Modelling selectional preferences in a lexical hierarchy

Diarmuid Ó Séaghdha
Computer Laboratory
University of Cambridge
Cambridge, UK
do242@cam.ac.uk

Anna Korhonen
Computer Laboratory
University of Cambridge
Cambridge, UK
Anna.Korhonen@cl.cam.ac.uk

Abstract

This paper describes Bayesian selectional preference models that incorporate knowledge from a lexical hierarchy such as WordNet. Inspired by previous work on modelling with WordNet, these approaches are based either on "cutting" the hierarchy at an appropriate level of generalisation or on a "walking" model that selects a path from the root to a leaf. In an evaluation comparing against human plausibility judgements, we show that the models presented here outperform previously proposed comparable WordNet-based models, are competitive with state-of-the-art selectional preference models and are particularly well-suited to estimating plausibility for items that were not seen in training.

1 Introduction

The concept of *selectional preference* captures the intuitive fact that predicates in language have a better semantic "fit" for certain arguments than others. For example, the direct object argument slot of the verb *eat* is more plausibly filled by a type of food (*I ate a pizza*) than by a type of vehicle (*I ate a car*), while the subject slot of the verb *laugh* is more plausibly filled by a person than by a vegetable. Human language users' knowledge about selectional preferences has been implicated in analyses of metaphor processing (Wilks, 1978) and in psycholinguistic studies of comprehension (Rayner et al., 2004). In Natural Language Processing, automatically acquired preference models have been shown to aid a number of tasks, including semantic role labelling (Zapirain et al., 2009), parsing (Zhou et al., 2011) and lexical disambiguation (Thater et al., 2010; Ó Séaghdha and Korhonen, 2011).

It is tempting to assume that with a large enough corpus, preference learning reduces to a simple language modelling task that can be solved by counting predicate-argument co-occurrences. Indeed, Keller and Lapata (2003) show that relatively good performance at plausibility estimation can be attained by submitting queries to a Web search engine. However, there are many scenarios where this approach is insufficient: for languages and language domains where Web-scale data is unavailable, for predicate types (e.g., inference rules or semantic roles) that cannot be retrieved by keyword search and for applications where accurate models of rarer words are required. Ó Séaghdha (2010) shows that the Web-based approach is reliably outperformed by more complex models trained on smaller corpora for less frequent predicate-argument combinations. Models that induce a level of semantic representation, such as probabilistic latent variable models, have a further advantage in that they can provide rich structured information for downstream tasks such as lexical disambiguation (Ó Séaghdha and Korhonen, 2011) and semantic relation mining (Yao et al., 2011).

Recent research has investigated the potential of Bayesian probabilistic models such as Latent Dirichlet Allocation (LDA) for modelling selectional preferences (Ó Séaghdha, 2010; Ritter et al., 2010; Reisinger and Mooney, 2011). These models are flexible and robust, yielding superior performance compared to previous approaches. In this paper we present a preliminary study of analogous

models that make use of a lexical hierarchy (in our case the WordNet hierarchy). We describe two broad classes of probabilistic models over WordNet and how they can be implemented in a Bayesian framework. The two main potential advantages of incorporating WordNet information are: (a) improved predictions about rare and out-of-vocabulary arguments; (b) the ability to perform syntactic word sense disambiguation with a principled probabilistic model and without the need for an additional step that heuristically maps latent variables onto WordNet senses. Focussing here on (a), we demonstrate that our models attain better performance than previously-proposed WordNet-based methods on a plausibility estimation task and are particularly well-suited to estimating plausibility for arguments that were not seen in training and for which LDA cannot make useful predictions.

2 Background and Related Work

The WordNet lexical hierarchy (Fellbaum, 1998) is one of the most-used resources in NLP, finding many applications in both the definition of tasks (e.g. the SENSEVAL/SemEval word sense disambiguation tasks) and in the construction of systems. The idea of using WordNet to define selectional preferences was first implemented by Resnik (1993), who proposed a measure of *associational strength* between a semantic class s and a predicate p corresponding to a relation type r:

$$A(s,p,r) = \frac{1}{\eta} P(s|p,r) \log_2 \frac{P(s|p,r)}{P(s|r)} \quad (1)$$

where η is a normalisation term. This measure captures the degree to which the probability of seeing s given the predicate p differs from the prior probability of s. Given that we are often interested in the preference of p for a word w rather than a class and words generally map onto multiple classes, Resnik suggests calculating $A(s,p,r)$ for all classes that could potentially be expressed by w and predicting the maximal value.

Cut-based models assume that modelling the selectional preference of a predicate involves finding the right "level of generalisation" in the WordNet hierarchy. For example, the direct object slot of the verb *eat* can be associated with the subhierarchy rooted at the synset **food#n#1**, as all hyponyms of that synset are assumed to be edible and the immediate hypernym of the synset, **substance#n#1**, is too general given that many substances are rarely eaten.[1] This leads to the notion of "cutting" the hierarchy at one or more positions (Li and Abe, 1998). The modelling task then becomes that of finding the cuts that are maximally general without overgeneralising. Li and Abe (1998) propose a model in which the appropriate cut c is selected according to the Minimum Description Length principle; this principle explicitly accounts for the trade-off between generalisation and accuracy by minimising a sum of *model description length* and *data description length*. The probability of a predicate p taking as its argument an synset s is modelled as:

$$P_{la}(s|p,r) = P(s|c_{s,p,r})P(c|p) \quad (2)$$

where $c_{s,p,r}$ is the portion of the cut learned for p that dominates s. The distribution $P(s|c_{s,p,r})$ is held to be uniform over all synsets dominated by $c_{s,p,r}$, while $P(c|p)$ is given by a maximum likelihood estimate.

Clark and Weir (2002) present a model that, while not explicitly described as cut-based, likewise seeks to find the right level of generalisation for an observation. In this case, the hypernym at which to "cut" is chosen by a chi-squared test: if the aggregate preference of p for classes in the subhierarchy rooted at c differs significantly from the individual preferences of p for the immediate children of c, the hierarchy is cut below c. The probability of p taking a synset s as its argument is given by:

$$P_{cw}(s|p,r) = \frac{P(p|c_{s,p,r},r)\frac{P(s|r)}{P(p|r)}}{\sum_{s' \in S} P(p|c_{s',p,r},r)\frac{P(s'|r)}{P(p|r)}} \quad (3)$$

where $c_{s,p,r}$ is the root node of the subhierarchy containing s that was selected for p.

An alternative approach to modelling with WordNet uses its hierarchical structure to define a Markov model with transitions from senses to senses and from senses to words. The intuition here is that each observation is generated by a "walk" from the root of the hierarchy to a leaf node and emitting the word

[1] In this paper we use WordNet version 3.0, except where stated otherwise.

corresponding to the leaf. Abney and Light (1999) proposed such a model for selectional preferences, trained via EM, but failed to achieve competitive performance on a pseudodisambiguation task.

The models described above have subsequently been used in many different studies. For example: McCarthy and Carroll (2003) use Li and Abe's method in a word sense disambiguation setting; Schulte im Walde et al. (2008) use their MDL approach as part of a system for syntactic and semantic subcategorisation frame learning; Shutova (2010) deploys Resnik's method for metaphor interpretation. Brockmann and Lapata (2003) report a comparative evaluation in which the methods of Resnik and Clark and Weir outpeform Li and Abe's method on a plausibility estimation task.

Much recent work on preference learning has focused on purely distributional methods that do not use a predefined hierarchy but learn to make generalisations about predicates and arguments from corpus observations alone. These methods can be vector-based (Erk et al., 2010; Thater et al., 2010), discriminative (Bergsma et al., 2008) or probabilistic (Ó Séaghdha, 2010; Ritter et al., 2010; Reisinger and Mooney, 2011). In the probabilistic category, Bayesian models based on the "topic modelling" framework (Blei et al., 2003b) have been shown to achieve state-of-the-art performance in a number of evaluation settings; the models considered in this paper are also related to this framework.

In machine learning, researchers have proposed a variety of topic modelling methods where the latent variables are arranged in a hierarchical structure (Blei et al., 2003a; Mimno et al., 2007). In contrast to the present work, these models use a relatively shallow hierarchy (e.g., 3 levels) and any hierarchy node can in principle emit any vocabulary item; they thus provide a poor match for our goal of modelling over WordNet. Boyd-Graber et al. (2007) describe a topic model that is directly influenced by Abney and Light's Markov model approach; this model (LDAWN) is described further in Section 3.3 below. Reisinger and Paşca (2009) investigate Bayesian methods for attaching attributes harvested from the Web at an appropriate level in the WordNet hierarchy; this task is related in spirit to the preference learning task.

3 Probabilistic modelling over WordNet

3.1 Notation

We assume that we have a lexical hierarchy in the form of a directed acyclic graph $G = (S, E)$ where each node (or *synset*) $s \in S$ is associated with a set of words W_n belonging to a large vocabulary V. Each edge $e \in E$ leads from a node n to its children (or *hyponyms*) Ch_n. As G is a DAG, a node may have more than one parent but there are no cycles. The ultimate goal is to learn a distribution over the argument vocabulary V for each predicate p in a set P, through observing predicate-argument pairs. A predicate is understood to correspond to a pairing of a lexical item v and a relation type r, for example $p = (eat, direct\_object)$. The list of observations for a predicate p is denoted by $Observations(p)$.

3.2 Cut-based models

Model 1 Generative story for WN-CUT

for cut $c \in \{1 \dots |C|\}$ **do**
  $\Phi_c \sim Multinomial(\beta_c)$
end for
for predicate $p \in \{1 \dots |P|\}$ **do**
  $\theta_p \sim Dirichlet(\alpha)$
  for argument instance $i \in Observations(p)$ **do**
    $c_i \sim Multinomial(\theta_p)$
    $w_i \sim Multinomial(\Phi_{c_i})$
  end for
end for

The first model we consider, WN-CUT, is directly inspired by Li and Abe's model (2). Each predicate p is associated with a distribution over "cuts", i.e., complete subgraphs of G rooted at a single node and containing all nodes dominated by the root. It follows that the number of possible cuts is the same as the number of synsets. Each cut c is associated with a non-uniform distribution over the set of words W_c that can be generated by the synsets contained in c. As well as the use of a non-uniform emission distribution and the placing of Dirichlet priors on the multinomial over cuts, a significant difference from Li and Abe's model is that overlapping cuts are permitted (indeed, every cut has non-zero probability given a predicate). For example, the

model may learn that the direct object slot of *eat* gives high probability to the cut rooted at **food#n#1** but also that the cut rooted at **substance#n#1** has non-negligible probability, higher than that assigned to **phenomenon#n#1**. It follows that the estimated probability of p taking argument w takes into account all possible cuts, weighted by their probability given p.

The generative story for WN-CUT is given in Algorithm 1; this describes the assumptions made by the model about how a corpus of observations is generated. The probability of predicate p taking argument w is defined as (4); an empirical posterior estimate of this quantity can be computed from a Gibbs sampling state via (5):

$$P(w|p) = \sum_c P(c|p)P(w|c) \tag{4}$$

$$\propto \sum_c \frac{f_{cp} + \alpha}{f_{\cdot p} + |C|\alpha} \frac{f_{wc} + \beta}{f_{\cdot c} + |W_c|\beta} \tag{5}$$

where f_{cw} is the number of observations containing argument w that have been assigned cut c, f_{cp} is the number of observations containing predicate p that have been assigned cut c and $f_{c\cdot}$, $f_{\cdot p}$ are the marginal counts for cut c and predicate p, respectively. The two terms that are multiplied in (4) play complementary roles analogous to those of the two description lengths in Li and Abe's MDL formulation; $P(c|p)$ will prefer to reuse more general cuts, while $P(w|c)$ will prefer more specific cuts with a smaller associated argument vocabulary.

As the number of words $|W_c|$ that can be emitted by a cut $|c|$ varies according to the size of the subhierarchy under the cut, the proportion of probability mass accorded to the likelihood and the prior in (5) is not constant. An alternative formulation is to keep the distribution of mass between likelihood and prior constant but vary the value of the individual β_c parameters according to cut size. Experiments suggest that this alternative does not differ in performance.

The second cut-based model, WN-CUT-TOPICS, extends WN-CUT by adding two extra layers of latent variables. Firstly, the choice of cut is conditional on a "topic" variable z rather than directly conditioned on the predicate; when the topic vocabulary Z is much smaller than the cut vocabulary C, this has the effect of clustering the cuts. Secondly, instead of immediately drawing a word once a cut has been chosen, the model first draws a synset s and then draws a word from the vocabulary W_s associated with that synset. This has two advantages; it directly disambiguates each observation to a specific synset rather than to a region of the hierarchy and it should also improve plausibility predictions for rare synonyms of common arguments. The generative story for WN-CUT-TOPICS is given in Algorithm 2; the distribution over arguments for p is given in (6) and the corresponding posterior estimate in (7):

Model 2 Generative story for WN-CUT-TOPICS

for topic $z \in \{1 \ldots |Z|\}$ **do**
  $\Psi_z \sim Dirichlet(\alpha)$
end for
for cut $c \in \{1 \ldots |C|\}$ **do**
  $\Phi_c \sim Dirichlet(\gamma_c)$
end for
for synset $s \in \{1 \ldots |S|\}$ **do**
  $\Xi_s \sim Dirichlet(\beta_s)$
end for
for predicate $p \in \{1 \ldots |P|\}$ **do**
  $\theta_p \sim Dirichlet(\kappa)$
  for argument instance $i \in Observations(p)$ **do**
    $z_i \sim Multinomial(\theta_p)$
    $c_i \sim Multinomial(\Psi_z)$
    $s_i \sim Multinomial(\Phi_c)$
    $w_i \sim Multinomial(\Xi_s)$
  end for
end for

$$P(w|p) = \sum_z P(z|p) \sum_c P(c|z) \sum_s P(s|c)P(w|s) \tag{6}$$

$$\propto \sum_z \frac{f_{zp} + \kappa_z}{f_{\cdot p} + \sum_{z'} \kappa_{z'}} \sum_c \frac{f_{cz} + \alpha}{f_{\cdot z} + |C|\alpha} \times \sum_s \frac{f_{sc} + \gamma}{f_{\cdot c} + |S_c|\gamma} \frac{f_{ws} + \beta}{f_{\cdot s} + |W_s|\beta} \tag{7}$$

As before, f_{zp}, f_{cz}, f_{sc} and f_{ws} are the respective co-occurrence counts of topics/predicates, cuts/topics, synsets/cuts and words/synsets in the sampling state and $f_{\cdot p}$, $f_{\cdot z}$, $f_{\cdot c}$ and $f_{\cdot s}$ are the corresponding marginal counts.

Since WN-CUT and WN-CUT-TOPICS are constructed from multinomials with Dirichlet priors, it is relatively straightforward to train them by collapsed Gibbs sampling (Griffiths and Steyvers, 2004), an iterative method whereby each latent variable in the model is stochastically updated according to the distribution given by conditioning on the current latent variable assignments of all other tokens. In the case of WN-CUT, this amounts to updating the cut assignment c_i for each token in turn. For WN-CUT-TOPICS there are three variables to update; c_i and s_i must be updated simultaneously, but z_i can be updated independently for the benefit of efficiency. Although WordNet contains 82,115 noun synsets, updates for c_i and s_i can be computed very efficiently, as there are typically few possible synsets for a given word type and few possible cuts for a given synset (the maximum synset depth is 19).

The hyperparameters for the various Dirichlet priors are also reestimated in the course of learning; the values of these hyperparameters control the degree of sparsity preferred by the model. The "top-level" hyperparameters α in WN-CUT and κ in WN-CUT-TOPICS are estimated using a fixed-point iteration proposed by Wallach (2008); the other hyperparameters are learned by slice sampling (Neal, 2003).

3.3 Walk-based models

Abney and Light (1999) proposed an approach to selectional preference learning in which arguments are generated for predicates by following a path $\lambda = (l_1, \ldots, l_{|\lambda|})$ from the root of the hierarchy to a leaf node and emitting the corresponding word. The path is chosen according to a Markov model with transition probabilities specific to each predicate. In this model, each leaf node is associated with a single word; the synsets associated with that word are the immediate parent nodes of the leaf. Abney and Light found that their model did not match the performance of Resnik's (1993) simpler method. We have had a similar lack of success with a Bayesian version of this model, which we do not describe further here.

Boyd-Graber et al. (2007) describe a related topic model, LDAWN, for word sense disambiguation that adds an intermediate layer of latent variables Z on which the Markov model parameters are conditioned. In their application, each document in a corpus is associated with a distribution over topics and each topic is associated with a distribution over paths. The clustering effect of the topic layer allows the documents to "share" information and hence alleviate problems due to sparsity. By analogy to Abney and Light, it is a short and intuitive step to apply LDAWN to selectional preference learning. The generative story for LDAWN is given in Algorithm 3; the probability model for $P(w|p)$ is defined by (8) and the posterior estimate is (9):

Model 3 Generative story for LDAWN

for topic $z \in \{1 \ldots |Z|\}$ **do**
  for synset $s \in \{1 \ldots |S|\}$ **do**
    Draw transition probabilities $\Psi_{z,s} \sim Dirichlet(\sigma\alpha_s)$
  end for
end for
for predicate $p \in \{1 \ldots |P|\}$ **do**
  $\theta_p \sim Dirichlet(\kappa)$
  for argument instance $i \in Observations(p)$ **do**
    $z_i \sim Multinomial(\theta_p)$
    Create a path starting at the root synset λ_0:
    while not at a leaf node **do**
      $\lambda_{t+1} \sim Multinomial(\Psi_{z_i,\lambda_t})$
    end while
    Emit the word at the leaf as w_i
  end for
end for

$$P(w|p) = \sum_z P(z|p) \sum_\lambda \mathbb{1}[\lambda \to w] P(\lambda|z) \quad (8)$$

$$\propto \sum_z \frac{f_{zp} + \kappa_z}{f_{\cdot p} + \sum_{z'} \kappa_{z'}} \sum_\lambda \mathbb{1}[\lambda \to w] \times \prod_{i=1}^{|\lambda|-1} \frac{f_{z,l_i \to l_{i+1}} + \sigma\alpha_{l_i \to l_{i+1}}}{f_{z,l_i \to \cdot} + \sigma} \quad (9)$$

where $\mathbb{1}[\lambda \to w] = 1$ when the path λ leads to leaf node w and has value 0 otherwise. Following Boyd-Graber et al. the Dirichlet priors on the transition probabilities are parameterised by the product of a strength parameter σ and a distribution α_s, the latter being fixed according to relative corpus frequencies to "guide" the model towards more fruitful paths.

Gibbs sampling updates for LDAWN are given in Boyd-Graber et al. (2007). As before, we reestimate

SEEN:	
staff morale	0.4889
team morale	0.5945
issue morale	0.0595
UNSEEN:	
pupil morale	0.4318
minute morale	-0.0352
snow morale	-0.2748

Table 1: Extract from the noun-noun section of Keller and Lapata's (2003) dataset, with human plausibility scores

the hyperparameters during learning; κ is estimated by Wallach's fixed-point iteration and σ is estimated by slice sampling.

4 Experiments

4.1 Experimental procedure

We evaluate our methods by comparing their predictions to human judgements of predicate-argument plausibility. This is a standard approach to selectional preference evaluation (Keller and Lapata, 2003; Brockmann and Lapata, 2003; Ó Séaghdha, 2010) and arguably yields a better appraisal of a model's intrinsic semantic quality than other evaluations such as pseudo-disambiguation or held-out likelihood prediction.[2] We use a set of plausibility judgements collected by Keller and Lapata (2003). This dataset comprises 180 predicate-argument combinations for each of three syntactic relations: verb-object, noun-noun modification and adjective-noun modification. The data for each relation is divided into a "seen" portion containing 90 combinations that were observed in the British National Corpus and an "unseen" portion containing 90 combinations that do not appear (though the predicates and arguments do appear separately). Plausibility judgements were elicited from a large group of human subjects, then normalised and log-transformed. Table 1 gives a representative illustration of the data. Following the evaluation in Ó Séaghdha (2010), with which we wish to compare, we use Pearson r and Spearman ρ correlation coefficients as performance measures.

All models were trained on the 90-million word written component of the British National Corpus,[3] lemmatised, POS-tagged and parsed with the RASP toolkit (Briscoe et al., 2006). We removed predicates occurring with just one argument type and all tokens containing non-alphabetic characters. The resulting datasets consist of 3,587,172 verb-object observations (7,954 predicate types, 80,107 argument types), 3,732,470 noun-noun observations (68,303 predicate types, 105,425 argument types) and 3,843,346 adjective-noun observations (29,975 predicate types, 62,595 argument types).

All the Bayesian models were trained by Gibbs sampling, as outlined above. For each model we run three sampling chains for 1,000 iterations and average the plausibility predictions for each to produce a final prediction $P(w|p)$ for each predicate-argument item. As the evaluation demands an estimate of the joint probability $P(w, p)$ we multiply the predicted $P(w|p)$ by a predicate probability $P(p|r)$ estimated from relative corpus frequencies. In training we use a burn-in period of 200 iterations, after which hyperparameters are reestimated and $P(p|r)$ predictions are sampled every 50 iterations. All probability estimates are log-transformed to match the gold standard judgements.

In order to compare against previously proposed selectional preference approaches based on WordNet we also reimplemented the methods that performed best in the evaluation of Brockmann and Lapata (2003): Resnik (1993) and Clark and Weir (2002). For Resnik's model we used WordNet 2.1 rather than WordNet 3.0 as the former has multiple roots, a property that turns out to be necessary for good performance. Clark and Weir's method requires that the user specify a significance threshold α to be used in deciding where to cut; to give it the best possible chance we tested with a range of values $(0.05, 0.3, 0.6, 0.9)$ and report results for the best-performing setting, which consistently was $\alpha = 0.9$. One can also use different statistical hypothesis tests; again we choose the test giving the best results, which was Pearson's chi-squared test. As this method produces a probability estimate conditioned on the predicate p we multiply by a MLE estimate of $P(p|r)$ and log-transform the result.

[2] For a related argument in the context of topic model evaluation, see Chang et al. (2009).

[3] http://www.natcorp.ox.ac.uk/

eat	**food#n#1, aliment#n#1, entity#n#1, solid#n#1, food#n#2**
drink	**fluid#n#1, liquid#n#1, entity#n#1, alcohol#n#1, beverage#n#1**
appoint	**individual#n#1, entity#n#1, chief#n#1, being#n#2, expert#n#1**
publish	**abstract_entity#n#1, piece_of_writing#n#1, communication#n#2, publication#n#1**

Table 2: Most probable cuts learned by WN-CUT for the object argument of selected verbs

	Verb-object				Noun-noun				Adjective-noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
WN-CUT	.593	.582	.514	.571	.550	.584	.564	.590	.561	.618	.453	.439
WN-CUT-100	.500	.529	**.575**	**.630**	**.619**	.639	**.662**	**.706**	.537	.510	.464	.431
WN-CUT-200	.538	.546	.557	.608	.595	.632	.639	.669	.585	.587	.435	.431
LDAWN-100	.497	.538	.558	.594	.605	.619	.635	.633	.549	.545	.459	.462
LDAWN-200	.546	.562	.508	.548	.610	**.654**	.526	.568	.578	.583	.453	.450
Resnik	.384	.473	.469	.470	.242	.187	.152	.037	.309	.388	.311	.280
Clark/Weir	.489	.546	.312	.365	.441	.521	.543	.576	.440	.476	.271	.242
BNC (MLE)	**.620**	**.614**	.196	.222	.544	.604	.114	.125	.543	**.622**	.135	.102
LDA	.504	.541	.558	.603	.615	.641	.636	.666	**.594**	.558	**.468**	**.459**

Table 3: Results (Pearson r and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data; underlining denotes the best-performing WordNet-based model, boldface denotes the overall best performance

4.2 Results

Table 2 demonstrates the top cuts learned by the WN-CUT model from the verb-object training data for a selection of verbs. Table 3 gives quantitative results for the WordNet-based models under consideration, as well as results reported by Ó Séaghdha (2010) for a purely distributional LDA model with 100 topics and a Maximum Likelihood Estimate model learned from the BNC. In general, the Bayesian WordNet-based models outperform the models of Resnik and Clark and Weir, and are competitive with the state-of-the-art LDA results. To test the statistical significance of performance differences we use the test proposed by Meng et al. (1992) for comparing correlated correlations, i.e., correlation scores with a shared gold standard. The differences between Bayesian WordNet models are not significant ($p > 0.05$, two-tailed) for any dataset or evaluation measure. However, all Bayesian models improve significantly over Resnik's and Clark and Weir's models for multiple conditions. Perhaps surprisingly, the relatively simple WN-CUT model scores the greatest number of significant improvements over both Resnik (7 out of 12 conditions) and Clark and Weir (8 out of 12), though the other Bayesian models do follow close behind. This may suggest that the incorporation of WordNet structure into the model in itself provides much of the clustering benefit provided by an additional layer of "topic" latent variables.[4]

In order to test the ability of the WordNet-based models to make predictions about arguments that are absent from the training vocabulary, we created an artificial out-of-vocabulary dataset by removing each of the Keller and Lapata argument words from the input corpus and retraining. An LDA selectional preference model will completely fail here, but we hope that the WordNet models can still make relatively accurate predictions by leveraging the additional lexical knowledge provided by the hierarchy. For example, if one knows that a tomatillo is classed as a vegetable in WordNet, one can predict a relatively high probability that it can be eaten, even though the word *tomatillo* does not appear in the BNC.

As a baseline we use a BNC-trained model that

[4] An alternative hypothesis is that samplers for the more complex models take longer to "mix". We have run some experiments with 5,000 iterations but did not observe an improvement in performance.

	Verb-object				Noun-noun				Adjective-noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
WN-CUT	**.334**	.326	**.518**	**.569**	.252	.212	.254	.274	**.451**	**.397**	**.471**	**.458**
WN-CUT-100	.308	**.357**	.459	.489	.223	.207	.126	.074	.285	.264	.234	.226
WN-CUT-200	.273	.321	.452	.482	.192	.174	.115	.053	.266	.212	.220	.214
LDAWN-100	.223	.235	.410	.391	**.259**	**.220**	.132	.138	.016	.037	.264	.254
LDAWN-200	.291	.285	.392	.379	.240	.163	.118	.131	.041	.078	.209	.212
Resnik	.203	.341	.472	.497	.054	-.054	.184	.089	.353	.393	.333	.365
Clark/Weir	.222	.287	.201	.235	.225	.162	**.279**	**.304**	.313	.202	.190	.148
BNC	.206	.224	.276	.240	.256	.240	.223	.225	.088	.103	.220	.231

Table 4: Forced-OOV results (Pearson r and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data

predicts $P(w, p)$ proportional to the MLE predicate probability $P(p)$; a distributional LDA model will make essentially the same prediction. Clark and Weir's method does not have full coverage; if no sense s of an argument appears in the data then $P(s|p)$ is zero for all senses and the resulting prediction is zero, which cannot be log-transformed. To sidestep this issue, unseen senses are assigned a pseudofrequency of 0.1. Results for this "forced-OOV" task are presented in Table 4. WN-CUT proves the most adept at generalising to unseen arguments, attaining the best performance on 7 of 12 dataset/evaluation conditions and a statistically significant improvement over the baseline on 6. We observe that estimating the plausibility of unseen arguments for noun-noun modifiers is particularly difficult. One obvious explanation is that the training data for this relation has fewer tokens per predicate, making it more difficult to learn their preferences. A second, more hypothetical, explanation is that the ontological structure of WordNet is a relatively poor fit for the preferences of nominal modifiers; it is well-known that almost any pair of nouns can combine to produce a minimally plausible noun-noun compound (Downing, 1977) and it may be that this behaviour is ill-suited by the assumption that preferences are sparse distributions over regions of WordNet.

5 Conclusion

In this paper we have presented a range of Bayesian selectional preference models that incorporate knowledge about the structure of a lexical hierarchy. One motivation for this work was to test the hypothesis that such knowledge can be helpful in constructing robust models that can handle rare and unseen arguments. To this end we have reported a plausibility-based evaluation in which our models outperform previously proposed WordNet-based preference models and make sensible predictions for out-of-vocabulary items. A second motivation, which we intend to explore in future work, is to apply our models in the context of a word sense disambiguation task. Previous studies have demonstrated the effectiveness of distributional Bayesian selectional preference models for predicting lexical substitutes (Ó Séaghdha and Korhonen, 2011) but these models lack a principled way to map a word onto its most likely WordNet sense. The methods presented in this paper offer a promising solution to this issue. Another potential research direction is integration of semantic relation extraction algorithms with WordNet or other lexical resources, along the lines of Pennacchiotti and Pantel (2006) and Van Durme et al. (2009).

Acknowledgements

The work in this paper was funded by the EPSRC (UK) grant EP/G051070/1, EU grant 7FP-ITC-248064 and the Royal Society, (UK).

References

Steven Abney and Marc Light. 1999. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the ACL-99 Workshop on Unsupervised Learning in NLP*, College Park, MD.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preferences from unlabeled text. In *Proceedings of EMNLP-08*, Honolulu, HI.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested Chinese Restaurant Process. In *Proceedings of NIPS-03*, Vancouver, BC.

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

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of EMNLP-CoNLL-07*, Prague, Czech Republic.

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.

Carsten Brockmann and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of EACL-03*, Budapest, Hungary.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of NIPS-09*, Vancouver, BC.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2), 187–206.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Xiao-Li Meng, Robert Rosenthal, and Donald B. Rubin. 1992. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172–175.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. In *Proceedings of ICML-07*, Corvallis, OR.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31(3):705–767.

Diarmuid Ó Séaghdha and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL-10*, Uppsala, Sweden.

Marco Pennacchiotti and Patrick Pantel. 2006. Ontologizing semantic relations. In *Proceedings of COLING-ACL-06*, Sydney, Australia.

Keith Rayner, Tessa Warren, Barbara J. Juhasz, and Simon P. Liversedge. 2004. The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30(6):1290–1301.

Joseph Reisinger and Raymond Mooney. 2011. Cross-cutting models of lexical semantics. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Joseph Reisinger and Marius Paşca. 2009. Latent variable models of concept-attribute attachment. In *Proceedings of ACL-IJCNLP-09*, Suntec, Singapore.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings ACL-10*, Uppsala, Sweden.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of ACL-08:HLT*, Columbus, OH.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL-HLT-10*, Los Angeles, CA.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of ACL-10*, Uppsala, Sweden.

Benjamin Van Durme, Philip Michalak, and Lenhart K. Schubert. 2009. Deriving generalized knowledge from corpora using WordNet abstraction. In *Proceedings of EACL-09*, Athens, Greece.

Hanna Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11:197–225.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using

generative models. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of ACL-IJCNLP-09*, Singapore.

Guangyou Zhou, Jun Zhao, Kang Liu, and Li Cai. 2011. Exploiting web-derived selectional preference to improve statistical dependency parsing. In *Proceedings of ACL-11*, Portland, OR.