18	100.0	48.7	0.21	100.0
W5 + $T \to C$	**48.6**	**0.21**	100.0	49.3	0.22	100.0
W5 + $T \leftrightarrow C$	48.1	0.20	100.0	**49.5**	**0.23**	100.0

Table 5: Results on the English Lexical Substitution Task dataset; boldface denotes best performance at full coverage for each corpus

and $P(z|n)$ is made by averaging over the predictions of all runs and over all settings of Z. Choosing a single setting of Z does not degrade performance significantly; however, averaging over settings is a convenient way to avoid having to pick a specific value.

We also investigate combinations of predictor types, once again produced by averaging: we combine $C \to T$ with $C \leftrightarrow T$ ($T \leftrightarrow C$) and combine each of these three models with W5.

5.3 Results

Table 5 presents the results attained by our models on the Lexical Substitution Task data. The dependency-based models have imperfect coverage (86% of the data); they can make no prediction when no syntactic context is provided for a target, perhaps as a result of parsing error. The window-based models have perfect coverage, but score noticeably lower. By combining dependency- and window-based models we can reach high performance with perfect coverage. All combinations outperform the corresponding W5 results to a statistically significant degree ($p < 0.01$). Performance at full coverage is already very good (GAP= 48.6, $\tau_b = 0.21$) on the BNC corpus, but the best results are attained by W5 + $T \leftrightarrow C$ trained on the combined corpus (GAP= 49.5, $\tau_b = 0.23$). The results for the W5 model trained on BNC data is comparable to that trained on the combined corpus; however the syntactic models show a clear benefit from the less sparse dependency data in the combined training corpus.

As remarked in Section 3.1, Dinu and Lapata (2010) use a slightly different formulation of $P(z|C, o)$. Using the window-based context model our formulation (5) outperforms (7) for both training corpora; the Dinu and Lapata (2010) version scores GAP $= 41.5, \tau_b = 0.15$ for the BNC corpus and GAP $= 42.0, \tau_b = 0.15$ for the combined corpus. The advantage of our formulation is statistically significant for all evaluation measures.

Table 6 gives a breakdown of performance by target part of speech for the BNC+Wikipedia-trained W5 and W5 + $T \leftrightarrow C$ models, as well as figures provided by previous researchers.[7] W5 + $T \leftrightarrow C$ outperforms W5 on all parts of speech using both

[7] The overall average GAP for Thater et al. (2010) does not appear in their paper but can be calculated from the score and number of instances listed for each POS.

	Nouns		Verbs		Adjectives		Adverbs		Overall	
	GAP	τ_b	GAP	τ_b	GAP	τ_b	GAP	τ_b	GAP	τ_b
W5	46.0	0.16	38.9	0.14	44.0	0.18	54.0	0.22	44.8	0.17
W5 + $T \leftrightarrow C$	**50.7**	**0.22**	45.1	**0.20**	**48.8**	**0.24**	**55.9**	0.24	**49.5**	**0.23**
Thater et al. (2010) (Model 1)	46.4	–	**45.9**	–	39.4	–	48.2	–	44.6	–
Thater et al. (2010) (Model 2)	42.5	–	–	–	43.2	–	51.4	–	–	–
Dinu and Lapata (2010) (LDA)	–	0.16	–	0.14	–	0.17	–	0.21	–	0.16
Dinu and Lapata (2010) (NMF)	–	0.15	–	0.14	–	0.16	–	**0.26**	–	0.16

Table 6: Performance by part of speech

evaluation metrics. As remarked above, previous researchers have used the corpus in slightly different ways; we believe that the results of Dinu and Lapata (2010) are fully comparable, while those of Thater et al. (2010) were attained on a slightly smaller dataset with parameters set through cross-validation. The results for W5 + $T \leftrightarrow C$ outperform all of Dinu and Lapata's per-POS and overall results except for a slightly superior score on adverbs attained by their NMF model ($\tau_b = 0.26$ compared to 0.24). Turning to Thater et al., we report higher scores for every POS with the exception of the verbs where their Model 1 achieves 45.9 GAP compared to 45.1; the overall average for W5 + $T \leftrightarrow C$ is substantially higher at 49.5 compared to 44.6. On balance, we suggest that our models do have an advantage over the current state of the art for lexical substitution.

6 Conclusion

In this paper we have proposed novel methods for modelling the effect of context on lexical meaning, demonstrating that information about syntactic context and textual proximity can fruitfully be integrated to produce state-of-the-art models of lexical choice. We have demonstrated the effectiveness of our techniques on two datasets but they are potentially applicable to a range of applications where semantic disambiguation is required. In future work, we intend to adapt our approach for word sense disambiguation as well as related domain-specific tasks such as gene name normalisation (Morgan et al., 2008). A further, more speculative direction for future research is to investigate more richly structured models of context, for example capturing correlations between words in a text within a framework similar to the Correlated Topic Model of Blei and Lafferty (2007) or more explicitly modelling polysemy effects as in Reisinger and Mooney (2010).

Acknowledgements

We are grateful to the EMNLP reviewers for their helpful comments. This research was supported by EPSRC grant EP/G051070/1.

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 (EMNLP-10)*, Cambridge, MA.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

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

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the ACL-06 Interactive Presentation Sessions*, Sydney, Australia.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of EACL-09*, Athens, Greece.

Stephen Clark, Ann Copestake, James R. Curran, Yue Zhang, Aurelie Herbelot, James Haggerty, Byung-Gyu Ahn, Curt Van Wyk, Jessika Roesner, Jonathan Kummerfeld, and Tim Dawborn. 2009. Large-scale syntactic processing: Parsing the web. Technical report, Final Report of the 2009 JHU CLSP Workshop.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, Cambridge,MA.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, Honolulu, HI.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2011. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS-11)*, Oxford, UK.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Kenneth E. Harper. 1965. Measurement of similarity between nouns. In *Proceedings of the 1965 International Conference on Computational Linguistics (COLING-65)*, New York, NY.

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

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

Kazuaki Kishida. 2005. Property of average precision and its generalisation: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

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 the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, Columbus, OH.

Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chengjie Sun, Heng hui Liu, Rafael Torres, Michael Krauthammer, William W Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K. Bretonnel Cohen, and Lynette Hirschman. 2008. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2).

Paul Nulty and Fintan Costello. 2010. UCD-PN: Selecting general paraphrases using conditional probability. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2)*, Uppsala, Sweden.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH.

Joseph Reisinger and Raymond Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, Cambridge,MA.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, MD.

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 (ACL-10)*, Uppsala, Sweden.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216.

Karen Spärck Jones. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge.

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 (ACL-10)*, Uppsala, Sweden.

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

Hanna Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of NIPS-09*, Vancouver, BC.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09), Paris, France.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics (COLING-00)*, Saarbrücken, Germany.