Cross-Cutting Models of Lexical Semantics

Joseph Reisinger
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
joeraii@cs.utexas.edu

Raymond Mooney
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
mooney@cs.utexas.edu

Abstract

Context-dependent word similarity can be measured over multiple cross-cutting dimensions. For example, lung and breath are similar thematically, while authoritative and superficial occur in similar syntactic contexts, but share little semantic similarity. Both of these notions of similarity play a role in determining word meaning, and hence lexical semantic models must take them both into account. Towards this end, we develop a novel model, Multi-View Mixture (MVM), that represents words as multiple overlapping clusterings. MVM finds multiple data partitions based on different subsets of features, subject to the marginal constraint that feature subsets are distributed according to Latent Dirichlet Allocation. Intuitively, this constraint favors feature partitions that have coherent topical semantics. Furthermore, MVM uses soft feature assignment, hence the contribution of each data point to each clustering view is variable, isolating the impact of data only to views where they assign the most features. Through a series of experiments, we demonstrate the utility of MVM as an inductive bias for capturing relations between words that are intuitive to humans, outperforming related models such as Latent Dirichlet Allocation.

1 Introduction

Humans categorize objects using multiple orthogonal taxonomic systems, where category generalization depends critically on what features are relevant to one particular system. For example, foods can be organized in terms of their nutritional value (high in fiber) or situationally (commonly eaten for Thanksgiving; Shafto et al. (2006)). Human knowledgebases such as Wikipedia also exhibit such multiple clustering structure (e.g. people are organized by occupation or by nationality). The effects of these overlapping categorization systems manifest themselves at the lexical semantic level (Murphy, 2002), implying that lexicographical word senses and traditional computational models of word-sense based on clustering or exemplar activation are too impoverished to capture the rich dynamics of word usage.

In this work, we introduce a novel probabilistic clustering method, Multi-View Mixture (MVM), based on cross-cutting categorization (Shafto et al., 2006) that generalizes traditional vector-space or distributional models of lexical semantics (Curran, 2004; Padó and Lapata, 2007; Schütze, 1998; Turney, 2006). Cross-cutting categorization finds multiple feature subsets (categorization systems) that produce high quality clusterings of the data. For example words might be clustered based on their part of speech, or based on their thematic usage. Context-dependent variation in word usage can be accounted for by leveraging multiple latent categorization systems. In particular, cross-cutting models can be used to capture both syntagmatic and paradigmatic notions of word relatedness, breaking up word features into multiple categorization systems and then computing similarity separately for each system.

MVM leverages primitives from Dirichlet-Process Mixture Models (DPMMs) and Latent Dirichlet Allocation (LDA). Each clustering (view) in MVM consists of a distribution over features and data and views are further subdivided into clusters based on a DPMM. View marginal distributions are determined by LDA, allowing data features to be distributed over multiple views, explaining subsets of features.
We evaluate MVM against several other model-based clustering procedures in a series of human evaluation tasks, measuring its ability to find meaningful syntagmatic and paradigmatic structure. We find that MVM finds more semantically and syntactically coherent fine-grained structure, using both common and rare n-gram contexts.

2 Mixture Modeling and Lexical Semantics

_Distributional_, or _vector space_ methods attempt to model word meaning by embedding words in a common metric space, whose dimensions are derived from, e.g., word collocations (Schütze, 1998), syntactic relations (Padó and Lapata, 2007), or latent semantic spaces (Finkelstein et al., 2001; Landauer and Dumais, 1997; Turian et al., 2010). The distributional hypothesis addresses the problem of modeling word similarity (Curran, 2004; Miller and Charles, 1991; Schütze, 1998; Turney, 2006), and can be extended to selectional preference (Resnik, 1997) and lexical substitution (McCarthy and Navigli, 2007) as well. Such methods are highly scalable (Gorman and Curran, 2006) and have been applied in information retrieval (Manning et al., 2008), large-scale taxonomy induction (Snow et al., 2006), and knowledge acquisition (Van Durme and Pașca, 2008).

Vector space models fail to capture the richness of word meaning since similarity is not a globally consistent metric. It violates, e.g., the triangle inequality: the sum of distances from _bat_ to _club_ and _club_ to _association_ is less than the distance from _bat_ to _association_ (Griffiths et al., 2007; Tversky and Gati, 1982).¹ Erk (2007) circumvents this problem by representing words as multiple exemplars derived directly from word occurrences and embedded in a common vector space to capture context-dependent usage. Likewise Reisinger and Mooney (2010) take a similar approach using mixture modeling combined with a background variation model to generate multiple prototype vectors for polysemous words.

Both of these approaches still ultimately embed all words in a single metric space and hence argue for globally consistent metrics that capture human intuitive notions of “similarity.” Rather than assuming a global metric embedding exists, in this work we simply leverage the _cluster assumption_, e.g. that similar words should appear in the same clusters, in particular extending it to multiple clusterings. The cluster assumption is a natural fit for lexical semantics, as partitions can account for metric violations. The end result is a model capable of representing multiple, overlapping similarity metrics that result in disparate valid clusterings leveraging the

**Subspace Hypothesis:** For any pair of words, the set of “active” features governing their apparent similarity differs. For example _wine_ and _bottle_ are similar and _wine_ and _vinegar_ are similar, but it would not be reasonable to expect that the features governing such similarity computations to overlap much, despite occurring in similar documents.

MVM can extract multiple competing notions of similarity, for example both _paradigmatic_, or thematic similarity, and _syntagmatic_ or syntactic similarity, in addition to more fine grained relations.

3 Multi-View Clustering with MVM

As feature dimensionality increases, the number of ways the data can exhibit interesting structure goes up exponentially. Clustering is commonly used to explain data, but often there are several equally valid, competing clusterings, keying off of different subsets of features, especially in high-dimensional settings such as text mining (Niu et al., 2010). For example, company websites can be clustered by sector or by geographic location, with one particular clustering becoming predominant when a majority of features correlate with it. In fact, informative features in one clustering may be noise in another, e.g. the occurrence of _CEO_ is not necessarily discriminative when clustering companies by industry sector, but may be useful in other clusterings. Multiple clustering is one approach to inferring feature subspaces that lead to high quality data partitions. Multiple clustering also improves the flexibility of generative clustering models, as a single model is no longer required to explain all the variance in the feature dimensions (Mansinghka et al., 2009).

¹Similarity also has been shown to violate symmetry (e.g. people have the intuition that _China_ is more similar to _North Korea_ than _North Korea_ is to _China_).
MVM is a multinomial-Dirichlet multiple clustering procedure for distributional lexical semantics that fits multiple, overlapping Dirichlet Process Mixture Models (DPMM) to a set of word clusters. Features are distributed across the set of clusterings (views) using LDA, and each DPMM is fit using a subset of the features. This reduces clustering noise and allows MVM to capture multiple ways in which the data can be partitioned. Figure 1 shows a simple example, and Figure 2 shows a larger sample of feature-view assignments from a 3-view MVM fit to contexts drawn from the Google n-gram corpus.

We implement MVM using generative model primitives drawn from Latent Dirichlet Allocation (LDA) and the Dirichlet Process (DP). \( |M| \) disparate clusterings (views) are inferred jointly from a set of data \( D = \{ \mathbf{w}_d | d \in [1 \cdots D] \} \). Each data vector \( \mathbf{w}_d \) is associated with a probability distribution over views \( \Theta_d^{[M]} \). Empirically, \( \Theta_d^{[M]} \) is represented as a set of feature-view assignments \( \mathbf{z}_d \), sampled via the standard LDA collapsed Gibbs sampler. Hence, each view maintains a separate distribution over features. The generative model for feature-view assignment is given by

\[
\begin{align*}
\Theta_d^{[M]} | \alpha & \sim \text{Dirichlet}(\alpha), \quad d \in D, \\
\phi_m | \beta & \sim \text{Dirichlet}(\beta), \quad m \in [M], \\
z_{d,m} | \Theta_d & \sim \text{Discrete}(\Theta_d), \quad n \in |\mathbf{w}_d|, \\
w_{d,n} | \phi_{z_{d,m}} & \sim \text{Discrete}(\phi_{z_{d,m}}), \quad n \in |\mathbf{w}_d|,
\end{align*}
\]

where \( \alpha \) and \( \beta \) are hyperparameters smoothing the per-document topic distributions and per-topic word distributions respectively.

Conditional on the feature-view assignment \( \{ \mathbf{z} \} \), a clustering is inferred for each view using the Chinese Restaurant Process representation of the DP. The clustering probability is given by

\[
p(c|z, w) \propto p(\{c_m\}, z, w) = \prod_{m=1}^{M} \prod_{d=1}^{|D|} p(w_d^{[z=m]}|c_m, z) p(c_m|z).
\]

where \( p(c_m|z) \) is a prior on the clustering for view \( m \), i.e. the DPMM, and \( p(w_d^{[z=m]}|c_m, z) \) is the likelihood of the clustering \( c_m \) given the data point \( w_d \) restricted to the features assigned to view \( m \):

\[
w_d^{[z=m]} \overset{\text{def}}{=} \{ w_{id}|z_{id} = m \}.
\]

Thus, we treat the \( m \) clusterings \( c_m \) as conditionally independent given the feature-view assignments.

The feature-view assignments \( \{ \mathbf{z} \} \) act as a set of marginal constraints on the multiple clusterings, and the impact that each data point can have on each clustering is limited by the number of features assigned to it. For example, in a two-view model, \( z_{id} = 1 \) might be set for all syntactic features (yielding a syntagmatic clustering) while \( z_{id} = 2 \) is set for document features (paradigmatic clustering).

By allowing the clustering model capacity to vary via the DPMM, MVM can naturally account for the semantic variance of the view. This provides a novel mechanism for handling feature noise: noisy features can be assigned to a separate view with potentially a small number of clusters. This phenomenon is apparent in cluster 1, view 1 in the example in figure 2, where place names and adjectives are clustered together using rare contexts.

From a topic modeling perspective, MVM finds topic refinements within each view, similar to hierarchical methods such as the nested Chinese Restaurant Process (Blei et al., 2003). The main difference is that the features assigned to the second “refined topics” level are constrained by the higher
View 1 cluster 2 and View 3 cluster 1 both contain past-tense verbs, but only overlap on a subset of syntactic features.
level, similar to hierarchical clustering. Unlike hierarchical clustering, however, the top level topics/views form an admixture, allowing individual features from a single data point to be assigned to multiple views.

The most similar model to ours is Cross-cutting categorization (CCC), which fits multiple DPMMS to non-overlapping partitions of features (Mansinghka et al., 2009; Shafto et al., 2006). Unlike MVM, CCC partitions features among multiple DPMMS, hence all occurrences of a particular feature will end up in a single clustering, instead of assigning them softly using LDA. Such hard feature partitioning does not admit an efficient sampling procedure, and hence Shafto et al. (2006) rely on Metropolis-Hastings steps to perform feature assignment, making the model less scalable.

3.1 Word Representation

MVM is trained as a lexical semantic model on Web-scale n-gram and semantic context data. N-gram contexts are drawn from a combination of the Google n-gram and Google books n-gram corpora, with the head word removed: e.g. for the term architect, we collect contexts such as the ___ of the house, an ___ is a, and the ___ of the universe. Semantic contexts are derived from word occurrence in Wikipedia documents: each document a word appears in is added as a potential feature for that word. This co-occurrence matrix is the transpose of the standard bag-of-words document representation.

In this paper we focus on two representations:

1. Syntax-only – Words are represented as bags of ngram contexts derived slot-filling procedure described above.

2. Syntax+Documents – The syntax-only representation is augmented with additional document contexts drawn from Wikipedia.

Models trained on the syntax-only set are only capable of capturing syntagmatic similarity relations, that is, words that tend to appear in similar contexts. In contrast, the syntax+documents set broadens the scope of modelable similarity relations, allowing for paradigmatic similarity (e.g. words that are topically related, but do not necessarily share common syntactic contexts).

Given such word representation data, MVM generates a fixed set of $M$ context views corresponding to dominant eigenvectors in local syntactic or semantic space. Within each view, MVM partitions words into clusters based on each word’s local representation in that view; that is, based on the set of context features it allocates to the view. Words have a non-uniform affinity for each view, and hence may not be present in every clustering (Figure 2). This is important as different ways of drawing distinctions between words do not necessarily apply to all words. In contrast, LDA finds locally consistent collections of contexts but does not further subdivide words into clusters given that set of contexts. Hence, it may miss more fine-grained structure, even with increased model complexity.

4 Experimental Setup

4.1 Corpora

We derive word features from three corpora: (1) the English Google Web n-gram corpus, containing n-gram contexts up to 5-gram that occur more than 40 times in a 1T word corpus of Web text, (2) the English Google Books n-gram corpus, consisting of n-gram contexts up to 5-gram that occur more than 40 times in a 500B word corpus of books, and (3) a snapshot of the English Wikipedia taken on October 11, 2010 containing over 3M articles.

MVM is trained on a sample of 20k English words drawn uniformly at random from the top 200k English terms appearing in Wikipedia (different parts of speech were sampled from the Google n-gram corpus according to their observed frequency). Two versions of the syntax-only dataset are created from different subsets of the Google n-gram corpora: (1) the common subset contains all syntactic contexts appearing more than 200 times in the combined corpus, and (2) the rare subset, containing only contexts that appear 50 times or fewer.

4.2 Human Evaluation

Our main goal in this work is to find models that capture aspects of the syntactic and semantic organization of word in text that are intuitive to humans.

---

2http://ngrams.googlelabs.com/datasets
3http://wikipedia.org
According to the use theory of meaning, lexical semantic knowledge is equivalent to knowing the contexts that words appear in, and hence being able to form reasonable hypotheses about the relatedness of syntactic contexts.

Vector space models are commonly evaluated by comparing their similarity predictions to a nominal set of human similarity judgments (Curran, 2004; Padó and Lapata, 2007; Schütze, 1998; Turney, 2006). In this work, since we are evaluating models that potentially yield many different similarity scores, we take a different approach, scoring clusters on their semantic and syntactic coherence using a set intrusion task (Chang et al., 2009).

In set intrusion, human raters are shown a set of options from a coherent group and asked to identify a single intruder drawn from a different group. We extend intrusion to three different lexical semantic tasks: (1) context intrusion, where the top contexts from each cluster are used, (3) document intrusion, where the top document contexts from each cluster are used, and (2) word intrusion, where the top words from each cluster are used. For each cluster, the top four contexts/words are selected and appended with another context/word from a different cluster. The resulting set is then shuffled, and the human raters are asked to identify the intruder, after being given a short introduction (with common examples) to the task. Table 1 shows sample questions of varying degrees of difficulty. As the semantic coherence and distinctness from other clusters increases, this task becomes easier.

Set intrusion is a more robust way to account for human similarity judgments than asking directly for a numeric score (e.g., the Miller and Charles (1991) set) as less calibration is required across raters. Furthermore, the additional cluster context significantly reduces the variability of responses.

Human raters were recruited from Amazon’s Mechanical Turk. A total of 1256 raters completed 30438 evaluations for 5780 unique intrusion tasks (5 evaluations per task). 2736 potentially fraudulent evaluations from 11 raters were rejected. Table 3 summarizes inter-annotator agreement. Overall we found \( \kappa \approx 0.4 \) for most tasks; a set of comments about the task difficulty is given in Table 2, drawn from an anonymous public message board.

5 Results
We trained DPMM, LDA and MVM models on the syntax-only and syntax+documents data across a wide range of settings for \( M \in \{3, 5, 7, 10, 20, 30, 50, 100, 200, 300, 500, 1000\} \).

Table 1: Example questions from the three intrusion tasks, in order of difficulty (left to right, easy to hard; computed from inter-annotator agreement). Italics show intruder items.
Figure 3: Average scores for each model broken down by parameterization and data source. Error bars depict 95% confidence intervals. X-axis labels show Model-views-α-β. Dots show average rater scores; bar-charts show standard quantile ranges and median score.
I just tried 30 of the what doesn’t belong ones. They took about 30 seconds each due to thinking time so not worth it for me.

I don’t understand the fill in the blank ones to be honest. I just kinda pick one, since I don’t know what’s expected lol

Your not filling in the blank just ignore the blank and think about how the words they show relate to each other and choose the one that relates least. Some have just words and no blanks.

These seem very subjective to mw. I hope there isn’t definite correct answers because some of them make me go [emoticon of head-scratching]

I looked and have no idea. I guess I’m a word idiot because I don’t see the relation between the words in the preview HIT - too scared to try any of these.

I didn’t dive in but I did more than I should have they were just too easy. Most of them I could tell what did not belong, some were pretty iffy though.

Table 2: Sample of comments about the task taken verbatim from a public Mechanical Turk user message board (TurkerNation). Overall the raters report the task to be difficult, but engaging.

α ∈ {0.1, 0.01}, and β ∈ {0.1, 0.05, 0.01} in order to understand how they perform relatively on the intrusion tasks and also how sensitive they are to various parameter settings. Models were run until convergence, defined as no increase in log-likelihood on the training set for 100 Gibbs samples. Average runtimes varied from a few hours to a few days, depending on the number of clusters or topics. There is little computational overhead for MVM compared to LDA or DPMM with a similar number of clusters.

Overall, MVM significantly outperforms both LDA and DPMM (measured as % of intruders correctly identified) as the number of clusters increases. Coarse-grained lexical semantic distinctions are easy for humans to make, and hence models with fewer clusters tend to outperform models with more clusters. Since high granularity predictions are more

...
<table>
<thead>
<tr>
<th>Model</th>
<th>Syntax</th>
<th>Syntax+Documents</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPMM</td>
<td>0.30</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td>LDA</td>
<td>0.33</td>
<td>0.39</td>
<td>0.35</td>
</tr>
<tr>
<td>MVM</td>
<td>0.44</td>
<td>0.49</td>
<td>0.46</td>
</tr>
<tr>
<td>Overall</td>
<td>0.37</td>
<td>0.43</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 3: Fleiss’ $\kappa$ scores for various model and data combinations. Results from MVM have higher $\kappa$ scores than LDA or DPMM; likewise Syntax+Documents data yields higher agreement, primarily due to the relative ease of the document intrusion task.

significantly more granular. Figure 4(a) breaks out model performance by model complexity, demonstrating that MVM has a significant edge over LDA as model complexity increases.

For rare n-gram contexts, we obtain similar results, with MVM scores being less variable across model parameterizations and complexity (Figure 3(b)). In general, LDA performance degrades faster as model complexity increases for rare contexts, due to the increased data sparsity (Figure 4(b)). For context intrusion, DPMM, LDA, and MVM average 45.9%, 36.1% and 50.9% accuracy respectively; for word intrusion, DPMM, LDA, and MVM average 67.4%, 45.6% and 67.9% accuracy; MVM performance does not differ significantly from DPMM, but both outperform LDA. Average cluster sizes are more uniform across model types for rare contexts: 384.0 for DPMM, 358.3 for LDA and 391 for MVM.

Human performance on the context intrusion task is significantly more variable than on the word-intrusion task, reflecting the additional complexity.

In all models, there is a high correlation between rater scores and per-cluster likelihood, indicating that model confidence reflects noise in the data.

### 5.2 Syntax+Documents Model

With the syntax+documents training set, MVM significantly outperforms LDA across a wide range of model settings. MVM also outperforms DPMM for word and document intrusion. For context intrusion, DPMM, LDA, and MVM average 68.0%, 51.3% and 66.9% respectively;\(^8\) for word intrusion, DPMM, LDA, and MVM average 56.3%, 64.0% and 74.9% respectively; for document intrusion, DPMM, LDA, and MVM average 41.5%, 49.7% and 60.6% respectively. Qualitatively, models trained on syntax+document yield a higher degree of paradigmatic clusters which have intuitive thematic structure. Performance on document intrusion is significantly lower and more variable, reflecting the higher degree of world knowledge required. As with the previous data set, performance of MVM models trained on syntax+documents data degrades less slowly as the cluster granularity increases (Figure 5).

One interesting question is to what degree MVM views partition syntax and document features versus LDA topics. That is, to what degree do the MVM views capture purely syntagmatic or purely paradigmatic variation? We measured view entropy for all three models, treating syntactic features and document features as different class labels. MVM with $M = 50$ views obtained an entropy score of 0.045, while LDA with $M = 50$ obtained 0.073, and the best DPMM model 0.082.\(^9\) Thus MVM views may indeed capture pure syntactic or thematic clusterings.

---

\(^8\)High DPMM accuracy is driven by the low number of clusters: 46.5 for DPMM vs. 358.3 for LDA and 725.6 for MVM.

\(^9\)The low entropy scores reflect the higher percentage of syntactic contexts overall.
5.3 Discussion

As cluster granularity increases, we find that MVM accounts for feature noise better than either LDA or DPMM, yielding more coherent clusters. (Chang et al., 2009) note that LDA performance degrades significantly on a related task as the number of topics increases, reflecting the increasing difficulty for humans in grasping the connection between terms in the same topic. This suggests that as topics become more refined in models with larger numbers of topics, they are less useful for humans. In this work, we find that although MVM and LDA perform similarity on average, MVM clusters are significantly more interpretable than LDA clusters as the granularity increases (Figures 4 and 5). We argue that models capable of making such fine-grained semantic distinctions are more desirable.

The results presented in the previous two sections hold both for unbiased cluster selection (e.g. where clusters are drawn uniformly at random from the model) and when cluster selection is biased based on model probability (results shown). Biased selection potentially gives an advantage to MVM, which generates many more small clusters than either LDA or DPMM, helping it account for noise.

6 Future Work

Models based on cross-cutting categorization is a novel approach to lexical semantics and hence should be evaluated on standard baseline tasks, e.g. contextual paraphrase or lexical substitution (McCarthy andNavigli, 2007). Additional areas for future work include:

(Latent Relation Modeling) Clusterings formed from feature partitions in MVM can be viewed as a form of implicit relation extraction; that is, instead of relying on explicit surface patterns in text, relations between words or concepts are identified indirectly based on common syntactic patterns. For example, clusterings that divide cities by geography or clusterings partition adjectives by their polarity.

(Latent Semantic Language Modeling) Generative models such as MVM can be used to build better priors for class-based language modeling (Brown et al., 1992). The rare n-gram results demonstrate that MVM is potentially useful for tail contexts; i.e. inferring tail probabilities from low counts.

(Explicit Feature Selection) In this work we rely on smoothing to reduce the noise of over-broad extraction rather than performing feature selection explicitly. All of the models in this paper can be combined with feature selection methods to remove noisy features, and it would be particularly interesting to draw parallels to models of “clutter” in vision.

(Hierarchical Cross-Categorization) Human concept organization consists of multiple overlapping local ontologies, similar to the loose ontological structure of Wikipedia. Furthermore, each ontological system has a different set of salient properties. It would be interesting to extend MVM to model hierarchy explicitly, and compare against baselines such as Brown clustering (Brown et al., 1992), the nested Chinese Restaurant Process (Blei et al., 2003) and the hierarchical Pachinko Allocation Model (Mimno et al., 2007).

7 Conclusion

This paper introduced MVM, a novel approach to modeling lexical semantic organization using multiple cross-cutting clusterings capable of capturing multiple lexical similarity relations jointly in the same model. In addition to robustly handling homonymy and polysemy, MVM naturally captures both syntagmatic and paradigmatic notions of word similarity. MVM performs favorably compared to other generative lexical semantic models on a set of human evaluations, over a wide range of model settings and textual data sources.

Acknowledgements

We would like to thank the anonymous reviewers for their extensive comments. This work was supported by a Google PhD Fellowship to the first author.

References


David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In ICML.


