# Unsupervised Clustering of Utterances using Non-parametric Bayesian Methods

*Ryuichiro Higashinaka*[1], *Noriaki Kawamae*[2], *Kugatsu Sadamitsu*[1], *Yasuhiro Minami*[3], *Toyomi Meguro*[3], *Kohji Dohsaka*[3], *and Hirohito Inagaki*[1]

[1]NTT Cyber Space Laboratories, NTT Corporation
[2]NTT Comware Corporation
[3]NTT Communication Science Laboratories, NTT Corporation

## Abstract

Unsupervised clustering of utterances can be useful for the modeling of dialogue acts for dialogue applications. Previously, the Chinese restaurant process (CRP), a non-parametric Bayesian method, has been introduced and has shown promising results for the clustering of utterances in dialogue. This paper newly introduces the infinite HMM, which is also a non-parametric Bayesian method, and verifies its effectiveness. Experimental results in two dialogue domains show that the infinite HMM, which takes into account the sequence of utterances in its clustering process, significantly outperforms the CRP. Although the infinite HMM outperformed other methods, we also found that clustering complex dialogue data, such as human-human conversations, is still hard when compared to human-machine dialogues.

**Index Terms**: Unsupervised clustering, Nonparametric Bayesian methods, Chinese restaurant process, Infinite HMM

## 1. Introduction

Dialogue acts, which are meaning representations for utterances, play key roles in dialogue systems research [1]. Since they serve as fundamental units for understanding user intentions, dialogue management, and language generation, they need to be carefully designed before any system deployment. However, the appropriate modeling of dialogue acts is often difficult because it requires an extensive effort by experts to achieve good coverage of user utterances.

One of the most difficult steps in dialogue act modeling is to determine the number of dialogue acts. This paper proposes applying unsupervised clustering methods to dialogue data in order to find appropriate clusters of utterances and estimate the number of dialogue acts. There have been several attempts to cluster utterances using such methods as K-means [2] and Kohonen self-organizing maps (SOMs) [3]; however, they require that the number of clusters (or the size of lattices for SOMs) be known in advance. Recently, a non-parametric Bayesian method called the Chinese restaurant process (CRP) has been utilized to infer the number of clusters in dialogue data and has shown some promising signs [4]. This paper builds on their findings and aims to find solutions to some of the remaining problems.

First, although the CRP has been shown to create better clustering results as iterations increase [4], it has not been compared with other clustering methods, making it difficult to judge the effectiveness of the CRP. Second, the CRP does not take into account the sequence of utterances in the clustering process, which may not be appropriate when considering that the function of an utterance is influenced by the context of a dialogue. Finally, the CRP has been applied to the Dihana corpus [5], which is a collection of dialogues between a system and human users. Since utterances of automated systems can be rigid because of the use of templates and rules, clustering can be easily performed for such data. We need to examine whether the CRP can also be used for more complex data, such as human-human conversations, so as to clarify its applicability.

To address the above three problems, we introduce another non-parametric Bayesian method called the **infinite HMM**, which can estimate the number of clusters while taking into account the sequence of utterances in its clustering process. We compare the CRP with the infinite HMM to investigate whether taking into account the sequence of utterances is effective. We also compare the CRP and the infinite HMM with K-means provided with the correct number of clusters to clarify the effectiveness of non-parametric Bayesian methods. In addition, we use human-human dialogue data we collected to investigate the applicability of the CRP and the infinite HMM.

## 2. Unsupervised Clustering Methods

We first describe K-means, which we regard as our baseline, and then describe the two non-parametric Bayesian methods; namely, the CRP and the infinite HMM. Both methods are related to the Dirichlet process [6], which is a non-parametric Bayesian model [7]. Dirichlet process mixture (DPM) models [8] and hierarchical Dirichlet processes (HDPs) [9] are two realizations of the Dirichlet process for handling mixture models, and the CRP and the infinite HMM are the implementations of the DPM and the HDP, respectively. They both assume infinite states in modeling data and have been applied to estimating the number of components in mixture models [10], the number of states in HMMs [9], and the number of dialogue acts in dialogue [4].

### 2.1. K-means

K-means is a standard method for clustering. Given a set of data and the number of clusters (K), it clusters the data by iteratively updating the centroids. Initially, K data are randomly chosen as centroids, and each datum is clustered to its nearest centroid using a distance function, such as the Euclidean distance. Then, centroids are updated to the mean of its cluster members. This process is repeated until convergence. Although K-means is widely used today for simplicity, its drawback is that K has to be determined in advance.



28 – 31 August 2011, Florence, Italy

### 2.2. Chinese Restaurant Process

In the CRP, a datum is called a customer and a cluster a table. The first customer becomes seated at the first table. Then, each following customer ($c_i$), sits at one of the currently occupied tables ($t_j$) or creates a new table ($t_{new}$; $new$ is the index given to a new table) with the probability

$$\mathrm{P}(t_j|c_i) \propto \begin{cases} \dfrac{\mathrm{n}(t_j)}{N+\alpha} \cdot \mathrm{P}(c_i|t_j) & (\text{if } j \neq new) \\ \dfrac{\alpha}{N+\alpha} \cdot \mathrm{P}(c_i|t_j) & (\text{if } j = new), \end{cases}$$

where '$\mathrm{n}(t_j)$' is a function that returns the number of customers for $t_j$, N the total number of customers already seated, $\alpha$ the hyper-parameter that determines how likely a new table is created, and $\mathrm{P}(c_i|t_j)$ the probability that $c_i$ is generated from $t_j$; that is,

$$\mathrm{P}(c_i|t_j) = \prod_{w \in W} \mathrm{P}(w|t_j)^{\mathrm{count}(c_i,w)}$$

$$\mathrm{P}(w|t_j) = \frac{\mathrm{count}(t_j, w) + \beta}{\sum_{w \in W} \mathrm{count}(t_j, w) + |W| \cdot \beta}.$$

Here, $W$ is a set of features, $\mathrm{count}(*, w)$ a function that returns the number of occurrences of a feature $w$ for a customer or a table, and $\beta$ the hyper-parameter. We use a uniform distribution for $\mathrm{P}(c_i|t_{new})$.

After all customers have been seated, Gibbs sampling is performed; that is, we repeat picking up one of the customers from its table and relocating the customer as if he/she were the last customer to be seated. After performing a sufficient number of samplings, we obtain the optimal number of tables together with their customers, which become the clustering results.

### 2.3. Infinite HMM

In the infinite HMM, a customer $c_i$ becomes seated at an already occupied table $t_j$ or creates a new table ($t_{j=new}$) with the probability

$$\mathrm{P}(t_j|c_i) \propto \mathrm{P}(t_{c_{i-1}}, t_j) \cdot \mathrm{P}(t_j, t_{c_{i+1}}) \cdot \mathrm{P}(c_i|t_j),$$

where $t_c$ means the table of a customer $c$. Here, we assume that the customers are given a sequential order, and $c_{i-1}$ and $c_{i+1}$ denote the previous and next data of $c_i$. $\mathrm{P}(t_j, t_k)$ is a simple transition probability:

$$\mathrm{P}(t_j, t_k) = \frac{\mathrm{transitions}(t_j, t_k) + \gamma}{\sum_{l=1}^{K} \mathrm{transitions}(t_j, t_l) + K \cdot \gamma + \alpha},$$

where $\alpha$ is the hyper-parameter that determines how likely a new table is created, K is the number of occupied tables, and $\mathrm{transitions}(t_j, t_k)$ returns the number of transitions from $t_j$ to $t_k$. $\gamma$ is a flooring value to avoid zero probability. The probability for creating a new table is

$$\mathrm{P}(t_{c_{i-1}}, t_{new}) \cdot \mathrm{P}(t_{new}, t_{c_{i+1}}) \cdot \mathrm{P}(c_i|t_{new}),$$

$$\mathrm{P}(t_{c_{i-1}}, t_{new}) = \frac{\alpha}{\sum_{l=1}^{K} \mathrm{transitions}(t_{c_{i-1}}, t_l) + \alpha},$$

$$\mathrm{P}(t_{new}, t_{c_{i+1}}) = \frac{\alpha}{0 + \alpha} = 1.$$

Here, we use a uniform distribution for $\mathrm{P}(c_i|t_{new})$.

Similarly to the CRP, Gibbs sampling is performed to obtain the optimal number of tables and the locations for the customers.

# 3. Experiment

We performed an experiment to verify the effectiveness of the non-parametric Bayesian methods. We first prepared dialogue data, and converted the utterances into feature vectors. Then, we applied the three clustering methods to the vectors. For the evaluation, we compare the clustering results against the human-annotated dialogue acts.

### 3.1. Dialogue Data

We used dialogues in an animal discussion (**AD**) domain and attentive listening (**AL**) domain. All dialogues are in Japanese and are text dialogues although we also plan to work on spoken dialogue data.

#### *3.1.1. Animal Discussion Domain*

We collected chat-like dialogues using an automated dialogue system [11]. In this domain, the conversational participants (the system and a human user) talked about likes and dislikes about animals via a text chat interface. The data consist of 1000 dialogues and all user/system utterances were annotated with 29 dialogue acts, which include those related to self-disclosure, question, response, and greetings. See [11] for the description of the system and the definitions of the dialogue acts.

#### *3.1.2. Attentive Listening*

We collected human-human listening-oriented dialogues [12]. In this AL domain, a listener attentively listens to the other in order to satisfy the speaker's desire to speak and to make himself/herself heard.

We collected such listening-oriented dialogues using a website where users taking the roles of listeners and speakers were matched up to have conversations. The conversations were done through a text-chat interface. The use of facial expressions was not allowed. The participants ended each conversation after approximately ten minutes. We collected 1260 such listening-oriented dialogues and annotated them with 38 dialogue acts. The dialogue acts include those for greetings, giving information, sympathy, self-disclosure, and so forth. See [12] for the complete list of dialogue acts.

### 3.2. Creating Feature Vectors

Since the computational cost of Gibbs sampling is rather high, we randomly sampled 50 dialogues from each domain for use in this experiment. There are 2894 and 2470 utterances in the AD and AL domains, respectively. Each utterance was converted into a feature vector using the bag-of-words (**BOW**) representation. As a morphological analyzer, we used ChaSen [1] and used word base forms as features. To avoid the influence of uncommon words, we did not use words that occurred less than ten times in the entire data set in each domain.

We conceived two other feature sets: (a) **CW-POS**, in which content words (nouns, verbs, adjectives, and unknown words) are abstracted to their part-of-speech (POS) tags before making a BOW representation, and (b) **POS**, in which all words are abstracted to their POS tags. The motivation behind (a) is to abstract proper nouns and numerical expressions and to put an emphasis on cue phrases, such as functional words (non-content words), which are known to be strong classification indicators of dialogue acts [13]. We have (b) as an extreme case of abstrac-

[1] http://chasen-legacy.sourceforge.jp/

Table 1: Evaluation results averaged over all 100 trials for the animal discussion domain. The asterisks, +, and † indicate statistical significance ($p<0.01$) by a non-paired t-test over K-means, the CRP, and the infinite HMM, respectively.

| | K-means | | | CRP | | | Infinite HMM | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature set | POS | CW-POS | BOW | POS | CW-POS | BOW | POS | CW-POS | BOW |
| Purity | 0.464 | 0.471 | 0.438 | 0.460 | $0.537^{*}$ | $0.551^{*}$ | $0.655^{*+}$ | $0.674^{*+}$ | $0.645^{*+}$ |
| F | $0.346^{+\dagger}$ | 0.281 | 0.231 | $0.326^{\dagger}$ | $0.343^{*}$ | $0.421^{*}$ | 0.215 | $0.384^{*+}$ | $0.476^{*+}$ |
| # Clusters | 27 | 27 | 27 | 27.96 | 32.01 | 35.03 | 453.70 | 180.40 | 143.62 |

Table 2: Evaluation results averaged over all 100 trials for the attentive listening domain. See Table 1 for the notations in the table.

| | K-means | | | CRP | | | Infinite HMM | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature set | POS | CW-POS | BOW | POS | CW-POS | BOW | POS | CW-POS | BOW |
| Purity | $0.270^{+}$ | 0.278 | 0.249 | 0.253 | $0.315^{*}$ | $0.326^{*}$ | $0.315^{*+}$ | $0.348^{*+}$ | $0.344^{*+}$ |
| F | $0.134^{\dagger}$ | 0.140 | 0.138 | $0.136^{\dagger}$ | $0.156^{*}$ | $0.159^{*}$ | 0.119 | $0.164^{*+}$ | $0.171^{*+}$ |
| # Clusters | 33 | 33 | 33 | 21.11 | 25.70 | 29.28 | 112.70 | 35.40 | 38.00 |

tion. The dimensions of the feature vectors for BOW, CW-POS, and POS in the AD domain are 564, 193, and 57, respectively, and those in the AL domain are 1673, 404, and 62, respectively.

### 3.3. Evaluation Procedure

We applied K-means, the CRP, and the infinite HMM to our feature vectors using the three sets of features. We used K-means as our baseline. Here, we provided K-means with the correct number of dialogue acts so that it could be a competitive baseline. The correct number of dialogue acts is the number of dialogue acts that are present in the sampled data. We had 27 and 33 dialogue acts in the AD and AL domains, respectively. We used the Euclidean distance.

Since the performance of K-means depends on the initial randomly created clusters and the CRP and the infinite HMM work probabilistically, we performed 100 trials for each clustering method and compared their averaged performance. For the CRP and the infinite HMM, we used tentative values of 0.1 for $\alpha$ and 0.01 for $\beta$ and $\gamma$. The number of iterations (sweeps) was set to 100, which means that each customer is considered 100 times to be relocated.

### 3.4. Evaluation Metrics

We used the purity and the F measure (harmonic mean of precision and recall) as our evaluation metrics [14]. The purity indicates how occupied each cluster is with the same dialogue act, and the F measure indicates how accurately each pair of data are clustered:

**Purity:**

$$\mathrm{purity}(\mathcal{C}, \mathcal{D}) = \frac{1}{N}\sum_{k=1}^{K}\max_{j}|c_k \cap d_j|,$$

where $\mathcal{C} = \{c_1 \dots c_K\}$ is the set of clusters, $\mathcal{D} = \{d_1 \dots d_N\}$ the set of dialogue act tags, and $N$ the number of data.

**F measure:**

$$\mathrm{precision} = \frac{TP}{TP+FP}, \quad \mathrm{recall} = \frac{TP}{TP+FN},$$

$$\mathrm{F} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}},$$

where TP, FP, and FN indicate true positive, false positive, and false negative. Here, true positive is the number of times a pair of data with the same dialogue act tag are clustered in the same cluster, false positive is the number of times a pair with a different dialogue act tag are clustered in the same cluster, and false negative is the number of times a pair with the same dialogue act tag are clustered in different clusters.

### 3.5. Results

Tables 1 and 2 show the clustering performance averaged over 100 trials. When we focus on the BOW feature set, we can confirm that the CRP performs better than K-means, but we also find that the infinite HMM performs significantly better than the CRP. When we look at the number of clusters, we see a similar number of clusters for K-means and the CRP, but a larger number of clusters for the infinite HMM, which is not surprising considering that the context is taken into account. The high performance of the infinite HMM suggests that similar utterances on the surface level have been successfully clustered into different clusters depending on the context. The number of clusters by the CRP is generally closer to the correct number, suggesting that humans design dialogue acts mostly irrespective of utterance sequences and that the number of dialogue acts can be redesigned to be more fine-grained by taking the context into account. Note that we have a very large number of clusters for the infinite HMM with the POS feature set, probably because the lack of word information is complemented by the transitions.

When we focus on the other feature sets, we see that they perform poorly compared to the BOW, suggesting that both content words and functional words are important in distinguishing dialogue acts. When we look at the differences in the domains, we see that it is much harder to cluster human-human utterances (i.e., the AL domain) as we expected, although the infinite HMM still performs reasonably better than other methods.

Table 3 shows the clustering results in the AL domain for the infinite HMM using the BOW feature set. As the table shows, each dialogue act is mostly clustered into single dominating clusters. This indicates that, by looking at these dominating clusters, we could obtain typical utterances of the same dialogue act, leading to a rapid modeling of dialogue acts from dialogue data.

## 4. Summary and Future Work

This paper introduced the infinite HMM, a non-parametric Bayesian method, for the clustering of utterances in dialogue. Experimental results showed that the infinite HMM significantly outperforms the previously proposed CRP in two dia-

Table 3: Clustering results in the AL domain by the infinite HMM using the BOW feature set. Shown here are the results for one of the 100 trials. There were 37 clusters made in this trial. This table shows to which clusters the reference dialogue acts (DAs) were clustered. Cluster IDs and their counts are shown as <cluster ID>:<count>. Reference DAs that occurred less than 20 times and clusters IDs with less than two counts are omitted for brevity. The largest counts in each row are shown in **bold**.

| Reference DA | Clusters |
|---|---|
| GREETING | 1:8, 3:38, 7:**147**, 9:5, 10:4, 13:60, 15:4, 19:15, 21:8, 26:7, 27:6 |
| SELF-DISC-FACT | 1:7, 2:3, 3:3, 4:44, 5:3, 6:14, 9:50, 10:**88**, 11:5, 14:3, 15:4, 17:12, 18:6, 19:20, 20:7, 22:11, 23:5, 24:4, 26:3 |
| SELF-DISC-PREF (positive) | 2:14, 3:3, 4:7, 6:3, 9:18, 10:**112**, 11:12, 16:3, 17:20, 18:11, 19:5, 24:3, 26:23, 28:15 |
| SYMPATHY | 1:3, 2:23, 3:6, 4:10, 6:3, 9:11, 10:28, 12:3, 13:5, 16:6, 17:23, 18:5, 19:6, 25:3, 26:**117**, 28:3 |
| INFORMATION | 1:4, 2:3, 3:4, 4:19, 6:15, 8:3, 9:21, 10:**67**, 11:3, 14:5, 15:5, 17:9, 18:11, 20:4, 24:3, 26:5, 28:4 |
| SELF-DISC-PREF (negative) | 3:3, 4:8, 6:4, 9:8, 10:**68**, 11:6, 16:4, 18:3, 19:4, 24:5, 26:4 |
| QUESTION-FACT | 6:24, 8:10, 9:12, 10:13, 11:3, 15:**34**, 18:5, 22:8, 24:7 |
| CONFIRMATION | 4:3, 6:6, 8:7, 9:9, 10:**20**, 11:5, 15:5, 17:3, 18:11, 20:3, 24:9, 26:16 |
| SELF-DISC-EXP | 1:4, 4:**25**, 5:3, 6:7, 9:6, 10:16, 19:7, 24:3, 30:5 |
| THANKS | 21:**77** |
| SELF-DISC-PREF (neutral) | 4:5, 6:5, 9:10, 10:**30** |
| ACKNOWLEDGMENT | 3:3, 13:6, 17:4, 26:**38** |
| QUESTION-PREF | 6:10, 8:7, 9:3, 10:**11**, 16:3, 18:8, 19:4, 24:4 |
| QUESTION-INFORMATION | 6:6, 8:4, 9:3, 10:**12**, 16:4, 18:8, 24:6 |
| SELF-DISC-DESIRE | 3:3, 5:3, 6:3, 9:6, 10:**10**, 19:9 |
| SELF-DISC-HABIT | 4:8, 6:7, 9:5, 10:**11**, 20:3 |
| REPEAT | 3:4, 6:3, 9:5, 10:6, 26:**7** |
| ADMIRATION | 3:5, 9:4, 10:4, 17:6, 26:**7**, 28:3 |
| NON-SYMPATHY | 1:3, 4:5, 9:4, 10:**7**, 19:5 |
| APPROVAL | 6:4, 9:4, 10:5, 17:**6**, 26:5 |
| SELF-DISC-PLAN | 9:4, 10:3, 19:**10** |
| QUESTION-HABIT | 8:3, 18:**4**, 22:**4**, 24:3 |
| PROPOSAL | 10:4, 19:**12** |
| PARAPHRASE | 9:**4**, 10:**4** |

logue domains. Regarding the problems mentioned in the introduction, we found that (1) the CRP works better than K-means but the infinite HMM performs significantly better, (2) the sequence of utterances should be incorporated to achieve better clustering performance, and (3) human-human dialogues are much harder to cluster compared to human-machine dialogues.

As future work, we need to investigate other features for better clustering performance, especially for human-human dialogues. In addition, we would like to adopt our findings in our modeling of dialogue acts for new domains. As a trial, we applied the CRP and the infinite HMM to estimate the number of dialogue acts for contact center simulation dialogues where an operator and a user perform transactions [15]. Table 4 shows the estimated number of clusters when we used 50 sampled dialogues (6550 utterances). The experimental condition was the same as we described in Section 3. It can be seen that the appropriate number of dialogue acts for this domain can be around 30 and can reach up to 80. We plan to use this insight in our modeling of dialogue acts in the development of our call summarization system [15]. In addition, we also need to optimize the hyper-parameters $\alpha$, $\beta$ and $\gamma$ for the CRP and the infinite HMM for better clustering accuracy.

Table 4: Averaged number of clusters over 100 trials for the contact center simulation dialogues.

| | CRP | | | Infinite HMM | | |
|---|---|---|---|---|---|---|
| Feature set | POS | CW-POS | BOW | POS | CW-POS | BOW |
| # Clusters | 27.46 | 30.00 | 31.18 | 911.03 | 61.96 | 78.70 |

## 5. References


[1] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. V. Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[2] K. Ohtake, "Unsupervised approach for dialogue act classification," in *Proc. PACLIC-22*, 2008, pp. 445–451.

[3] T. Andernach, M. Poel, and E. Salomons, "Finding classes of dialogue utterances with Kohonen networks," in *Proc. the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, 1997, pp. 85–94.

[4] N. Crook, R. Granell, and S. Pulman, "Unsupervised classification of dialogue acts using a Dirichlet process mixture model," in *Proc. SIGDIAL*, 2009, pp. 341–348.

[5] D. Griol, L. F. Hurtado, E. Sanchis, and E. Segarra, "Acquiring and evaluating a dialog corpus through a dialog simulation technique," in *Proc. SIGDIAL*, 2007, pp. 39–42.

[6] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[7] J. K. Ghosh, *Bayesian Nonparametrics*. Springer, 2003.

[8] D. Aldous, "Exchangeability and related topics," in *Lecture Notes in Math*, vol. 1117, 1985, pp. 1–198.

[9] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. NIPS*, 2004.

[10] C. E. Rasmussen and Z. Ghahramani, "Infinite mixtures of Gaussian process experts," in *Proc. NIPS*, 2001, pp. 881–888.

[11] R. Higashinaka, K. Dohsaka, and H. Isozaki, "Effects of self-disclosure and empathy in human-computer dialogue," in *Proc. SLT*, 2008, pp. 109–112.

[12] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable Markov decision processes," in *Proc. COLING*, 2010, pp. 761–769.

[13] N. Webb and M. Ferguson, "Automatic extraction of cue phrases for cross-corpus dialogue act classification," in *Proc. COLING*, vol. Poster, 2010, pp. 1310–1317.

[14] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, (Chapter16: Flat clustering).

[15] R. Higashinaka, Y. Minami, H. Nishikawa, K. Dohsaka, T. Meguro, S. Takahashi, and G. Kikui, "Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues," in *Proc. COLING*, vol. Poster, 2010, pp. 400–408.