Latent variable models of selectional preference

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

Abstract

This paper describes the application of so-called *topic models* to selectional preference induction. Three models related to Latent Dirichlet Allocation, a proven method for modelling document-word co-occurrences, are presented and evaluated on datasets of human plausibility judgements. Compared to previously proposed techniques, these models perform very competitively, especially for infrequent predicate-argument combinations where they exceed the quality of Web-scale predictions while using relatively little data.

1 Introduction

Language researchers have long been aware that many words place semantic restrictions on the words with which they can co-occur in a syntactic relationship. Violations of these restrictions make the sense of a sentence odd or implausible:

(1) Colourless green ideas sleep furiously.

(2) The deer shot the hunter.

Recognising whether or not a selectional restriction is satisfied can be an important trigger for metaphorical interpretations (Wilks, 1978) and also plays a role in the time course of human sentence processing (Rayner et al., 2004). A more relaxed notion of *selectional preference* captures the idea that certain classes of entities are more likely than others to fill a given argument slot of a predicate. In Natural Language Processing, knowledge about probable, less probable and wholly infelicitous predicate-argument pairs is of value for numerous applications, for example semantic role labelling (Gildea and Jurafsky, 2002; Zapirain et al., 2009). The notion of selectional preference is not restricted to surface-level predicates such as verbs and modifiers, but also extends to semantic frames (Erk, 2007) and inference rules (Pantel et al., 2007).

The fundamental problem that selectional preference models must address is data sparsity: in many cases insufficient corpus data is available to reliably measure the plausibility of a predicate-argument pair by counting its observed frequency. A rarely seen pair may be fundamentally implausible (*a carrot laughed*) or plausible but rarely expressed (*a manservant laughed*).[1] In general, it is beneficial to smooth plausibility estimates by integrating knowledge about the frequency of other, similar predicate-argument pairs. The task thus share some of the nature of language modelling; however, it is a task less amenable to approaches that require very large training corpora and one where the semantic quality of a model is of greater importance.

This paper takes up tools ("topic models") that have been proven successful in modelling document-word co-occurrences and adapts them to the task of selectional preference learning. Advantages of these models include a well-defined generative model that handles sparse data well, the ability to jointly induce semantic classes and predicate-specific distributions over those classes, and the enhanced statistical strength achieved by sharing knowledge across predicates. Section 2 surveys prior work on selectional preference modelling and on semantic applications of topic models. Section 3 describes the models used in our experiments. Section 4 provides details of the experimental design. Section 5 presents results for our models on the task of predicting human plausibility judgements for predicate-argument combinations; we show that performance is generally competi-

[1] At time of writing, Google estimates 855 hits for "*a|the carrot|carrots laugh|laughs|laughed*" and 0 hits for "*a|the manservant|manservants|menservants laugh|laughs|laughed*"; many of the *carrot* hits are false positives but a significant number are true subject-verb observations.

tive with or superior to a number of other models, including models using Web-scale resources, especially for low-frequency examples. In Section 6 we wrap up by summarising the paper's conclusions and sketching directions for future research.

2 Related work

2.1 Selectional preference learning

The representation (and latterly, learning) of selectional preferences for verbs and other predicates has long been considered a fundamental problem in computational semantics (Resnik, 1993). Many approaches to the problem use lexical taxonomies such as WordNet to identify the semantic classes that typically fill a particular argument slot for a predicate (Resnik, 1993; Clark and Weir, 2002; Schulte im Walde et al., 2008). In this paper, however, we focus on methods that do not assume the availability of a comprehensive taxonomy but rather induce semantic classes automatically from a corpus of text. Such methods are more generally applicable, for example in domains or languages where handbuilt semantic lexicons have insufficient coverage or are non-existent.

Rooth et al. (1999) introduced a model of selectional preference induction that casts the problem in a probabilistic latent-variable framework. In Rooth et al.'s model each observed predicate-argument pair is probabilistically generated from a latent variable, which is itself generated from an underlying distribution on variables. The use of latent variables, which correspond to coherent clusters of predicate-argument interactions, allow probabilities to be assigned to predicate-argument pairs which have not previously been observed by the model. The discovery of these predicate-argument clusters and the estimation of distributions on latent and observed variables are performed simultaneously via an Expectation Maximisation procedure. The work presented in this paper is inspired by Rooth et al.'s latent variable approach, most directly in the model described in Section 3.3. Erk (2007) and Padó et al. (2007) describe a corpus-driven smoothing model which is not probabilistic in nature but relies on similarity estimates from a "semantic space" model that identifies semantic similarity with closeness in a vector space of co-occurrences. Bergsma et al. (2008) suggest learning selectional preferences in a discriminative way, by training a collection of SVM classifiers to recognise likely and unlikely arguments for predicates of interest.

Keller and Lapata (2003) suggest a simple alternative to smoothing-based approaches. They demonstrate that noisy counts from a Web search engine can yield estimates of plausibility for predicate-argument pairs that are superior to models learned from a smaller parsed corpus. The assumption inherent in this approach is that given sufficient text, all plausible predicate-argument pairs will be observed with frequency roughly correlated with their degree of plausibility. While the model is undeniably straightforward and powerful, it has a number of drawbacks: it presupposes an extremely large corpus, the like of which will only be available for a small number of domains and languages, and it is only suitable for relations that are identifiable by searching raw text for specific lexical patterns.

2.2 Topic modelling

The task of inducing coherent semantic clusters is common to many research areas. In the field of document modelling, a class of methods known as "topic models" have become a de facto standard for identifying semantic structure in documents. These include the Latent Dirichlet Allocation (LDA) model of Blei et al. (2003) and the Hierarchical Dirichlet Process model of Teh et al. (2006). Formally seen, these are hierarchical Bayesian models which induce a set of latent variables or topics that are shared across documents. The combination of a well-defined probabilistic model and Gibbs sampling procedure for estimation guarantee (eventual) convergence and the avoidance of degenerate solutions. As a result of intensive research in recent years, the behaviour of topic models is well-understood and computationally efficient implementations have been developed. The tools provided by this research are used in this paper as the building blocks of our selectional preference models.

Hierarchical Bayesian modelling has recently gained notable popularity in many core areas of natural language processing, from morphological segmentation (Goldwater et al., 2009) to opinion modelling (Lin et al., 2006). Yet so far there have been relatively few applications to traditional lexical semantic tasks. Boyd-Graber et al. (2007) integrate a model of random walks on the WordNet graph into an LDA topic model to build an unsupervised word sense disambiguation system. Brody

and Lapata (2009) adapt the basic LDA model for application to unsupervised word sense induction; in this context, the topics learned by the model are assumed to correspond to distinct senses of a particular lemma. Zhang et al. (2009) are also concerned with inducing multiple senses for a particular term; here the goal is to identify distinct entity types in the output of a pattern-based entity set discovery system. Reisinger and Paşca (2009) use LDA-like models to map automatically acquired attribute sets onto the WordNet hierarchy. Griffiths et al. (2007) demonstrate that topic models learned from document-word co-occurrences are good predictors of semantic association judgements by humans.

Simultaneously to this work, Ritter et al. (2010) have also investigated the use of topic models for selectional preference learning. Their goal is slightly different to ours in that they wish to model the probability of a binary predicate taking two specified arguments, i.e., $P(n_1, n_2|v)$, whereas we model the joint and conditional probabilities of a predicate taking a single specified argument. The model architecture they propose, LinkLDA, falls somewhere between our LDA and DUAL-LDA models. Hence LinkLDA could be adapted to estimate $P(n, v|r)$ as DUAL-LDA does, but a preliminary investigation indicates that it does not perform well in this context. The most likely explanation is that LinkLDA generates its two arguments independently, which may be suitable for distinct argument positions of a given predicate but is unsuitable when one of those "arguments" is in fact the predicate.

The models developed in this paper, though intended for semantic modelling, also bear some similarity to the internals of generative syntax models such as the "infinite tree" (Finkel et al., 2007). In some ways, our models are less ambitious than comparable syntactic models as they focus on specific fragments of grammatical structure rather than learning a more general representation of sentence syntax. It would be interesting to evaluate whether this restricted focus improves the quality of the learned model or whether general syntax models can also capture fine-grained knowledge about combinatorial semantics.

3 Three selectional preference models

3.1 Notation

In the model descriptions below we assume a predicate vocabulary of V types, an argument vocabulary of N types and a relation vocabulary of R types. Each predicate type is associated with a singe relation; for example the predicate type *eat:V:dobj* (the direct object of the verb *eat*) is treated as distinct from *eat:V:subj* (the subject of the verb *eat*). The training corpus consists of W observations of argument-predicate pairs. Each model has at least one vocabulary of Z arbitrarily labelled latent variables. f_{zn} is the number of observations where the latent variable z has been associated with the argument type n, f_{zv} is the number of observations where z has been associated with the predicate type v and f_{zr} is the number of observations where z has been associated with the relation r. Finally, $f_{z\cdot}$ is the total number of observations associated with z and $f_{\cdot v}$ is the total number of observations containing the predicate v.

3.2 Latent Dirichlet Allocation

As noted above, LDA was originally introduced to model sets of documents in terms of topics, or clusters of terms, that they share in varying proportions. For example, a research paper on bioinformatics may use some vocabulary that is shared with general computer science papers and some vocabulary that is shared with biomedical papers. The analogical move from modelling document-term cooccurrences to modelling predicate-argument cooccurrences is intuitive: we assume that each predicate is associated with a distribution over semantic classes ("topics") and that these classes are shared across predicates. The high-level "generative story" for the LDA selectional preference model is as follows:

(1) For each predicate v, draw a multinomial distribution Θ_v over argument classes from a Dirichlet distribution with parameters α.

(2) For each argument class z, draw a multinomial distribution Φ_z over argument types from a Dirichlet with parameters β.

(3) To generate an argument for v, draw an argument class z from Θ_v and then draw an argument type n from Φ_z

The resulting model can be written as:

$$P(n|v, r) = \sum_z P(n|z)P(z|v, r) \quad (1)$$

$$\propto \sum_z \frac{f_{zn} + \beta}{f_{z\cdot} + N\beta} \frac{f_{zv} + \alpha_z}{f_{\cdot v} + \sum_{z'} \alpha_{z'}} \quad (2)$$

Due to multinomial-Dirichlet conjugacy, the distributions Θ_v and Φ_z can be integrated out and do not appear explicitly in the above formula. The first term in (2) can be seen as a smoothed estimate of the probability that class z produces the argument n; the second is a smoothed estimate of the probability that predicate v takes an argument belonging to class z. One important point is that the smoothing effects of the Dirichlet priors on Θ_v and Φ_z are greatest for predicates and arguments that are rarely seen, reflecting an intuitive lack of certainty. We assume an asymmetric Dirichlet prior on Θ_v (the α parameters can differ for each class) and a symmetric prior on Φ_z (all β parameters are equal); this follows the recommendations of Wallach et al. (2009) for LDA. This model estimates predicate-argument probabilities conditional on a given predicate v; it cannot by itself provide joint probabilities $P(n, v|r)$, which are needed for our plausibility evaluation.

Given a dataset of predicate-argument combinations and values for the hyperparameters α and β, the probability model is determined by the class assignment counts f_{zn} and f_{zv}. Following Griffiths and Steyvers (2004), we estimate the model by Gibbs sampling. This involves resampling the topic assignment for each observation in turn using probabilities estimated from all other observations. One efficiency bottleneck in the basic sampler described by Griffiths and Steyvers is that the entire set of topics must be iterated over for each observation. Yao et al. (2009) propose a reformulation that removes this bottleneck by separating the probability mass $p(z|n, v)$ into a number of buckets, some of which only require iterating over the topics currently assigned to instances of type n, typically far fewer than the total number of topics. It is possible to apply similar reformulations to the models presented in Sections 3.3 and 3.4 below; depending on the model and parameterisation this can reduce the running time dramatically.

Unlike some topic models such as HDP (Teh et al., 2006), LDA is *parametric*: the number of topics Z must be set by the user in advance. However, Wallach et al. (2009) demonstrate that LDA is relatively insensitive to larger-than-necessary choices of Z when the Dirichlet parameters α are optimised as part of model estimation. In our implementation we use the optimisation routines provided as part of the Mallet library, which use an iterative procedure to compute a maximum likelihood estimate of these hyperparameters.[2]

3.3 A Rooth et al.-inspired model

In Rooth et al.'s (1999) selectional preference model, a latent variable is responsible for generating both the predicate and argument types of an observation. The basic LDA model can be extended to capture this kind of predicate-argument interaction; the generative story for the resulting ROOTH-LDA model is as follows:

(1) For each relation r, draw a multinomial distribution Θ_r over interaction classes from a Dirichlet distribution with parameters α.

(2) For each class z, draw a multinomial Φ_z over argument types from a Dirichlet distribution with parameters β and a multinomial Ψ_z over predicate types from a Dirichlet distribution with parameters γ.

(3) To generate an observation for r, draw a class z from Θ_r, then draw an argument type n from Φ_z and a predicate type v from Ψ_z.

The resulting model can be written as:

$$P(n, v|r) = \sum_z P(n|z)P(v|z)P(z|r) \tag{3}$$

$$\propto \sum_z \frac{f_{zn} + \beta}{f_{z\cdot} + N\beta} \frac{f_{zv} + \gamma}{f_{z\cdot} + V\gamma} \frac{f_{zr} + \alpha_z}{f_{\cdot r} + \sum_{z'} \alpha_{z'}} \tag{4}$$

As suggested by the similarity between (4) and (2), the ROOTH-LDA model can be estimated by an LDA-like Gibbs sampling procedure.

Unlike LDA, ROOTH-LDA does model the joint probability $P(n, v|r)$ of a predicate and argument co-occurring. Further differences are that information about predicate-argument co-occurrence is only shared within a given interaction class rather than across the whole dataset and that the distribution Φ_z is not specific to the predicate v but rather to the relation r. This could potentially lead to a loss of model quality, but in practice the ability to induce "tighter" clusters seems to counteract any deterioration this causes.

3.4 A "dual-topic" model

In our third model, we attempt to combine the advantages of LDA and ROOTH-LDA by clustering arguments and predicates according to separate

[2] `http://mallet.cs.umass.edu/`

class vocabularies. Each observation is generated by two latent variables rather than one, which potentially allows the model to learn more flexible interactions between arguments and predicates.:

(1) For each relation r, draw a multinomial distribution Ξ_r over predicate classes from a Dirichlet with parameters κ.

(2) For each predicate class c, draw a multinomial Ψ_c over predicate types and a multinomial Θ_c over argument classes from Dirichlets with parameters γ and α respectively.

(3) For each argument class z, draw a multinomial distribution Φ_z over argument types from a Dirichlet with parameters β.

(4) To generate an observation for r, draw a predicate class c from Ξ_r, a predicate type from Ψ_c, an argument class z from Θ_c and an argument type from Φ_z.

The resulting model can be written as:

$$P(n,v|r) = \sum_c \sum_z P(n|z)P(z|c)P(v|c)P(c|r) \quad (5)$$

$$\propto \sum_c \sum_z \frac{f_{zn}+\beta}{f_{z\cdot}+N\beta} \frac{f_{zc}+\alpha_z}{f_{\cdot c}+\sum_{z'}\alpha_{z'}} \times \frac{f_{cv}+\gamma}{f_{c\cdot}+V\gamma} \frac{f_{cr}+\kappa_c}{f_{\cdot r}+\sum_{c'}\kappa_{c'}} \quad (6)$$

To estimate this model, we first resample the class assignments for all arguments in the data and then resample class assignments for all predicates. Other approaches are possible – resampling argument and then predicate class assignments for each observation in turn, or sampling argument and predicate assignments together by blocked sampling – though from our experiments it does not seem that the choice of scheme makes a significant difference.

4 Experimental setup

In the document modelling literature, probabilistic topic models are often evaluated on the likelihood they assign to unseen documents; however, it has been shown that higher log likelihood scores do not necessarily correlate with more semantically coherent induced topics (Chang et al., 2009). One popular method for evaluating selectional preference models is by testing the correlation between their predictions and human judgements of plausibility on a dataset of predicate-argument pairs. This can be viewed as a more semantically relevant measurement of model quality than likelihood-based methods, and also permits comparison with non-probabilistic models. In Section 5, we use two plausibility datasets to evaluate our models and compare to other previously published results.

We trained our models on the 90-million word written component of the British National Corpus (Burnard, 1995), parsed with the RASP toolkit (Briscoe et al., 2006). Predicates occurring with just one argument type were removed, as were all tokens containing non-alphabetic characters; no other filtering was done. The resulting datasets consisted of 3,587,172 verb-object observations with 7,954 predicate types and 80,107 argument types, 3,732,470 noun-noun observations with 68,303 predicate types and 105,425 argument types, and 3,843,346 adjective-noun observations with 29,975 predicate types and 62,595 argument types.

During development we used the verb-noun plausibility dataset from Padó et al. (2007) to direct the design of the system. Unless stated otherwise, all results are based on runs of 1,000 iterations with 100 classes, with a 200-iteration burnin period after which hyperparameters were reestimated every 50 iterations.[3] The probabilities estimated by the models ($P(n|v,r)$ for LDA and $P(n,v|r)$ for ROOTH- and DUAL-LDA) were sampled every 50 iterations post-burnin and averaged over three runs to smooth out variance. To compare plausibility scores for different predicates, we require the joint probability $P(n,v|r)$; as LDA does not provide this, we approximate $P_{LDA}(n,v|r) = P_{BNC}(v|r)P_{LDA}(n|v,r)$, where $P_{BNC}(v|r)$ is proportional to the frequency with which predicate v is observed as an instance of relation r in the BNC.

For comparison, we reimplemented the methods of Rooth et al. (1999) and Padó et al. (2007). As mentioned above, Rooth et al. use a latent-variable model similar to (4) but without priors, trained via EM. Our implementation (henceforth ROOTH-EM) chooses the number of classes from the range $(20, 25, \ldots, 50)$ through 5-fold cross-validation on a held-out log-likelihood measure. Settings outside this range did not give good results. Again, we run for 1,000 iterations and average predictions over

[3] These settings were based on the MALLET defaults; we have not yet investigated whether modifying the simulation length or burnin period is beneficial.

LDA	0	Nouns:	*agreement*, *contract*, *permission*, *treaty*, *deal*, ...
	1	Nouns	*information*, *datum*, *detail*, *evidence*, *material*, ...
	2	Nouns	*skill, knowledge, country, technique, understanding*, ...
ROOTH-LDA	0	Nouns	*force, team, army, group, troops*, ...
	0	Verbs	*join, arm, lead, beat, send*, ...
	1	Nouns	*door, eye, mouth, window, gate*, ...
	1	Verbs	*open, close, shut, lock, slam*, ...
DUAL-LDA	0N	Nouns	*house, building, site, home, station*, ...
	1N	Nouns	*stone, foot, bit, breath, line*, ...
	0V	Verbs	*involve, join, lead, represent, concern*, ...
	1V	Verbs	*see, break, have, turn, round*, ...
ROOTH-EM	0	Nouns	*system, method, technique, skill, model*, ...
	0	Verbs	*use, develop, apply, design, introduce*, ...
	1	Nouns	*eye, door, page, face, chapter*,...
	1	Verbs	*see, open, close, watch, keep*,...

Table 1: Most probable words for sample semantic classes induced from verb-object observations

three runs. Padó et al. (2007), a refinement of Erk (2007), is a non-probabilistic method that smooths predicate-argument counts with counts for other observed arguments of the same predicate, weighted by the similarity between arguments. Following their description, we use a 2,000-dimensional space of syntactic co-occurrence features appropriate to the relation being predicted, weight features with the G^2 transformation and compute similarity with the cosine measure.

5 Results

5.1 Induced semantic classes

Table 1 shows sample semantic classes induced by models trained on the corpus of BNC verb-object co-occurrences. LDA clusters nouns only, while ROOTH-LDA and ROOTH-EM learn classes that generate both nouns and verbs and DUAL-LDA clusters nouns and verbs separately. The LDA clusters are generally sensible: class 0 is exemplified by *agreement* and *contract* and class 1 by *information* and *datum*. There are some unintuitive blips, for example *country* appears between *knowledge* and *understanding* in class 2. The ROOTH-LDA classes also feel right: class 0 deals with nouns such as *force*, *team* and *army* which one might *join*, *arm* or *lead* and class 1 corresponds to "things that can be opened or closed" such as a *door*, an *eye* or a *mouth* (though the model also makes the questionable prediction that all these items can plausibly be locked or slammed). The DUAL-LDA classes are notably less coherent, especially when it comes to clustering verbs: DUAL-LDA's class 0V, like ROOTH-LDA's class 0, has verbs that take groups as objects but its class 1V mixes sensible conflations (*turn*, *round*) with very common verbs such as *see* and *have* and the unrelated *break*. The general impression given by inspection of the DUAL-LDA model is that it has problems with mixing and does not manage to learn a good model; we have tried a number of solutions (e.g., blocked sampling of argument and predicate classes), without overcoming this brittleness. Unsurprisingly, ROOTH-EM's classes have a similar feel to ROOTH-LDA; our general impression is that some of ROOTH-EM's classes look even more coherent than the LDA-based models, presumably because it does not use priors to smooth its per-class distributions.

5.2 Comparison with Keller and Lapata (2003)

Keller and Lapata (2003) collected a dataset of human plausibility judgements for three classes of grammatical relation: verb-object, noun-noun modification and adjective-noun modification. The items in this dataset were not chosen to balance plausibility and implausibility (as in prior psycholinguistic experiments) but according to their corpus frequency, leading to a more realistic task. 30 predicates were selected for each relation; each predicate was matched with three arguments from different co-occurrence bands in the BNC, e.g., *naughty-girl* (high frequency), *naughty-dog* (medium) and *naughty-lunch* (low). Each predicate was also matched with three random arguments

	Verb-object				Noun-noun				Adjective-noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
AltaVista (KL)	**.641**	–	.551	–	**.700**	–	.578	–	**.650**	–	.480	–
Google (KL)	.624	–	.520	–	.692	–	.595	–	.641	–	.473	–
BNC (RASP)	.620	**.614**	.196	.222	.544	.604	.114	.125	.543	**.622**	.135	.102
ROOTH-EM	.455	.487	.479	.520	.503	.491	.586	.625	.514	.463	.395	.355
Padó et al.	.484	.490	.398	.430	.431	.503	.558	.533	.479	.570	.120	.138
LDA	.504	.541	.558	.603	.615	**.641**	.636	.666	.594	.558	.468	.459
ROOTH-LDA	.520	.548	**.564**	**.605**	.607	.622	**.691**	**.722**	.575	.599	**.501**	**.469**
DUAL-LDA	.453	.494	.446	.516	.496	.494	.553	.573	.460	.400	.334	.278

Table 2: Results (Pearson r and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data

with which it does not co-occur in the BNC (e.g., *naughty-regime, naughty-rival, naughty-protocol*). In this way two datasets (*Seen* and *Unseen*) of 90 items each were assembled for each predicate.

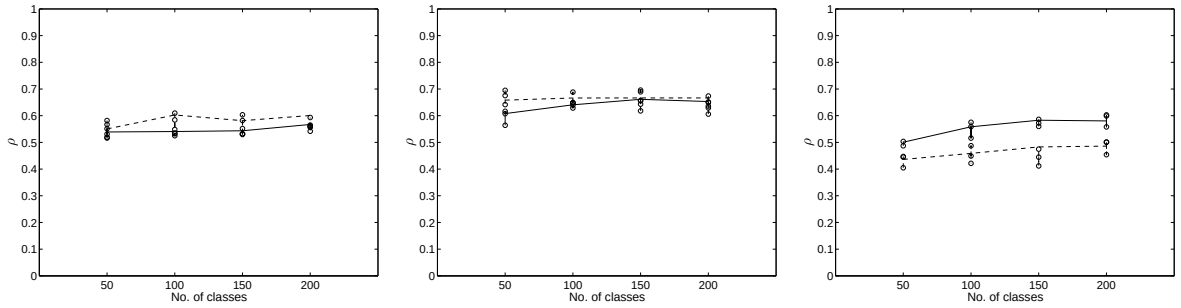
Table 2 presents results for a variety of predictive models – the Web frequencies reported by Keller and Lapata (2003) for two search engines, frequencies from the RASP-parsed BNC,[4] the reimplemented methods of Rooth et al. (1999) and Padó et al. (2007), and the LDA, ROOTH-LDA and DUAL-LDA topic models. Following Keller and Lapata, we report Pearson correlation coefficients between log-transformed predicted frequencies and the gold-standard plausibility scores (which are already log-transformed). We also report Spearman rank correlations except where we do not have the original predictions (the Web count models), for completeness and because the predictions of preference models are may not be log-normally distributed as corpus counts are. Zero values (found only in the BNC frequency predictions) were smoothed by 0.1 to facilitate the log transformation; it seems natural to take a zero prediction as a non-specific prediction of very low plausibility rather than a "missing value" as is done in other work (e.g., Padó et al., 2007).

Despite their structural differences, LDA and ROOTH-LDA perform similarly - indeed, their predictions are highly correlated. ROOTH-LDA scores best overall, outperforming Padó et al.'s (2007) method and ROOTH-EM on every dataset and evaluation measure, and outperforming Keller and Lapata's (2003) Web predictions on every Unseen dataset. LDA also performs consistently well, surpassing ROOTH-EM and Padó et al. on all but one occasion. For frequent predicate-argument pairs (Seen datasets), Web counts are clearly better; however, the BNC counts are unambiguously superior to LDA and ROOTH-LDA (whose predictions are based entirely on the generative model even for observed items) for the Seen verb-object data only. As might be suspected from the mixing problems observed with DUAL-LDA, this model does not perform as well as LDA and ROOTH-LDA, though it does hold its own against the other selectional preference methods.

To identify significant differences between models, we use the statistical test for correlated correlation coefficients proposed by Meng et al. (1992), which is appropriate for correlations that share the same gold standard.[5] For the seen data there are few significant differences: ROOTH-LDA and LDA are significantly better ($p < 0.01$) than Padó et al.'s model for Pearson's r on seen noun-noun data, and ROOTH-LDA is also significantly better ($p < 0.01$) using Spearman's ρ. For the unseen datasets, the BNC frequency predictions are unsurprisingly significantly worse at the $p < 0.01$ level than all smoothing models. LDA and ROOTH-LDA are significantly better ($p < 0.01$) than Padó et al. on every unseen dataset; ROOTH-EM is significantly better ($p < 0.01$) than Padó et al. on Unseen adjectives for both correlations. Meng et al.'s test does not find significant differences between ROOTH-EM and the LDA models despite the latter's clear advantages (a number of conditions do come close). This is because their predictions are highly correlated, which is perhaps

[4] The correlations presented here for BNC counts are notably better than those reported by Keller and Lapata (2003), presumably reflecting our use of full parsing rather than shallow parsing.

[5] We cannot compare our data to Keller and Lapata's Web counts as we do not possess their per-item scores.



(a) Verb-object (b) Noun-noun (c) Adjective-noun

Figure 1: Effect of number of argument classes on Spearman rank correlation with LDA: the solid and dotted lines show the Seen and Unseen datasets respectively; bars show locations of individual samples

unsurprising given that they are structurally similar models trained on the same data. We hypothesise that the main reason for the superior numerical performance of the LDA models over EM is the principled smoothing provided by the use of Dirichlet priors, which has a small but discriminative effect on model predictions. Collating the significance scores, we find that ROOTH-LDA achieves the most positive outcomes, followed by LDA and then by ROOTH-EM. DUAL-LDA is found significantly better than Padó et al.'s model on unseen adjective-noun combinations, and significantly worse than the same model on seen adjective-noun data.

Latent variable models that use EM for inference can be very sensitive to the number of latent variables chosen. For example, the performance of ROOTH-EM worsens quickly if the number of clusters is overestimated; for the Keller and Lapata datasets, settings above 50 classes lead to clear overfitting and a precipitous drop in Pearson correlation scores. On the other hand, Wallach et al. (2009) demonstrate that LDA is relatively insensitive to the choice of topic vocabulary size Z when the α and β hyperparameters are optimised appropriately during estimation. Figure 1 plots the effect of Z on Spearman correlation for the LDA model. In general, Wallach et al.'s finding for document modelling transfers to selectional preference models; within the range $Z = 50$–200 performance remains at a roughly similar level. In fact, we do not find that performance becomes significantly less robust when hyperparameter reestimation is deactiviated; correlation scores simply drop by a small amount (1–2 points), irrespective of the Z chosen. ROOTH-LDA (not graphed) seems slightly more sensitive to Z; this may be because the α parameters in this model operate on the relation level rather than the document level and thus fewer "observations" of class distributions are available when reestimating them.

5.3 Comparison with Bergsma et al. (2008)

As mentioned in Section 2.1, Bergsma et al. (2008) propose a discriminative approach to preference learning. As part of their evaluation, they compare their approach to a number of others, including that of Erk (2007), on a plausibility dataset collected by Holmes et al. (1989). This dataset consists of 16 verbs, each paired with one plausible object (e.g., *write-letter*) and one implausible object (*write-market*). Bergsma et al.'s model, trained on the 3GB AQUAINT corpus, is the only model reported to achieve perfect accuracy on distinguishing plausible from implausible arguments. It would be interesting to do a full comparison that controls for size and type of corpus data; in the meantime, we can report that the LDA and ROOTH-LDA models trained on verb-object observations in the BNC (about 4 times smaller than AQUAINT) also achieve a perfect score on the Holmes et al. data.[6]

6 Conclusions and future work

This paper has demonstrated how Bayesian techniques originally developed for modelling the topical structure of documents can be adapted to learn probabilistic models of selectional preference. These models are especially effective for estimating plausibility of low-frequency items, thus distinguishing rarity from clear implausibility.

The models presented here derive their predictions by modelling predicate-argument plausibility through the intermediary of latent variables. As observed in Section 5.2 this may be a suboptimal

[6] Bergsma et al. report that all plausible pairs were seen in their corpus; three were unseen in ours, as well as 12 of the implausible pairs.

strategy for frequent combinations, where corpus counts are probably reliable and plausibility judgements may be affected by lexical collocation effects. One principled method for folding corpus counts into LDA-like models would be to use hierarchical priors, as in the n-gram topic model of Wallach (2006). Another potential direction for system improvement would be an integration of our generative model with Bergsma et al.'s (2008) discriminative model – this could be done in a number of ways, including using the induced classes of a topic model as features for a discriminative classifier or using the discriminative classifier to produce additional high-quality training data from noisy unparsed text.

Comparison to plausibility judgements gives an intrinsic measure of model quality. As mentioned in the Introduction, selectional preferences have many uses in NLP applications, and it will be interesting to evaluate the utility of Bayesian preference models in contexts such as semantic role labelling or human sentence processing modelling. The probabilistic nature of topic models, coupled with an appropriate probabilistic task model, may facilitate the integration of class induction and task learning in a tight and principled way. We also anticipate that latent variable models will prove effective for learning selectional preferences of semantic predicates (e.g., FrameNet roles) where direct estimation from a large corpus is not a viable option.

Acknowledgements

This work was supported by EPSRC grant EP/G051070/1. I am grateful to Frank Keller and Mirella Lapata for sharing their plausibility data, and to Andreas Vlachos and the anonymous ACL and CoNLL reviewers for their helpful comments.

References

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, Andrew Y. Ng, and Michael I. Jordan. 2003. 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.

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

Lou Burnard, 1995. *Users' Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service, Oxford, UK.

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.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of ACL-07*, Prague, Czech Republic.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2007. The infinite tree. In *Proceedings of ACL-07*, Prague, Czech Republic.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

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

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

Virginia M. Holmes, Laurie Stowe, and Linda Cupples. 1989. Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language*, 28(6):668–689.

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

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of CoNLL-06*, New York, NY.

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

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP-CoNLL-07*, Prague, Czech Republic.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of NAACL-HLT-07*, Rochester, NY.

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

Philip S. 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 of 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 ACL-99*, College Park, MD.

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.

Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

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

Hanna Wallach. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of ICML-06*, Pittsburgh, PA.

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

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of KDD-09*, Paris, France.

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.

Huibin Zhang, Mingjie Zhu, Shuming Shi, and Ji-Rong Wen. 2009. Employing topic models for pattern-based semantic class discovery. In *Proceedings of ACL-IJCNLP-09*, Singapore.