# Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering

**Andreas Vlachos**
Computer Laboratory
University of Cambridge
Cambridge CB3 0FD, UK
av308l@cl.cam.ac.uk

**Anna Korhonen**
Computer Laboratory
University of Cambridge
Cambridge CB3 0FD, UK
alk23@cl.cam.ac.uk

**Zoubin Ghahramani**
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK
zoubin@eng.cam.ac.uk

## Abstract

In this work, we apply Dirichlet Process Mixture Models (DPMMs) to a learning task in natural language processing (NLP): lexical-semantic verb clustering. We thoroughly evaluate a method of guiding DPMMs towards a particular clustering solution using pairwise constraints. The quantitative and qualitative evaluation performed highlights the benefits of both standard and constrained DPMMs compared to previously used approaches. In addition, it sheds light on the use of evaluation measures and their practical application.

## 1 Introduction

Bayesian non-parametric models have received a lot of attention in the machine learning community. These models have the attractive property that the number of components used to model the data is not fixed in advance but is actually determined by the model and the data. This property is particularly interesting for NLP where many tasks are aimed at discovering novel, previously unknown information in corpora. Recent work has applied Bayesian non-parametric models to anaphora resolution (Haghighi and Klein, 2007), lexical acquisition (Goldwater, 2007) and language modeling (Teh, 2006) with good results.

Recently, Vlachos et al. (2008) applied the basic models of this class, Dirichlet Process Mixture Models (DPMMs) (Neal, 2000), to a typical learning task in NLP: lexical-semantic verb clustering. The task involves discovering classes of verbs similar in terms of their syntactic-semantic properties (e.g. MOTION class for *travel, walk, run*, etc.). Such classes can provide important support for other NLP tasks, such as word sense disambiguation, parsing and semantic role labeling (Dang, 2004; Swier and Stevenson, 2004). Although some fixed classifications are available (e.g. VerbNet (Kipper-Schuler, 2005)) these are not comprehensive and are inadequate for specific domains (Korhonen et al., 2006b).

Unlike the clustering algorithms applied to this task before, DPMMs do not require the number of clusters as input. This is important because even if the number of classes in a particular task was known (e.g. in the context of a carefully controlled experiment), a particular dataset may not contain instances for all the classes. Moreover, each class is not necessarily contained in one cluster exclusively, since the target classes are defined manually without taking into account the feature representation used. The fact that DPMMs do not require the number of target clusters in advance, renders them promising for the many NLP tasks where clustering is used for learning purposes.

While the results of Vlachos et al. (2008) are promising, the use of a clustering approach which discovers the number of clusters in data presents a new challenge to existing evaluation measures. In this work, we investigate optimal evaluation for such approaches, using the dataset and the basic method of Vlachos et al. as a starting point. We review the applicability of existing evaluation measures and propose a modified version of the newly introduced V-measure (Rosenberg and Hirschberg, 2007). We complement the quantitative evaluation with thorough qualitative assessment, for which we introduce a method to summarize samples obtained from a clustering algorithm.

In preliminary work by Vlachos et al. (2008), a constrained version of DPMMs which takes advantage of *must-link* and *cannot-link* pairwise constraints was introduced. It was demonstrated how such constraines can guide the clustering solution towards some prior intuition or considerations relevant to the specific NLP application in mind. We explain the inference algorithm for the constrained DPMM in greater detail and evaluate quantita-

*Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, pages 74–82,
Athens, Greece, 31 March 2009. 

tively the contribution of each constraint type of independently, complementing it with qualitative analysis. The latter demonstrates how the pairwise constraints added affects instances beyond those involved directly. Finally, we discuss how the unsupervised and the constrained version of DPMMs can be used in a real-world setup.

The results from our comprehensive evaluation show that both versions of DPMMs are capable of learning novel information not in the gold standard, and that the constrained version is more accurate than a previous verb clustering approach which requires setting the number of clusters in advance and is therefore less realistic.

## 2 Unsupervised clustering with DPMMs

With DPMMs, as with other Bayesian non-parametric models, the number of mixture components is not fixed in advance, but is determined by the model and the data. The parameters of each component are generated by a Dirichlet Process (DP) which can be seen as a distribution over the parameters of other distributions. In turn, each instance is generated by the chosen component given the parameters defined in the previous step:

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_i|G &\sim G \\ x_i|\theta_i &\sim F(\theta_i) \end{aligned} \quad (1)$$

In Eq. 1, $G_0$ and $G$ are probability distributions over the component parameters ($\theta$), and $\alpha > 0$ is the concentration parameter which determines the variance of the Dirichlet process. We can think of $G$ as a randomly drawn probability distribution with mean $G_0$. Intuitively, the larger $\alpha$ is, the more similar $G$ will be to $G_0$. Instance $x_i$ is generated by distribution $F$, parameterized by $\theta_i$. The graphical model is depicted in Figure 1.

The prior probability of assigning an instance to a particular component is proportionate to the number of instances already assigned to it ($n_{-i,z}$). In other words, DPMMs exhibit the "rich get richer" property. In addition, the probability that a new cluster is created is dependent on the concentration parameter $\alpha$. A popular metaphor to describe DPMMs which exhibits an equivalent clustering property is the Chinese Restaurant Process (CRP). Customers (instances) arrive at a Chinese restaurant which has an infinite number of tables (components). Each customer sits at one of the tables that is either occupied or vacant with popular tables attracting more customers.

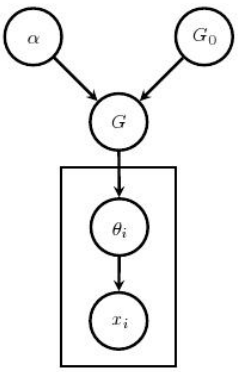


Figure 1: Graphical representation of DPMMs.

In this work, the distribution used to model the components is the multinomial and the prior used is the Dirichlet distribution ($F$ and $G_0$ in Eq. 1). The conjugacy between them allows for the analytic integration over the component parameters. Following Neal (2000), the component assignments $z_i$ are sampled using the following scheme:

$$\begin{aligned} P(z_i = z|z_{-i}, x_i) \propto \\ p(z_i = z|z_{-i}) DirM(x_i|z_i = z, x_{-i,z}, \lambda) \end{aligned} \quad (2)$$

In Eq. 2 $DirM$ is the Dirichlet-Multinomial distribution, $\lambda$ are the parameters of the Dirichlet prior $G_0$ and $x_{-i,z}$ are the instances assigned already to component $z$ (none if we are sampling the probability of assignment to a new component). This sampling scheme is possible due to the fact that the instances in the model are exchangeable, i.e. the order in which they are generated is not relevant.

In terms of the CRP metaphor, we consider each instance $x_i$ as the last customer to arrive and he chooses to sit together with other customers at an existing table or to sit at a new table. Following Navarro et al. (2006) who used the same model to analyze individual differences, we sample the concentration parameter $\alpha$ using the inverse Gamma distribution as a prior.

## 3 Evaluation measures

The evaluation of unsupervised clustering against a gold standard is not straightforward because the clusters found are not explicitly labelled. Formally defined, an unsupervised clustering algorithm partitions a set of instances $X = \{x_i|i = 1, ..., N\}$ into a set of clusters $K = \{k_j|j = 1, ..., |K|\}$. The standard approach to evaluate the quality of the clusters is to use an external gold standard in which the instances are partitioned into a set of

classes $C = \{c_l | l = 1, ..., |C|\}$. Given this, the goal is to find a partitioning of the instances $K$ that is as close as possible to the gold standard $C$.

Most work on verb clustering has used the F-measure or the Rand Index (RI) (Rand, 1971) for evaluation, which rely on counting pairwise links between instances. However, Rosenberg and Hirschberg (2007) pointed out that F-measure assumes (the missing) mapping between $c_l$ and $k_j$. In practice, RI values concentrate in a small interval near 100% (Meilă, 2007).

Rosenberg & Hirschberg (2007) proposed an information-theoretic metric: V-measure. V-measure is the harmonic mean of homogeneity and completeness which evaluate the quality of the clustering in a complementary way. Homogeneity assesses the degree to which each cluster contains instances from a single class of $C$. This is computed as the conditional entropy of the class distribution of the gold standard given the clustering discovered by the algorithm, $H(C|K)$, normalized by the entropy of the class distribution in the gold standard, $H(C)$. Completeness assesses the degree to which each class is contained in a single cluster. This is computed as the conditional entropy of the cluster distribution discovered by the algorithm given the class, $H(K|C)$, normalized by the entropy of the cluster distribution, $H(K)$. In both cases, we subtract the resulting ratios from 1 to associate higher scores with better solutions:

$$h = 1 - \frac{H(C|K)}{H(C)}$$
$$c = 1 - \frac{H(K|C)}{H(K)}$$
$$V_\beta = \frac{(1+\beta) * h * c}{(\beta * h) + c} \quad (3)$$

The parameter $\beta$ in Eq. 3 regulates the balance between homogeneity and completeness. Rosenberg & Hirschberg set it to 1 in order to obtain the harmonic mean of these qualities. They also note that V-measure favors clustering solutions with a large number of clusters (large $|K|$), since such solutions can achieve very high homogeneity while maintaining reasonable completeness. This effect is more prominent when a dataset includes a small number of instaces for gold standard classes. While increasing $|K|$ does not guarantee an increase in V-measure (splitting homogeneous clusters would reduce completeness without improving homogeneity), it is easier to achieve higher scores when more clusters are produced.

Another relevant measure is the Variation of Information (VI) (Meilă, 2007). Like V-measure, it assesses homogeneity and completeness using the quantities $H(C|K)$ and $H(K|C)$ respectively, however it simply adds them up to obtain a final result (higher scores are worse). It is also a metric, i.e. VI scores can be added, subtracted, etc, since the quantities involved are measured in bits. However, it can be observed that if $|C|$ and $|K|$ are very different then the terms $H(C|K)$ and $H(K|C)$ will not necessarily be in the same range. In particular, if $|K| \ll |C|$ then $H(K|C)$ (and $VI$) will be low. In addition, VI scores are not normalized and therefore their interpretation is difficult.

Both V-measure and VI have important advantages over RI and F-measure: they do not assume a mapping between classes and clusters and their scores depend only on the relative sizes of the clusters. However, V-measure and VI can be misleading if the number of clusters found ($|K|$) is substantially different than the number of gold standard classes ($|C|$). In order to ameliorate this, we suggest to take advantage of the $\beta$ parameter in Eq. 3 in order to balance homogeneity and completeness. More specifically, setting $\beta = |K|/|C|$ assigns more weight to completeness than to homogeneity in case $|K| > |C|$ since the former is harder to achieve and the latter is easier when the clustering solution has more clusters than the gold standard has classes. The opposite occurs when $|K| < |C|$. In case $|K| = |C|$ the score is the same as the original V-measure. Achieving 100% score according to any of these measures requires correct prediction of the number of clusters.

In this work, we evaluate our results using the three measures described above (V-measure, VI, V-beta). We complement this evaluation with qualitative evaluation which assesses the potential of DPMMs to discover novel information that might not be included in the gold standard.

## 4 Experiments

To perform lexical-semantic verb clustering we used the dataset of Sun et al. (2008). It contains 204 verbs belonging to 17 fine-grained classes in Levin's (1993) taxonomy so that each class contains 12 verbs. The classes and their verbs were selected randomly. The features for each verb are its subcategorization frames (SCFs) and associated frequencies in corpus data, which capture the

| | DPMM | Sun et al. |
|---|---|---|
| no. of clusters | 37.79 | 17 |
| homogeneity | 60.23% | 57.57% |
| completeness | 55.82% | 60.19% |
| V-measure | 57.94% | 58.85% |
| V-beta | 57.11% | 58.85% |
| VI (bits) | 3.5746 | 3.3598 |

Table 1: Clustering performances.

syntactic context in which the verb occurs. SCFs were extracted from the publicly available VALEX lexicon (Korhonen et al., 2006a). VALEX was acquired automatically using a domain-independent statistical parsing toolkit, RASP (Briscoe and Carroll, 2002), and a classifier which identifies verbal SCFs. As a consequence, it includes some noise due to standard text processing and parsing errors and due to the subtlety of argument-adjunct distinction. In our experiments, we used the SCFs obtained from VALEX1, parameterized for the prepositional frame, which had the best performance in the experiments of Sun et al. (2008).

The feature sets based on verbal SCFs are very sparse and the counts vary over a large range of values. This can be problematic for generative models like DPMMs, since a few dominant features can mislead the model. To reduce the sparsity, we applied non-negative matrix factorization (NMF) (Lin, 2007) which decomposes the dataset in two dense matrices with non-negative values. It has proven useful in a variety of tasks, e.g. information retrieval (Xu et al., 2003) and image processing (Lee and Seung, 1999).

We use a symmetric Dirichlet prior with parameters of 1 ($\lambda$ in Equation 2). The number of dimensions obtained using NMF was 35. We run the Gibbs sampler 5 times, using 100 iterations for burn-in and draw 20 samples from each run with 5 iterations lag between samples. Table 1 shows the average performances. The DPMM discovers 37.79 verb clusters on average with its performance ranging between 53% and 58% depending on the evaluation measure used. Homogeneity is 4.5% higher than completeness, which is expected since the number of classes in the gold standard is 17. The fact that the DPMM discovers more than twice the number of classes is reflected in the difference between the V-measure and V-beta, the latter being lower. In the same table, we show the results of Sun et al. (2008), who used pairwise clustering (PC) (Puzicha et al., 2000) which involves determining the number of clusters in advance.

The performance of the DPMM is 1%-3% lower than that of Sun et al. As expected, the difference in V-measure is smaller since the DPMM discovers a larger number of clusters, while for VI it is larger. The slightly better performance of PC can be attributed to two factors. First, the (correct) number of clusters is given as input to the PC algorithm and not discovered like by the DPMM. Secondly, PC uses the similarities between the instances to perform the clustering, while the DPMM attempts to find the parameters of the process that generated the data, which is a different and typically a harder task. In addition, the DPMM has two clear advantages which we illustrate in the following sections: it can be used to discover novel information and it can be modified to incorporate intuitive human supervision.

## 5 Qualitative evaluation

The gold standard employed in this work (Sun et al., 2008) is not fully accurate or comprehensive. It classifies verbs according to their predominant senses in the fairly small SemCor data. Individual classes are relatively coarse-grained in terms of syntactic-semantic analysis[1] and they capture some of the meaning components only. In addition, the gold standard does not capture the semantic relatedness of distinct classes. In fact, the main goal of clustering is to improve such existing classifications with novel information and to create classifications for new domains. We performed qualitative analysis to investigate the extent to which the DPMM meets this goal.

We prepared the data for qualitative analysis as follows: We represented each clustering sample as a linking matrix between the instances of the dataset and measured the frequency of each pair of instances occurring in the same cluster. We constructed a partial clustering of the instances using only those links that occur with frequency higher than a threshold $prob\_link$. Singleton clusters were formed by considering instances that are not linked with any other instances more frequently than a threshold $prob\_single$. The lower the $prob\_link$ threshold, the larger the clusters will be, since more instances get linked. Note that including more links in the solution can either in-

[1] Many original Levin classes have been manually refined in VerbNet.

crease the number of clusters when instances involved were not linked otherwise, or decrease it when linking instances that already belong to other clusters. The higher the $prob\_single$ threshold, the more instances will end up as singletons. By adjusting these two thresholds we can affect the coverage of the analysis. This approach was chosen because it enables to conduct qualitative analysis of data relevant to most clustering samples and irrespective of individual samples. It can also be useful in order to use the output of the clustering algorithm as a component in a pipeline which requires a single result rather than multiple samples.

Using this method, we generated data sets for qualitative analysis using 4 sets of values for $prob\_link$ and $prob\_single$, respectively: (99%, 1%), (95%, 5%), (90%, 10%) and (85%, 15%). Table 1 shows the number of a) verbs, b) clusters (2 or more instances) and c) singletons in each resulting data set, along with the percentage and size of the clusters which represent 1, 2, or multiple gold standard classes. As expected, higher threshold values produce high precision clusters for a smaller set of verbs (e.g. (99%,1%) produces 5 singletons and assigns 70 verbs to 20 clusters, 55% of which represent a single gold standard class), while less extreme threshold values yield higher recall clusters for a larger set of verbs (e.g. (85%,15%) produces 10 singletons and assigns 140 verbs to 25 clusters, 20% of which contain verbs from several gold standard classes).

We conducted the qualitative analysis by comparing the four data sets against the gold standard, SCF distributions, and WordNet (Fellbaum, 1998) senses for each test verb. We first analysed the 5-10 singletons in data sets and discovered that while 3 of the verbs resist classification because of syntactic idiosyncrasy (e.g. *unite* takes intransitive SCFs with frequency higher than other members of class 22.2), the majority of them (7) end up in singletons for valid semantic reasons: taking several frequent WordNet senses they are "too polysemous" to be realistically clustered according to their predominant sense (e.g. *get* and *look*).

We then examined the clusters, and discovered that even in the data set created with the lowest $prob\_link$ threshold of 85%, almost half of the "errors" are in fact novel semantic patterns discovered by clustering. Many of these could be new sub-classes of existing gold standard classes. For example, looking at the 13 high accuracy clusters which correspond to a single gold standard class each, they only represent 9 gold standard classes because as many as 4 classes been divided into two clusters, suggesting that the gold standard is too coarse-grained. Interestingly, each such subdivision seems semantically justified (e.g. the 11.1 PUT verbs *bury* and *immerse* appear in a different cluster than the semantically slightly different *place* and *situate*).

In addition, the DPMM discovers semantically similar gold standard classes. For example, in the data set created with the $prob\_link$ threshold of 99%, 6 of the clusters include members from 2 different gold standard classes. 2 occur due to syntactic idiosyncrasy, but the majority (4) occur because of true semantic relatedness (e.g. the clustering relates 22.2 AMALGAMATE and 36.1 CORRESPOND classes which share similar meaning components). Similarly, in the data set produced by the $prob\_link$ threshold of 85%, one of the largest clusters includes 26 verbs from 5 gold standard classes. The majority of them belong to 3 classes which are related by the meaning component of "motion": 43.1 LIGHT EMISSION, 47.3 MODES OF BEING INVOLVING MOTION, and 51.3.2 RUN verbs:

- **class 22.2** AMALGAMATE: *overlap*
- **class 36.1** CORRESPOND: *banter, concur, dissent, haggle*
- **class 43.1** LIGHT EMISSION: *flare, flicker, gleam, glisten, glow, shine, sparkle*
- **class 47.3** MODES OF BEING INVOLVING MOTION: *falter, flutter, quiver, swirl, wobble*
- **class 51.3.2** RUN: *fly, gallop, glide, jog, march, stroll, swim, travel, trot*

Thus many of the singletons and the clusters in the different outputs capture finer or coarser-grained lexical-semantic differences than those captured in the gold standard. It is encouraging that this happens despite us focussing on a relatively small set of 204 verbs and 17 classes only.

## 6 Constrained DPMMs

While the ability to discover novel information is attractive in NLP, in many cases it is also desirable to influence the solution with respect to some prior intuition or consideration relevant to the application in mind. For example, while discovering finer-grained classes than those included in the gold standard is useful for some applications, others may benefit from a coarser clustering or a clustering that reveals a specific aspect of the dataset.

| THR | verbs | clusters | singletons | % and size of clusters containing | | |
|---|---|---|---|---|---|---|
| | | | | 1 class | 2 classes | multiple classes |
| 99%,1% | 70 | 20 | 5 | 55% (3.0) | 30% (2.8) | 15% (4.5) |
| 95%,5% | 104 | 25 | 9 | 40% (3.7) | 44% (2.8) | 16% (6.8) |
| 90%,10% | 128 | 28 | 9 | 46% (3.4) | 39% (2.5) | 14% (11.0) |
| 85%,15% | 140 | 25 | 10 | 44% (3.7) | 28% (3.3) | 20% (13.0) |

Table 2: An overview of the data sets generated for qualitative analysis

Preliminary work by Vlachos et al. (2008) introduced a constrained version of DPMMs that enables human supervision to guide the clustering solution when needed. We model the human supervision as pairwise constraints over instances, following Wagstaff & Cardie (2000): given a pair of instances, they are either linked together (*must-link*) or not (*cannot-link*). For example, *charge* and *run* should form a *must-link* if the aim is to cluster 51.3 MOTION verbs together, but they should form a *cannot-link* if we are interested in 54.5 BILL verbs. In the discussion and the experiments that follow, we assume that all links are consistent with each other. This information can be obtained by asking human experts to label links, or by extracting it from extant lexical resources. Specifying the relations between the instances results in a partial labeling of the instances. Such labeling is likely to be re-usable, since relations between the instances are likely to be useful for a wider range of tasks which might not have identical labels but could still have similar relations.

In order to incorporate the constraints in the DPMM, we modify the underlying generative process to take them into account. In particular *must-linked* instances are generated by the same component and *cannot-linked* instances always by different ones. In terms of the CRP metaphor, customers connected with *must-links* arrive at the restaurant together and choose a table jointly, respecting their *cannot-links* with other customers. They get seated at the same table successively one after the other. Customers without *must-links* with others choose tables avoiding their *cannot-links*.

In order to sample the component assignments according to this model, we restrict the Gibbs sampler to take them into account using the sampling scheme of Fig. 2. First we identify *linked-groups* of instances, taking into account transitivity[2]. We then sample the component assignments only from distributions that respect the links provided. More specifically, for each instance that does not belong to a *linked-group*, we restrict the sampler to choose components that do not contain instances *cannot-linked* with it. For instances in a *linked-group*, we sample their assignment jointly, again taking into account their *cannot-links*. This is performed by adding each instance of the *linked-group* successively to the same component. In Fig. 2, $\mathcal{C}_i$ are the *cannot-links* for instance(s) $i$, $\ell$ are the indices of the instances in a *linked-group*, and $z_{<i}$ and $x_{<i}$ are the assignments and the instances of a *linked-group* that have been assigned to a component before instance $i$.

**Input:** data $\mathcal{X}$, *must-links* $\mathcal{M}$, *cannot-links* $\mathcal{C}$
$linked\_groups$ = find_linked_groups($\mathcal{X}$, $\mathcal{M}$)
Initialize $Z$ according to $\mathcal{M}$, $\mathcal{C}$
**for** $i$ **not in** $linked\_groups$
  **for** $z = 1$ **to** $|Z| + 1$
    **if** $x_{-i,z} \cap \mathcal{C}_i = \emptyset$
      $P(z_i = z|z_{-i}, x_i)$ (Eq. 2)
    **else**
      $P(z_i = z|z_{-i}, x_i) = 0$
  Sample from $P(z_i)$
**for** $\ell$ **in** $linked\_groups$
  **for** $z = 1$ **to** $|Z| + 1$
    **if** $x_{-\ell,z} \cap \mathcal{C}_\ell = \emptyset$
      Set $P(z_\ell = z|z_{-\ell}, x_\ell) = 1$
      **for** $i$ **in** $\ell$
        $P(z_\ell{=}\, z|z_{-\ell}, x_\ell)* =$
          $P(z_i = z|z_{-\ell}, x_{-\ell,z}, z_{<i}, x_{<i})$
    **else**
      $P(z_\ell = z|z_{-\ell}, x_\ell) = 0$
  Sample from $P(z_\ell)$

Figure 2: Gibbs sampler incorporating *must-links* and *cannot-links*.

## 7 Experiments using constraints

To investigate the impact of pairwise constraints on clustering by the DPMM, we conduct exper-

[2]If A is linked to B and B to C, then A is linked to C.

iments in which the links are sampled randomly from the gold standard. The number of links varied from 10 to 50 and the random choice was repeated 5 times without checking for redundancy due to transitivity. All the other experimental settings are identical to those in Section 4. Following Wagstaff & Cardie (2000), in Table 3 we show the impact of each link type independently (labeled "must" and "cannot" accordingly), as well as when mixed in equal proportions ("mix").

Adding randomly selected pairwise links is beneficial. In particular, *must-links* improve the clustering rapidly. Incorporating 50 *must-links* improves the performance by 7-8% according to the evaluation measures. In addition, it reduces the average number of clusters by approximately 4. The *cannot-links* are rather ineffective, which is expected as the clustering discovered by the unsupervised DPMM is more fine-grained than the gold standard. For the same reason, it is more likely that the randomly selected *cannot-links* are already discovered by the DPMM and are thus redundant. Wagstaff & Cardie also noted that the impact of the two types of links tends to vary across data sets. Nevertheless, a minor improvement is observed in terms of homogeneity. The balanced mix improves the performance, but less rapidly than the *must-links*.

In order to assess how the links added help the DPMM learn other links we use the Constrained Rand Index (CRI), which is a modification of the Rand Index that takes into account only the pairwise decisions that are not dictated by the constraints added (Wagstaff and Cardie, 2000; Klein et al., 2002). We evaluate the constrained DPMM with CRI (Table 3, bottom right graph) and our results show that the improvements obtained using pairwise constraints are due to learning links beyond the ones enforced.

In a real-world setting, obtaining the mixed set of links is equivalent to asking a human expert to give examples of verbs that should be clustered together or not. Such information could be extracted from a lexical resource (e.g. ontology). Alternatively, the DPMM could be run without any constraints first and if a human expert judges the clustering too coarse (or fine) then *cannot-links* (or *must-links*) could help, since they can adapt the clustering rapidly. When 20 randomly selected *must-links* are integrated, the DPMM reaches or exceeds the performance of PC used by Sun et al. (2008) according to all the evaluation measures. We also argue that it is more realistic to guide the clustering algorithm using pairwise constraints than by defining the number of clusters in advance. Instead of using pairwise constraints to affect the clustering solution, one could alter the parameters for the Dirichlet prior $G_0$ (Eq. 1) or experiment with varying concentration parameter values. However, it is difficult to predict in advance the exact effect such changes would have in the solution discovered.

Finally, we conducted qualitative analysis of the samples obtained constraining the DPMM with 10 randomly selected *must-links*. We first prepared the data according to the method described in Section 5, using $prob\_link$ and $prob\_single$ thresholds of 99% and 1% respectively. This resulted in 26 clusters and one singleton for 79 verbs. Recall that without constraining the DPMM these thresholds produced 20 clusters and 5 singletons for 70 verbs. 49 verbs are shared in both outputs, while the average cluster size is similar.

The resulting clusters are highly accurate. As many as 16 (i.e. 62%) of them represent a single gold standard class, 7 of which contain (only) the pairs of *must-linked* verbs. Interestingly, only 11 out of 17 gold standard classes are exemplified among the 16 clusters, with 5 classes subdivided into finer-grained classes. Each of these sub-divisions seems semantically fully motivated (e.g. 30.3 PEER verbs were subdivided so that *peep* and *peek* were assigned to a different cluster than the semantically different *gaze*, *glance* and *stare*) and 4 of them can be directly attributed to the use of *must-links*.

From the 6 clusters that contained members from two different gold standard classes, the majority (5) make sense as well. 3 of these contain members of *must-link* pairs together with verbs from semantically related classes (e.g. 37.7 SAY and 40.2 NONVERBAL EXPRESSION classes). 3 of the clusters that contain members of several gold standard classes include *must-link* pairs as well. In two cases *must-links* have helped to bring together verbs which belong to the same class (e.g. the members of the *must-link* pair *broaden-freeze* which represent 45.4 CHANGE OF STATE class appear now in the same cluster with other class members *dampen*, *soften* and *sharpen*). Thus, DPMMs prove useful in learning novel information taking into account pairwise constraints. Only 4

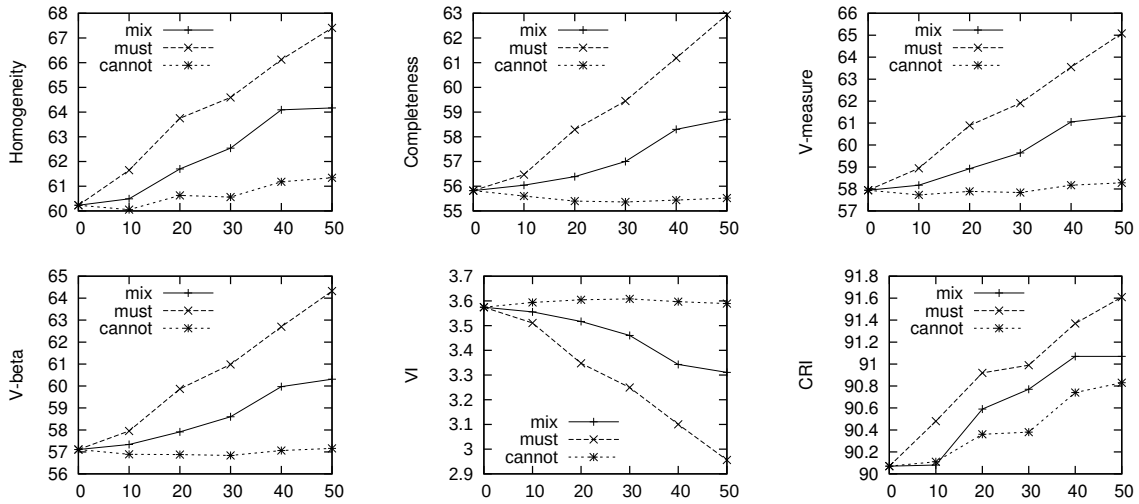


Table 3: Performance of constrained DPMMs incorporating pairwise links.

(i.e. 15%) of the clusters in the output examined are not meaningful (mostly due to the mismatch between the syntax and semantics of verbs).

## 8 Related work

Previous work on unsupervised verb clustering used algorithms that require the number of clusters as input e.g. PC, Information Bottleneck (Korhonen et al., 2006b) and spectral clustering (Brew and Schulte im Walde, 2002). In terms of applying non-parametric Bayesian approaches to NLP, Haghighi and Klein (2007) evaluated the clustering properties of DPMMs by performing anaphora resolution with good results.

There is a large body of work on semi-supervised learning (SSL), but relatively little work has been done on incorporating some form of supervision in clustering. It is important to note that the pairwise links used in this work constitute a weak form of supervision since they cannot be used to infer class labels which are required for SSL. However, the opposite can be done. Wagstaff & Cardie (2000) employed *must-links* and *cannot-links* to constrain the COBWEB algorithm, while Klein et al. (2002) applied them to complete-link hierarchical agglomerative clustering. The latter also studied how the added links affect instances not directly involved in them.

It can be argued that one could use clustering algorithms that require the number of clusters to be known in advance to discover interesting subclasses such as those discovered by the DPMMs. However, this would normally require multiple runs and manual inspection of the results, while DPMMs discover them automatically. Apart from the fact that fixing the number of clusters in advance restricts the discovery of novel information in the data, such algorithms cannot take full advantage of the pairwise constraints, since the latter are likely to change the number of clusters.

## 9 Conclusions - Future Work

In this work, following Vlachos et al. (2008) we explored the application of DPMMs to the task of verb clustering. We modified V-measure (Rosenberg and Hirschberg, 2007) to deal more appropriately with the varying number of clusters discovered by DPMMs and presented a method of agregating the generated samples which allows for qualitative evaluation. The quantitative and qualitative evaluation demonstrated that they achieve performance comparable with that of previous work and in addition discover novel information in the data. Furthermore, we evaluated the incorporation of constraints to guide the DPMM obtaining promising results and we discussed their application in a real-world setup.

The results obtained encourage the application of DPMMs and non-parametric Bayesian methods to other NLP tasks. We plan to extend our experiments to larger datasets and further domains. While the improvements achieved using randomly selected pairwise constraints were promising, an active constraint selection scheme as in Klein et al. (2002) could increase their impact. Finally, an extrinsic evaluation of the clustering provided by DPMMs in the context of an NLP application would be informative on their practical potential.

## Acknowledgments

We are grateful to Diarmuid Ó Séaghdha and Jurgen Van Gael for helpful discussions.

## References

Chris Brew and Sabine Schulte im Walde. 2002. Spectral Clustering for German Verbs. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

Hoa Trang Dang. 2004. *Investigations into the role of lexical semantics in word sense disambiguation*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Sharon J. Goldwater. 2007. *Nonparametric bayesian models of lexical acquisition*. Ph.D. thesis, Brown University, Providence, RI, USA.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Dan Klein, Sepandar Kamvar, and Chris Manning. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006a. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2006b. Automatic classification of verbs in biomedical texts. In *Proceedings of the COLING-ACL*, pages 345–352.

Daniel D. Lee and Sebastian H. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October.

Beth Levin. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago.

Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Daniel J. Navarro, Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. 2006. Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122, April.

Radford M. Neal. 2000. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, June.

Jan Puzicha, Thomas Hofmann, and Joachim Buhmann. 2000. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CoNLL*, pages 410–420, Prague, Czech Republic.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 95–102.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of COLING-ACL*, pages 985–992, Sydney, Australia.

Andreas Vlachos, Zoubin Ghahramani, and Anna Korhonen. 2008. Dirichlet process mixture models for verb clustering. In *Proceedings of the ICML workshop on Prior Knowledge for Text and Language*.

Kiri Wagstaff and Claire Cardie. 2000. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA. ACM Press.