In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2010), pp. 1173--1182, October 9-11, 2010, MIT, Massachusetts, USA.

A Mixture Model with Sharing for Lexical Semantics

Joseph Reisinger
Department of Computer Science
University of Texas at Austin
1616 Guadalupe, Suite 2.408
Austin, TX, 78701
joeraii@cs.utexas.edu

Raymond Mooney
Department of Computer Science
University of Texas at Austin
1616 Guadalupe, Suite 2.408
Austin, TX, 78701
mooney@cs.utexas.edu

Abstract

We introduce *tiered clustering*, a mixture model capable of accounting for varying degrees of shared (context-independent) feature structure, and demonstrate its applicability to inferring distributed representations of word meaning. Common tasks in lexical semantics such as word relatedness or selectional preference can benefit from modeling such structure: Polysemous word usage is often governed by some common background metaphoric usage (e.g. the senses of *line* or *run*), and likewise modeling the selectional preference of verbs relies on identifying commonalities shared by their typical arguments. Tiered clustering can also be viewed as a form of soft feature selection, where features that do not contribute meaningfully to the clustering can be excluded. We demonstrate the applicability of tiered clustering, highlighting particular cases where modeling shared structure is beneficial and where it can be detrimental.

1 Introduction

Word meaning can be represented as high-dimensional vectors inhabiting a common space whose dimensions capture semantic or syntactic properties of interest (e.g. Erk and Pado, 2008; Lowe, 2001). Such *vector-space* representations of meaning induce measures of word similarity that can be tuned to correlate well with judgements made by humans. Previous work has focused on designing feature representations and semantic spaces that capture salient properties of word meaning (e.g. Curran, 2004; Gabrilovich and Markovitch, 2007; Landauer and Dumais, 1997), often leveraging the distributional hypothesis, i.e. that similar words appear in similar contexts (Miller and Charles, 1991; Pereira et al., 1993).

Since vector-space representations are constructed at the lexical level, they conflate multiple word meanings into the same vector, e.g. collapsing occurrences of *bank*$_{\text{institution}}$ and *bank*$_{\text{river}}$. Methods such as *Clustering by Committee* (Pantel, 2003) and *multi-prototype* representations (Reisinger and Mooney, 2010) address this issue by performing word-sense disambiguation across word occurrences, and then building meaning vectors from the disambiguated words. Such approaches can readily capture the structure of homonymous words with several unrelated meanings (e.g. *bat* and *club*), but are not suitable for representing the common metaphor structure found in highly polysemous words such as *line* or *run*.

In this paper, we introduce *tiered clustering*, a novel probabilistic model of the *shared structure* often neglected in clustering problems. Tiered clustering performs *soft* feature selection, allocating features between a Dirichlet Process clustering model and a background model consisting of a single component. The background model accounts for features commonly shared by all occurrences (i.e. context-independent feature variation), while the clustering model accounts for variation in word usage (i.e. context-dependent variation, or *word senses*; Table 1).

Using the tiered clustering model, we derive a multi-prototype representation capable of capturing varying degrees of sharing between word senses, and demonstrate its effectiveness in lexical semantic tasks where such sharing is desirable. In particular we show that tiered clustering outperforms the multi-prototype approach for (1) selectional preference (Resnik, 1997; Pantel et al., 2007), i.e. predicting the typical filler of an argument slot of a verb, and (2) word-relatedness in the presence of highly polysemous words. The former case exhibits a high degree of explicit structure, especially for more selectionally restrictive verbs (e.g. the set of things that can be *eaten* or can *shoot*).

The remainder of the paper is organized as follows: Section 2 gives relevant background on the methods compared, Section 3 outlines the multi-prototype model based on the Dirichlet Process mixture model, Section 4 derives the tiered clustering model, Section 5 discusses similarity metrics, Section 6 details the experimental setup and includes a micro-analysis of feature selection, Section 7 presents results applying tiered clustering to word relatedness and selectional preference, Section 8 discusses future work, and Section 9 concludes.

2 Background

Models of the *attributional similarity* of concepts, i.e. the degree to which concepts overlap based on their attributes (Turney, 2006), are commonly implemented using vector-spaces derived from (1) word collocations (Schütze, 1998), directly leveraging the distributional hypothesis (Miller and Charles, 1991), (2) syntactic relations (Padó and Lapata, 2007), (3) structured corpora (e.g. Gabrilovich and Markovitch (2007)) or (4) latent semantic spaces (Finkelstein et al., 2001; Landauer and Dumais, 1997). Such models can be evaluated based on their correlation with human-reported lexical similarity judgements using e.g. the WordSim-353 collection (Finkelstein et al., 2001). Distributional methods exhibit a high degree of scalability (Gorman and Curran, 2006) and have been applied broadly in information retrieval (Manning et al., 2008), large-scale taxonomy induction (Snow et al., 2006), and knowledge acquisition (Van Durme and Paşca, 2008).

Reisinger and Mooney (2010) introduced a *multi-prototype* approach to vector-space lexical semantics where individual words are represented as collections of “prototype” vectors. This representation is capable of accounting for homonymy and polysemy, as well as other forms of variation in word usage, like similar context-dependent methods (Erk and Pado, 2008). The set of vectors for a word is determined by unsupervised *word sense discovery* (Schütze, 1998), which clusters the contexts in which a word appears. Average prototype vectors are then computed separately for each cluster, producing a distributed representation for each word.

LIFE	
	all, about, life, would, death
	my, you, real, your, about
	spent, years, rest, lived, last
	sentenced, imprisonment, sentence, prison
	insurance, peer, Baron, member, company
	Guru, Rabbi, Baba, la, teachings
RADIO	
	station, radio, stations, television
	amateur, frequency, waves, system
	show, host, personality, American
	song, single, released, airplay
	operator, contact, communications, message
WIZARD	
	evil, powerful, magic, wizard
	Merlin, King, Arthur, Arthurian
	fairy, wicked, scene, tale
	Harry, Potter, Voldemort, Dumbledore
STOCK	
	stock, all, other, company, new
	market, crash, markets, price, prices
	housing, breeding, fish, water, horses
	car, racing, cars, NASCAR, race, engine
	card, cards, player, pile, game, paper
	rolling, locomotives, line, new, railway

Table 1: Example tiered clustering representation of words with varying degrees of polysemy. Each boxed set shows the most common background (shared) features, and each prototype captures one thematic usage of the word. For example, *wizard* is broken up into a background cluster describing features common to all usages of the word (e.g., *magic* and *evil*) and several genre-specific usages (e.g. *Merlin*, *fairy tales* and *Harry Potter*).

Distributional methods have also proven to be a powerful approach to modeling *selectional preference* (Padó et al., 2007; Pantel et al., 2007), rivaling methods based on existing semantic resources such as WordNet (Clark and Weir, 2002; Resnik, 1997) and FrameNet (Padó, 2007) and performing nearly as well as supervised methods (Herdağdelen and Baroni, 2009). Selectional preference has been shown to be useful for, e.g., resolving ambiguous attachments (Hindle and Rooth, 1991), word sense disambiguation (McCarthy and Carroll, 2003) and semantic role labeling (Gildea and Jurafsky, 2002).

3 Multi-Prototype Models

Representing words as mixtures over several prototypes has proven to be a powerful approach to vector-space lexical semantics (Pantel, 2003; Pantel et al., 2007; Reisinger and Mooney, 2010). In this section we briefly introduce a version of the multi-prototype model based on the Dirichlet Process Mixture Model (DPMM), capable of inferring automatically the number of prototypes necessary for each word (Rasmussen, 2000). Similarity between two DPMM word-representations is then computed as a function of their cluster centroids (§5), instead of the centroid of all the word's occurrences.

Multiple prototypes for each word w are generated by clustering feature vectors $v(c)$ derived from each occurrence $c \in \mathcal{C}(w)$ in a large textual corpus and collecting the resulting cluster centroids $\pi_k(w), k \in [1, K_w]$. This approach is commonly employed in unsupervised word sense discovery; however, we do not assume that clusters correspond to word senses. Rather, we only rely on clusters to capture meaningful variation in word usage.

Instead of assuming all words can be represented by the same number of clusters, we allocate representational flexibility dynamically using the DPMM. The DPMM is an infinite capacity model capable of assigning data to a variable, but finite number of clusters K_w, with probability of assignment to cluster k proportional to the number of data points previously assigned to k. A single parameter η controls the degree of smoothing, producing more uniform clusterings as $\eta \to \infty$. Using this model, the number of clusters no longer needs to be fixed a priori, allowing the model to allocate expressivity dynamically to concepts with richer structure. Such a model naturally allows the word representation to allocate additional capacity for highly polysemous words, with the number of clusters growing logarithmically with the number of occurrences. The DPMM has been used for rational models of concept organization (Sanborn et al., 2006), but to our knowledge has not yet been applied directly to lexical semantics.

4 Tiered Clustering

Tiered clustering allocates features between two submodels: a (context-dependent) DPMM and a single (context-independent) *background* component. This model is similar structurally to the feature selective clustering model proposed by Law et al. (2002). However, instead of allocating entire feature *dimensions* between model and background components, assignment is done at the level of individual feature occurrences, much like topic assignment in Latent Dirichlet Allocation (LDA; Griffiths et al., 2007). At a high level, the tiered model can be viewed as a combination of a multi-prototype model and a single-prototype back-off model. However, by leveraging both representations in a joint framework, uninformative features can be removed from the clustering, resulting in more semantically tight clusters.

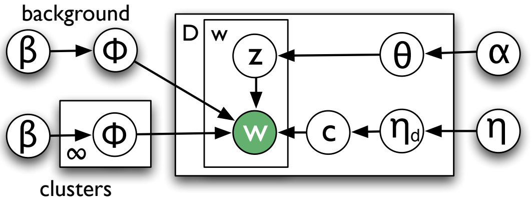

Figure 1: Plate diagram for the tiered clustering model with cluster indicators drawn from the Chinese Restaurant Process.

Concretely, each word occurrence $\mathbf{w}_d$ first selects a cluster $\boldsymbol{\phi}_d$ from the DPMM; then each feature $w_{i,d}$ is generated from either the background model $\boldsymbol{\phi}_{\text{back}}$ or the selected cluster $\boldsymbol{\phi}_d$, determined by the tier indicator $z_{i,d}$. The full generative model for tiered clustering is given by

$$\begin{array}{rcll}
\theta_d|\alpha & \sim & \text{Beta}(\alpha) & d \in D, \\
\boldsymbol{\phi}_d|\boldsymbol{\beta}, G_0 & \sim & \text{DP}(\boldsymbol{\beta}, G_0) & d \in D, \\
\boldsymbol{\phi}_{\text{back}}|\boldsymbol{\beta}_{\text{back}} & \sim & \text{Dirichlet}(\boldsymbol{\beta}_{\text{back}}) & \\
z_{i,d}|\theta_d & \sim & \text{Bernoulli}(\theta_d) & i \in |\mathbf{w}_d|, \\
w_{i,d}|\boldsymbol{\phi}_d, z_{i,d} & \sim & \begin{cases} \text{Mult}(\boldsymbol{\phi}_{\text{back}}) \\ \quad (z_{i,d} = 1) \\ \text{Mult}(\boldsymbol{\phi}_d) \\ \quad \text{(otherwise)} \end{cases} & i \in |\mathbf{w}_d|,
\end{array}$$

where $\boldsymbol{\alpha}$ controls the per-data tier distribution smoothing and $\boldsymbol{\beta}$ controls the uniformity of the DP cluster allocation. The DP is parameterized by a base measure G_0, controlling the per-cluster term distribution smoothing; which use a Dirichlet with hyperparameter η, as is common (Figure 1).

Since the background topic is shared across all occurrences, it can account for features with *context-independent* variance, such as stop words and other high-frequency noise, as well as the central tendency of the collection (Table 1). Furthermore, it is possible to put an asymmetric prior on η, yielding more fine-grained control over the assumed *uniformity* of the occurrence of noisy features, unlike in the model proposed by Law et al. (2002).

Although exact posterior inference is intractable in this model, we derive an efficient *collapsed Gibbs sampler* via analogy to LDA (Appendix 1).

5 Measuring Semantic Similarity

Due to its richer representational structure, computing similarity in the multi-prototype model is less straightforward than in the single prototype case. Reisinger and Mooney (2010) found that simply averaging all similarity scores over all pairs of prototypes (sampled from the cluster distributions) performs reasonably well and is robust to noise. Given two words w and w', this *AvgSim* metric is

$$\text{AvgSim}(w, w') \stackrel{\text{def}}{=} \frac{1}{K_w K_{w'}} \sum_{j=1}^{K_w} \sum_{k=1}^{K_{w'}} d(\pi_k(w), \pi_j(w'))$$

K_w and $K_{w'}$ are the number of clusters for w and w' respectively, and $d(\cdot, \cdot)$ is a standard distributional similarity measure (e.g. cosine distance). As cluster sizes become more uniform, AvgSim tends towards the single prototype similarity,[1] hence the effectiveness of AvgSim stems from boosting the influence of small clusters.

Tiered clustering representations offer more possibilities for computing semantic similarity than multi-prototype, as the background prototype can be treated separately from the other prototypes. We make use of a simple sum of the distance between the two background components, and the AvgSim of the two sets of clustering components.

6 Experimental Setup

6.1 Corpus

Word occurrence statistics are collected from a snapshot of English Wikipedia taken on Sept. 29th, 2009. Wikitext markup is removed, as are articles with fewer than 100 words, leaving 2.8M articles with a total of 2.05B words. Wikipedia was chosen due to its semantic breadth.

6.2 Evaluation Methodology

We evaluate the tiered clustering model on two problems from lexical semantics: word relatedness and selectional preference. For the word relatedness evaluation, we compared the predicted similarity of word pairs from each model to two collections of human similarity judgements: WordSim-353 (Finkelstein et al., 2001) and the Princeton Evocation relations (WN-Evocation, Ma et al., 2009).

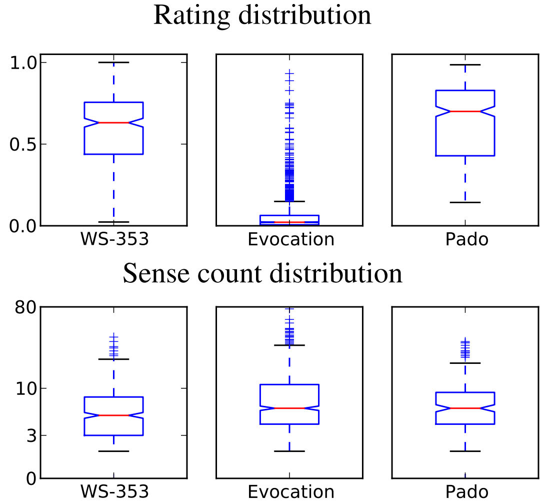

Figure 2: (**top**) The distribution of ratings (scaled [0,1]) on WS-353, WN-Evocation and Padó datasets. (**bottom**) The distribution of sense counts for each data set (log-domain), collected from WordNet 3.0.

WS-353 contains between 13 and 16 human similarity judgements for each of 353 word pairs, rated on a 1–10 integer scale. WN-Evocation is significantly larger than WS-353, containing over 100k similarity comparisons collected from trained human raters. Comparisons are assigned to only 3-5 human raters on average and contain a significantly higher fraction of zero- and low-similarity items than WS-353 (Figure 2), reflecting more accurately real-world lexical semantics settings. In our experiments we discard all comparisons with fewer than 5 ratings and then sample 10% of the remaining pairs uniformly at random, resulting in a test set with 1317 comparisons.

For selectional preference, we employ the *Padó* dataset, which contains 211 verb-noun pairs with human similarity judgements for how plausible the noun is for each argument of the verb (2 arguments per verb, corresponding roughly to subject and object). Results are averaged across 20 raters; typical inter-rater agreement is $\rho = 0.7$ (Padó et al., 2007).

In all cases correlation with human judgements is computed using Spearman's nonparametric rank correlation (ρ) with average human judgements

[1] This can be problematic for certain clustering methods that specify uniform priors over cluster sizes; however the DPMM naturally exhibits a linear decay in cluster sizes with the $\mathbb{E}[\#\text{ clusters of size } M] = \eta/M$.

(Agirre et al., 2009).

6.3 Feature Representation

In the following analyses we confine ourselves to representing word occurrences using unordered unigrams collected from a window of size T=10 centered around the occurrence, represented using *tf-idf* weighting. Feature vectors are pruned to a fixed length f, discarding all but the highest-weight features (f is selected via empirical validation, as described in the next section). Finally, semantic similarity between word pairs is computed using cosine distance (ℓ_2-normalized dot-product).[2]

6.4 Feature Pruning

Feature pruning is one of the most significant factors in obtaining high correlation with human similarity judgements using vector-space models, and has been suggested as one way to improve sense disambiguation for polysemous verbs (Xue et al., 2006). In this section, we calibrate the single prototype and multi-prototype methods on WS-353, reaching the limit of human and oracle performance and demonstrating robust performance gains even with semantically impoverished features. In particular we obtain ρ=0.75 correlation on WS-353 using *only* unigram collocations and ρ=0.77 using a fixed-K multi-prototype representation (Figure 3; Reisinger and Mooney, 2010). This result rivals average human performance, obtaining correlation near that of the supervised oracle approach of Agirre et al. (2009).

The optimal pruning cutoff depends on the feature weighting and number of prototypes as well as the feature representation. *t-test* and χ^2 features are most robust to feature noise and perform well even with no pruning; *tf-idf* yields the best results but is most sensitive to the pruning parameter (Figure 3). As the number of features increases, more pruning is required to combat feature noise.

Figure 4 breaks down the similarity pairs into four quantiles for each data set and then shows correlation separately for each quantile. In general the more polarized data quantiles (1 and 4) have higher correlation, indicating that fine-grained distinctions in semantic distance are easier for those sets.[3] Feature pruning improves correlations in quantiles 2–4 while reducing correlation in quantile 1 (lowest similarity). This result is to be expected as more features are necessary to make fine-grained distinctions between dissimilar pairs.

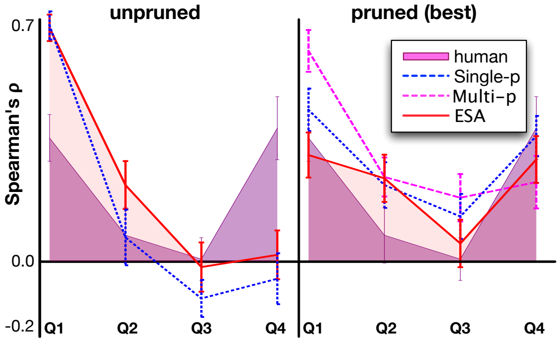

Figure 4: Correlation results on WS-353 broken down over quantiles in the human ratings. Quantile ranges are shown in Figure 2. In general ratings for highly similar (dissimilar) pairs are more predictable (quantiles 1 and 4) than middle similarity pairs (quantiles 2, 3). ESA shows results for a more semantically rich feature set derived using Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007).

7 Results

We evaluate four models: (1) the standard single-prototype approach, (2) the DPMM multi-prototype approach outlined in §3, (3) a simple combination of the multi-prototype and single-prototype approaches (MP+SP)[4] and (4) the tiered clustering approach (§4). Each data set is divided into 5 quantiles based on per-pair average sense counts,[5] collected from WordNet 3.0 (Fellbaum, 1998); examples of pairs in the *high-polysemy* quantile are shown in Table 2. Unless otherwise specified, both DPMM multi-prototype and tiered clustering

[2] (**Parameter robustness**) We observe lower correlations on average for T=25 and T=5 and therefore observe T=10 to be near-optimal. Substituting weighted Jaccard similarity for cosine does not significantly affect the results in this paper.

[3] The fact that the per-quantile correlation is significantly lower than the full correlation e.g. in the human case indicates that fine-grained ordering (within quantile) is more difficult than coarse-grained (between quantile).

[4] (**MP+SP**) Tiered clustering's ability to model both shared and idiosyncratic structure can be easily approximated by using the single prototype model as the shared component and multi-prototype model as the clustering. However, unlike in the tiered model, all features are assigned to *both* components. We demonstrate that this simplification actually hurts performance.

[5] Despite many skewed pairs (e.g. line has 36 senses while insurance has 3), we found that arithmetic average and geometric average perform the same.

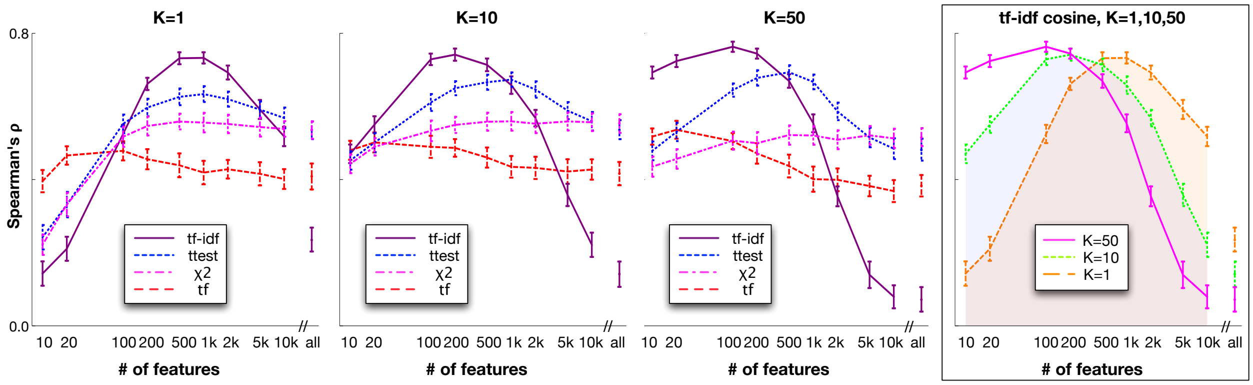

Figure 3: Effects of feature pruning and representation on WS-353 correlation broken down across multi-prototype representation size. In general *tf-idf* features are the most sensitive to pruning level, yielding the highest correlation for moderate levels of pruning and significantly lower correlation than other representations without pruning. The optimal amount of pruning varies with the number of prototypes used, with fewer features being optimal for more clusters. Bars show 95% confidence intervals.

WordSim-353
stock-live, start-match, line-insurance, game-round, street-place, company-stock
Evocation
break-fire, clear-pass, take-call, break-tin, charge-charge, run-heat, social-play
Padó
see-drop, see-return, hit-stock, raise-bank, see-face, raise-firm, raise-question

Table 2: Examples of highly polysemous pairs from each data set using sense counts from WordNet.

Method	$\rho \cdot 100$	$\mathbb{E}[C]$	background
Single prototype	73.4±0.5	1.0	-
high polysemy	76.0±0.9	1.0	-
Multi-prototype	76.8±0.4	14.8	-
high polysemy	79.3±1.3	12.5	-
MP+SP	75.4±0.5	14.8	-
high polysemy	80.1±1.0	12.5	-
Tiered	76.9±0.5	27.2	43.0%
high polysemy	83.1±1.0	24.2	43.0%

Table 3: Spearman's correlation on the WS-353 data set. *All* refers to the full set of pairs, *high polysemy* refers to the top 20% of pairs, ranked by sense count. $\mathbb{E}[C]$ is the average number of clusters employed by each method and *background* is the average percentage of features allocated by the tiered model to the background cluster. 95% confidence intervals are computed via bootstrapping.

use symmetric Dirichlet hyperparameters, β=0.1, η=0.1, and tiered clustering uses α=10 for the background/clustering allocation smoother.

7.1 WordSim-353

Correlation results for WS-353 are shown in Table 3. In general the approaches incorporating multiple prototypes outperform single prototype ($\rho = 0.768$ vs. $\rho = 0.734$). The tiered clustering model does not significantly outperform either the multi-prototype or MP+SP models on the full set, but yields significantly higher correlation on the high-polysemy set.

The tiered model generates more clusters than DPMM multi-prototype (27.2 vs. 14.8), despite using the same hyperparameter settings: Since words commonly shared across clusters have been allocated to the background component, the cluster components have less overlap and hence the model naturally allocates more clusters.

Examples of the tiered clusterings for several words from WS-353 are shown in Table 1 and corresponding clusters from the multi-prototype approach are shown in Table 4. In general the background component does indeed capture commonalities between all the sense clusters (e.g. all wizards use magic) and hence the tiered clusters are more semantically pure. This effect is most visible in *thematically polysemous* words, e.g. *radio* and *wizard*.

7.2 Evocation

Compared to WS-353, the WN-Evocation pair set is sampled more uniformly from English word pairs and hence contains a significantly larger fraction of unrelated words, reflecting the fact that word sim-

LIFE	my, you, real, about, your, would
	years, spent, rest, lived, last
	sentenced, imprisonment, sentence, prison
	years, cycle, life, all, expectancy, other
	all, life, way, people, human, social, many
RADIO	station, FM, broadcasting, format, AM
	radio, station, stations, amateur,
	show, station, host, program, radio
	stations, song, single, released, airplay
	station, operator, radio, equipment, contact
WIZARD	evil, magic, powerful, named, world
	Merlin, King, Arthur, powerful, court
	spells, magic, cast, wizard, spell, witch
	Harry, Dresden, series, Potter, character
STOCK	market, price, stock, company, value, crash
	housing, breeding, all, large, stock, many
	car, racing, company, cars, summer, NASCAR
	stock, extended, folded, card, barrel, cards
	rolling, locomotives, new, character, line

Table 4: Example DPMM multi-prototype representation of words with varying degrees of polysemy. Compared to the tiered clustering results in Table 1 the multi-prototype clusters are significantly less pure for *thematically polysemous* words such as *radio* and *wizard*.

ilarity is a sparse relation (Figure 2 top). Furthermore, it contains proportionally more highly polysemous words relative to WS-353 (Figure 2 bottom).

On WN-Evocation, the single prototype and multi-prototype do not differ significantly in terms of correlation (ρ=0.198 and ρ=0.201 respectively; Table 5), while SP+MP yields significantly lower correlation (ρ=0.176), and the tiered model yields significantly higher correlation (ρ=0.224). Restricting to the top 20% of pairs with highest human similarity judgements yields similar outcomes, with single prototype, multi-prototype and SP+MP statistically indistinguishable (ρ=0.239, ρ=0.227 and ρ=0.235), and tiered clustering yielding significantly higher correlation (ρ=0.277). Likewise tiered clustering achieves the most significant gains on the high polysemy subset.

7.3 Selectional Preference

Tiered clustering is a natural model for verb selectional preference, especially for more selectionally restrictive verbs: the set of words that appear in a particular argument slot naturally have some kind of commonality (i.e. they can be *eaten* or can *promise*). The background component of the tiered clustering model can capture such general argument structure. We model each verb argument slot in the Padó set with a separate tiered clustering model, separating terms co-occurring with the target verb according to which slot they fill.

Method	$\rho \cdot 100$	$\mathbb{E}[C]$	background
Single prototype	19.8±0.6	1.0	-
high similarity	23.9±1.1	1.0	-
high polysemy	11.5±1.2	1.0	-
Multi-prototype	20.1±0.5	14.8	-
high similarity	22.7±1.2	14.1	-
high polysemy	13.0±1.3	13.2	-
MP+SP	17.6±0.5	14.8	-
high similarity	23.5±1.2	14.1	-
high polysemy	11.4±1.0	13.2	-
Tiered	22.4±0.6	29.7	46.6%
high similarity	27.7±1.3	29.9	47.2%
high polysemy	15.4±1.1	27.4	46.6%

Table 5: Spearman's correlation on the Evocation data set. The *high similarity* subset contains the top 20% of pairs sorted by average rater score.

Method	$\rho \cdot 100$	$\mathbb{E}[C]$	background
Single prototype	25.8±0.8	1.0	-
high polysemy	17.3±1.7	1.0	-
Multi-prototype	20.2±1.0	18.5	-
high polysemy	14.1±2.4	17.4	-
MP+SP	19.7±1.0	18.5	-
high polysemy	10.5±2.5	17.4	-
Tiered	29.4±1.0	37.9	41.7%
high polysemy	28.5±2.4	37.4	43.2%

Table 6: Spearman's correlation on the Padó data set.

On the Padó set, the performance of the DPMM multi-prototype approach breaks down and it yields significantly lower correlation with human norms than the single prototype (ρ=0.202 vs. ρ=0.258; Table 6), due to its inability to capture the shared structure among verb arguments. Furthermore combining with the single prototype does not significantly change its performance (ρ=0.197). Moving to the tiered model, however, yields significant improvements in correlation over the other models (ρ=0.294), primarily improving correlation in the case of highly polysemous verbs and arguments.

8 Discussion and Future Work

We have demonstrated a novel model for distributional lexical semantics capable of capturing both shared (context-independent) and idiosyncratic (context-dependent) structure in a set of word occurrences. The benefits of this tiered model were most pronounced on a selectional preference task, where there is significant shared structure imposed by conditioning on the verb. Although our results on the Padó are not state of the art,[6] we believe this to be due to the impoverished vector-space design; tiered clustering can be applied to more expressive vector spaces, such as those incorporating dependency parse and FrameNet features.

One potential explanation for the superior performance of the tiered model vs. the DPMM multi-prototype model is simply that it allocates more clusters to represent each word (Reisinger and Mooney, 2010). However, we find that decreasing the hyperparameter β (decreasing vocabulary smoothing and hence increasing the effective number of clusters) beyond $\beta = 0.1$ actually hurts multi-prototype performance. The additional clusters do not provide more semantic content due to significant background similarity.

Finally, the DPMM multi-prototype and tiered clustering models allocate clusters based on the variance of the underlying data set. We observe a *negative* correlation (ρ=-0.33) between the number of clusters allocated by the DPMM and the number of word senses found in WordNet. This result is most likely due to our use of unigram context window features, which induce clustering based on thematic rather than syntactic differences. Investigating this issue is future work.

(**Future Work**) The word similarity experiments can be expanded by breaking pairs down further into highly homonymous and highly polysemous pairs, using e.g. WordNet to determine how closely related the senses are. With this data it would be interesting to validate the hypothesis that the percentage of features allocated to the background cluster is correlated with the degree of homonymy.

The basic tiered clustering can be extended with additional background tiers, allocating more expressivity to model background feature variation. This class of models covers the spectrum between a pure topic model (all background tiers) and a pure clustering model and may be reasonable when there is believed to be more background structure (e.g. when jointly modeling all verb arguments). Furthermore, it is straightforward to extend the model to a two-tier, two-clustering structure capable of additionally accounting for commonalities *between* arguments.

Applying more principled feature selection approaches to vector-space lexical semantics may yield more significant performance gains. Towards this end we are currently evaluating two classes of approaches for setting pruning parameters per-word instead of globally: (1) *subspace clustering*, i.e. unsupervised feature selection (e.g., Parsons et al., 2004) and (2) *multiple clustering*, i.e. finding feature partitions that lead to disparate clusterings (e.g., Shafto et al., 2006).

9 Conclusions

This paper introduced a simple probabilistic model of *tiered clustering* inspired by feature selective clustering that leverages feature exchangeability to allocate data features between a clustering model and shared component. The ability to model background variation, or shared structure, is shown to be beneficial for modeling words with high polysemy, yielding increased correlation with human similarity judgements modeling word relatedness and selectional preference. Furthermore, the tiered clustering model is shown to significantly outperform related models, yielding qualitatively more precise clusters.

Acknowledgments

Thanks to Yinon Bentor and Bryan Silverthorn for many illuminating discussions. This work was supported by an NSF Graduate Research Fellowship to the first author, and a Google Research Award.

A Collapsed Gibbs Sampler

In order to sample efficiently from this model, we leverage the Chinese Restaurant Process representation of the DP (cf., Aldous, 1985), introducing a per-word-occurrence cluster indicator c_d. Word occurrence features are then drawn from a combination of a single cluster component indicated by c_d and the background topic.

By exploiting conjugacy, the latent variables $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and η_d can be integrated out, yielding an efficient

[6] E.g., Padó et al. (2007) report ρ=0.515 on the same data.

collapsed Gibbs sampler. The likelihood of word occurrence d is given by

$$P(\mathbf{w}_d|\mathbf{z}, c_d, \boldsymbol{\phi}) = \prod_i P(w_{i,d}|\boldsymbol{\phi}_{c_d})^{\delta(z_{d,i}=0)} P(w_{i,d}|\boldsymbol{\phi}_{\text{noise}})^{\delta(z_{d,i}=1)}.$$

Hence, this model can be viewed as a two-topic variant of LDA with the addition of a per-word-occurrence (i.e. document) cluster indicator.[7] The update rule for the latent tier indicator $\mathbf{z}$ is similar to the update rule for 2-topic LDA, with the background component as the first topic and the second topic being determined by the per-word-occurrence cluster indicator $\boldsymbol{c}$.

We can efficiently approximate $p(\mathbf{z}|\mathbf{w})$ via Gibbs sampling, which requires the complete conditional posteriors for all $z_{i,d}$. These are

$$P(z_{i,d} = t|\mathbf{z}_{-(i,d)}, \mathbf{w}, \alpha, \beta) = \frac{n_t^{(w_{i,d})} + \beta}{\sum_w (n_t^{(w)} + \beta)} \frac{n_t^{(d)} + \alpha}{\sum_j (n_j^{(d)} + \alpha)}.$$

where $\mathbf{z}_{-(i,d)}$ is shorthand for the set $\mathbf{z} - \{z_{i,d}\}$, $n_t^{(w)}$ is the number of occurrences of word w in topic t not counting $w_{i,d}$ and $n_t^{(d)}$ is the number of features in occurrence d assigned to topic t, not counting $w_{i,d}$.

Likewise sampling the cluster indicators conditioned on the data $p(\boldsymbol{c}_d|\mathbf{w}, \boldsymbol{c}_{-d}, \alpha, \eta)$ decomposes into the DP posterior over cluster assignments and the cluster-conditional Multinomial-Dirichlet word-occurrence likelihood $p(\boldsymbol{c}_d|\mathbf{w}, \boldsymbol{c}_{-d}, \alpha, \eta) = p(\boldsymbol{c}_d|\boldsymbol{c}_{-d}, \eta) p(\mathbf{w}_d|\mathbf{w}_{-d}, \boldsymbol{c}, \mathbf{z}, \alpha)$ given by

$$P(c_d = k_{\text{old}}|\boldsymbol{c}_{-d}, \alpha, \eta) \propto \underbrace{\left(\frac{m_k^{(-d)}}{m_{\blacksquare}^{(-d)} + \eta}\right)}_{p(\boldsymbol{c}_d|\boldsymbol{c}_{-d},\eta)} \underbrace{\frac{C(\alpha + \vec{\boldsymbol{n}}_k^{(-d)} + \vec{\boldsymbol{n}}_{\blacksquare}^{(d)}))}{C(\alpha + \vec{\boldsymbol{n}}_k^{(-d)})}}_{p(\mathbf{w}_d|\mathbf{w}_{-d},\boldsymbol{c},\mathbf{z},\alpha)}$$

$$P(c_d = k_{\text{new}}|\boldsymbol{c}_{-d}, \alpha, \eta) \propto \frac{\eta}{m_{\blacksquare}^{(-d)} + \eta} \frac{C(\alpha + \vec{\boldsymbol{n}}_{\blacksquare}^{(d)})}{C(\alpha)}$$

where $m_k^{(-d)}$ is the number of occurrences assigned to k not including d, $\vec{\boldsymbol{n}}_k^{(d)}$ is the vector of counts of words from occurrence $\mathbf{w}_d$ assigned to cluster k (i.e. words with $\mathbf{z}_{i,d} = 0$) and $C(\cdot)$ is the normalizing constant for the Dirichlet $C(\boldsymbol{a}) = \Gamma(\sum_{j=1}^m a_j)^{-1} \prod_{j=1}^m \Gamma(a_j)$ operating over vectors of counts $\boldsymbol{a}$.

[7] Effectively, the tiered clustering model is a special case of the *nested* Chinese Restaurant Process with the tree depth fixed to two (Blei et al., 2003).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and Wordnet-based approaches. In *Proc. of NAACL-HLT-09*, pages 19–27.

David J. Aldous. 1985. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117, pages 1–198. Springer, Berlin.

David Blei, Thomas Griffiths, Michael Jordan, and Joshua Tenenbaum. 2003. Hierarchical topic models and the nested Chinese restaurant process. In *Proc. NIPS-2003*.

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

James Richard Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh. College of Science.

Katrin Erk and Sebastian Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP 2008*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proc. of WWW 2001*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of IJCAI-07*, pages 1606–1611.

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

James Gorman and James R. Curran. 2006. Scaling distributional similarity to large corpora. In *Proc. of ACL 2006*.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:2007.

Amaç Herdağdelen and Marco Baroni. 2009. Bagpack: A general framework to represent semantic relations. In *Proc. of GEMS 2009*.

Donald Hindle and Mats Rooth. 1991. Structural ambiguity and lexical relations. In *Proc. of ACL 1991*.

Thomas Landauer and Susan 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.

Martin H. C. Law, Anil K. Jain, and Mário A. T. Figueiredo. 2002. Feature selection in mixture-based clustering. In *Proc. of NIPS 2002*.

Will Lowe. 2001. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 576–581.

Xiaojuan Ma, Jordan Boyd-Graber, Sonya S. Nikolova, and Perry Cook. 2009. Speaking through pictures: Images vs. icons. In *ACM Conference on Computers and Accessibility*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proc. of EMNLP 2007*.

Ulrike Padó. 2007. *The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Sentence Processing*. Ph.D. thesis, Saarland University, Saarbrücken.

Patrick Pantel, Rahul Bhagat, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *In Proceedings of NAACL 2007*.

Patrick Andre Pantel. 2003. *Clustering by committee*. Ph.D. thesis, Edmonton, Alta., Canada.

Lance Parsons, Ehtesham Haque, and Huan Liu. 2004. Subspace clustering for high dimensional data: A review. *SIGKDD Explor. Newsl.*, 6(1).

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proc. of ACL 1993*.

Carl E. Rasmussen. 2000. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*. MIT Press.

Joseph Reisinger and Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proc. of NAACL 2010*.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 52–57. ACL.

Adam N. Sanborn, Thomas L. Griffiths, and Daniel J. Navarro. 2006. A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Patrick Shafto, Charles Kemp, Vikash Mansinghka, Matthew Gordon, and Joshua B. Tenenbaum. 2006. Learning cross-cutting systems of categories. In *Proc. CogSci 2006*.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proc. of ACL 2006*.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Benjamin Van Durme and Marius Paşca. 2008. Finding cars, goddesses and enzymes: Parametrizable acquisition of labeled instances for open-domain information extraction. In *Proc. of AAAI 2008*.

Nianwen Xue, Jinying Chen, and Martha Palmer. 2006. Aligning features with sense distinction dimensions. In *Proc. of COLING/ACL 2006*.