Information and Media Technologies 2(1): 317–325 (2007)
reprinted from: Transactions of the Japanese Society for Artificial Intelligence 22 (2007) (2): 218 –226 (2007)

特集論文

Discovering Concepts from Word Co-occurrences with a Relational Model

Kenichi Kurihara — Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
kurihara@mi.cs.titech.ac.jp, http://mi.cs.titech.ac.jp/kurihara

Yoshitaka Kameya — (affiliation as previous author)
kameya@mi.cs.titech.ac.jp, http://mi.cs.titech.ac.jp/kameya

Taisuke Sato — (affiliation as previous author)
sato@mi.cs.titech.ac.jp, http://mi.cs.titech.ac.jp/sato

keywords: clustering, Dirichlet process, variational inference, relational learning

Summary

Clustering word co-occurrences has been studied to discover clusters as latent concepts. Previous work has applied the semantic aggregate model (SAM), and reports that discovered clusters seem semantically significant. The SAM assumes a co-occurrence arises from one latent concept. This assumption seems moderately natural. However, to analyze latent concepts more deeply, the assumption may be too restrictive. We propose to make clusters for each part of speech from co-occurrence data. For example, we make adjective clusters and noun clusters from adjective–noun co-occurrences while the SAM builds clusters of “co-occurrences.” The proposed approach allows us to analyze adjectives and nouns independently.

To take this approach, we propose a frequency-based infinite relational model (FIRM) for word co-occurrences. The FIRM is a stochastic block model that takes into account the frequency of observations although traditional stochastic blockmodels ignore it. The FIRM also utilizes the Dirichlet process so that the number of clusters is inferred. We derive a variational inference algorithm for the model to apply to a large dataset. Experimental results show that the FIRM is more helpful to analyze adjectives and nouns independently, and the FIRM clusters capture the SAM clusters better than a stochastic blockmodel.

1. Introduction

Clustering word co-occurrences has been studied to discover clusters as latent concepts. Pereira et al. reported discovered such clusters seem semantically significant [Pereira 93]. Mochihashi and Matsumoto experimentally discovered word meaning by clusters [Mochihashi 02]. Nakagawa et al. have proposed metaphor understanding based on word co-occurrence clustering [Nakagawa 06]. In these studies, the semantic aggregate model (SAM) has been applied, which assumes a co-occurrence comes from a single latent concept. This assumption seems moderately natural. However, to analyze latent concepts more deeply, the assumption may be too restrictive.

The goal of this study is to give deeper analysis of clusters as concepts than the semantic aggregate model (SAM). We propose to make clusters for each part of speech from co-occurrence data. For example, we make adjective clusters and noun clusters from adjective–noun co-occurrences unlike the SAM which builds clusters of “co-occurrences.” In other words, the adjective clusters and the noun clusters discovered by the proposed approach give interpretations of adjectives and nouns independently, but the co-occurrence clusters discovered by the SAM only explains “co-occurrences.” The proposed approach is thus more useful to analyze each of the adjectives and the nouns.

Recently, relational learning has received a great deal of attention[*1], for example, to find social roles in social network data. The stochastic blockmodel is a well-known model for relational learning in sociology. Kemp et al. have proposed an infinite relational

[*1] Recent workshops on statistical relational learning,
- http://kdl.cs.umass.edu/srl2003/
- http://www.cs.umd.edu/projects/srl2004/
- http://www.cs.umd.edu/projects/srl2006/

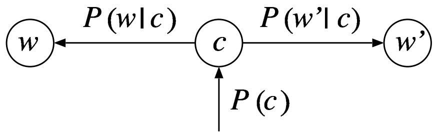


Fig. 1 Graphical representation of Semantic Aggregate Model.

model (IRM) [Kemp 06], which is a stochastic blockmodel exploiting the Dirichlet process (DP) [Ferguson 73, Antoniak 74]. The IRM partitions each type into clusters, and the number of clusters is estimated using the Dirichlet process.

A word co-occurrence can also be regarded as a relation. If adjective a and noun n have a co-occurrence, put $R(a,n)=1$, otherwise $R(a,n)=0$. Therefore, it is straightforward to apply relational models to word co-occurrences.

To take the proposed approach, we propose a model called frequency-based infinite relational model (FIRM) for word co-occurrences. The FIRM is a stochastic block model. However, it takes into account the frequency of observations which is statistically informative while traditional stochastic blockmodels ignore frequency. The FIRM also utilizes the Dirichlet process. Since the inference using the Dirichlet process is computationally expensive, we derive a variational inference algorithm for the FIRM to apply to one million datacases. In experiments, we evaluate the FIRM and the IRM using the SAM as the gold standard because it has experimentally been shown that the SAM makes clusters that are consistent with psychological experiments [Nakagawa 06]. Experimental results show that the FIRM discovers adjective clusters and noun clusters that capture those discovered by the SAM better than the IRM.

2. Semantic Aggregate Model

We briefly review the semantic aggregate model (SAM), which has been applied to clustering of word co–occurrences [Pereira 93, Mochihashi 02, Nakagawa 06][*2]. The SAM is a generative probability model for word co-occurrences, in which a co-occurrence of two words comes from a concept we implicitly have. Let c be a concept, w and w' be words. The SAM assumes the following factorization of $p(w,w',c)$,

$$p(w,w',c)=p(w|c)p(w'|c)p(c). \tag{1}$$

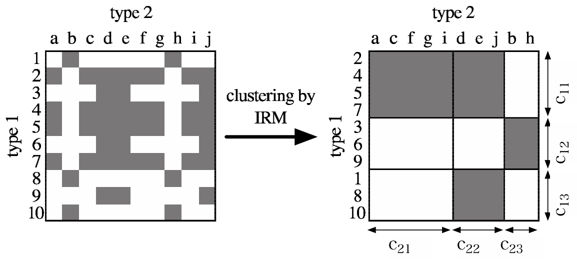


Fig. 2 A toy example of the infinite relational model (IRM). The IRM partitions type 1 (1–10) into clusters $c_{11}-c_{13}$ and type 2 (a–j) into $c_{21}-c_{23}$.

Figure 1 graphically depicts the SAM. $p(w|c)$ and $p(c)$ can be seen as parameters of the SAM. Given these parameters, we can compute the membership distribution $p(c|w)$,

$$p(c|w)=\frac{p(c)p(w|c)}{p(w)}=\frac{p(c)p(w|c)}{\sum_c p(c)p(w|c)}. \tag{2}$$

This membership distribution indicates how often concept c occurs when word w occurs. Therefore, the membership distribution may allow us to capture conceptual characteristics of word w. For example, we can use a similarity measure δ between word w and word w': $\delta(w,w')=\exp(-\mathrm{KL}(p(c|w)||p(c|w')))$ where KL is the Kullback–Liebler divergence.

Nakagawa et al. conducted psychological experiments on the clustering results by the SAM, and showed that the clusters are consistent with the results of psychological experiments [Nakagawa 06]. So, in this paper, we evaluate co-occurrence models by comparing them with the SAM as the gold standard.

3. Infinite Relational Model

Kemp et al. proposed the infinite relational model (IRM) in the context of statistical relational learning [Kemp 06]. The IRM is a general model to partition each *type* into clusters. Figure 2 is a toy example. The left hand side matrix shows relation R : $(type\ 1)\times(type\ 2)\rightarrow\{0,1\}$ (each black and white dot shows $R(\cdot,\cdot)=1$ and $R(\cdot,\cdot)=0$, respectively). For example, *type* 1, *type* 2 and R can be a set of people, a set of animals and predicate "like", i.e. $R(i,j)=1$ means person i likes animal j. Given relation R as an input, the IRM makes clusters like the right hand side matrix in Figure 2. In this example, *type* 1 is partitioned into clusters c_{11}–c_{13}, and *type* 2 is also partitioned into clusters c_{21}–c_{23}.

We apply the IRM to word co-occurrence clustering. The dataset we use consists of adjective–noun

[*2] Pereira et al. and Nakagawa et al. put different parameters on w and w' in Figure 1, but Mochihashi and Matsumoto used the same parameter set.

Table 1 Notation

Distributions	
DP	the Dirichlet process
$\mathrm{Beta}(x;a,b)$	the beta distribution; $\propto x^{a-1}(1-x)^{b-1}$
$\mathrm{Bern}(x;a)$	the Bernoulli distribution; $= a^x(1-a)^{x-1}$
$\mathrm{Mult}(\boldsymbol{x};\boldsymbol{a})$	the multinomial distribution; $= \prod_{i=1}^{n} \boldsymbol{a}_{\boldsymbol{x}_i}$ where $\boldsymbol{x} = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)$ and $\sum_c \boldsymbol{a}_c = 1$
$\mathrm{Dir}(\boldsymbol{x};\boldsymbol{a})$	the Dirichlet distribution; $\propto \prod_i \boldsymbol{x}_i^{\boldsymbol{a}_i - 1}$
Observation	
A	the set of adjectives
N	the set of nouns
(a,n)	a co-occurrence consisting of adjective a and noun n
D	training data; $D = \{(a_i,n_i) \mid i = 1...m, a_i \in A, n_i \in N\}$
$f(a,n)$	the number of observations of (a,n) in D
Hidden Variables	
η	the parameter of the Bernoulli for co-occurrences
Z^A	the assignments of adjectives; $Z^A = \{z_a^A \mid a \in A\}$ where z_a^A indicates a cluster.
Z^N	the assignments of nouns; $Z^N = \{z_n^N \mid n \in N\}$ where z_n^N indicates a cluster.
$\boldsymbol{u}^A$, $\boldsymbol{u}^N$	the parameter of a multinomial for adjectives and nouns; $\sum_{a\in A} \boldsymbol{u}_a^A = 1$, $\sum_{n \in N} \boldsymbol{u}_n^N = 1$
$\boldsymbol{v}^A$, $\boldsymbol{v}^N$	the stick lengths for the Dirichlet process
Hyperparameters	
T^A, T^N	truncation levels for variational inference
α	the parameter of the Dirichlet dist. for $\boldsymbol{u}^A$ and $\boldsymbol{u}^N$
β	the parameter of the beta distribution for η
γ	the parameter of the Dirichlet process for Z^A and Z^N

co-occurrences. To apply the IRM to adjective–noun co-occurrences, we define the following relation R,

$$R : \text{adjectives} \times \text{nouns} \rightarrow \{0,1\}.$$

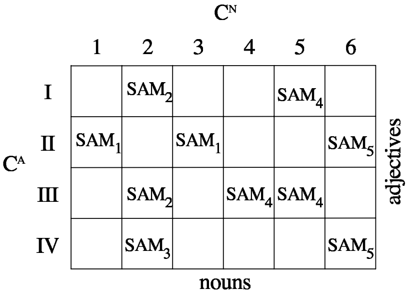
If co-occurrence (a,n) exists in a dataset, put $R(a,n) = 1$, otherwise put $R(a,n) = 0$. Kemp et al. stated that they ignore missing relations whereas we regard them as negative relations, i.e. $R(a,n) = 0$. This is because the number of observed adjective–noun co-occurrences is quite smaller than that of possible co-occurrences[*3].

Using the notation in Table 1, the IRM for adjective–noun co-occurrence is modeled as,

$$
\begin{aligned}
Z^A|\gamma &\sim \mathrm{DP}(\gamma) && (3)\\
Z^N|\gamma &\sim \mathrm{DP}(\gamma) && (4)\\
\eta(t^A,t^N)|\beta &\sim \mathrm{Beta}(\eta(t^A,t^N);\beta,\beta) && (5)\\
R(a,n)|z_a^A,z_n^N,\eta &\sim \mathrm{Bern}(R(a,n);\eta(z_a^A,z_n^N)), && (6)
\end{aligned}
$$

The IRM does not need to specify the number of clusters with the Dirichlet process (DP) whereas the traditional stochastic blockmodel requires the number of clusters. From (6), the joint probability of R is,

$$p(R|Z^A,Z^N,\eta) = \prod_{a\in A}\prod_{n \in B} \eta(z_a^A,z_n^N)^{R(a,n)} (1-\eta(z_a^A,z_n^N))^{1-R(a,n)}. \quad (7)$$



	1	2	3	4	5	6
I		SAM$_2$			SAM$_4$	
II	SAM$_1$		SAM$_1$			SAM$_5$
III		SAM$_2$		SAM$_4$	SAM$_4$	
IV		SAM$_3$				SAM$_5$

Fig. 3 A conceptual image of the relational clustering and clustering by the SAM. $C^{\mathrm{SAM}} = \{\mathrm{SAM}_1,...,\mathrm{SAM}_5\}$ is the clusters by the SAM. $C^A = \{\mathrm{I},...,\mathrm{IV}\}$ and $C^N = \{1,...,6\}$ are clusters of adjective and nouns by the a relational model. In this example, SAM_1 is explained by adjective cluster II and noun clusters 1 and 3.

We are interested in inferring Z^A and Z^N. Although it can be carried out using a Markov chain Monte Carlo (MCMC) method, Kemp et al. simply inferred the best partition Z^A and Z^N of the IRM by hill climbing [Kemp 06].

4. Semantic Aggregate Model and Relational Models

The SAM assumes that words in one co-occurrence comes out of the same latent concept (see Figure 1) though adjectives and nouns could have different sets of concepts, intuitively. On the other hand, relational models, e.g. the IRM, discover clusters for each *type*. In the case of adjective-noun co-occurrences, they discover adjective clusters and noun clusters.

[*3] Our dataset has 210,605 distinct co-occurrences consisting of 1,291 adjectives and 3,705 nouns. The number of possible co-occurrences is 4,783,155. Therefore, 210,605 distinct co-occurrences are just 4.4% of possible co-occurrences.

We expect that clusters discovered by a relational model lead to a deeper analysis of adjectives and nouns than the SAM as shown by an example in Figure 3. Let C^{SAM} be the set of clusters discovered by the SAM and C^A and C^N be the sets of adjective and noun clusters by a relational model, respectively. Cluster $\mathrm{SAM}_1 \in C^{\mathrm{SAM}}$ is described by noun cluster II$\in C^N$ and adjective clusters 1 and 3$\in C^A$, i.e. $\mathrm{SAM}_1 = (\{\mathrm{II}\}, \{1,3\})$. As this example shows, a relational model could help us to analyze adjectives and nouns independently. For this reason, we propose a relational model based on the IRM for word co-occurrence in the following section.

5. Clustering of Word Co-occurrences with Frequency-based IRM

5・1 Frequency-based Infinite Relational Model

We propose a frequency-based infinite relational model (FIRM). Figure 4 shows the difference of inputs between the FIRM and the IRM. As this figure shows, the FIRM takes into account the frequency of co-occurrences, which are statistically informative. On the other hand, the IRM assumes we just have a single observation for each relation, thereby ignoring frequency.

The generative model of the FIRM is described as,

$$\begin{array}{rcll} a|\boldsymbol{u}^A & \sim & \mathrm{Mult}(a;\boldsymbol{u}^A) & (8)\\ n|\boldsymbol{u}^N & \sim & \mathrm{Mult}(n;\boldsymbol{u}^N) & (9)\\ Z^A|\gamma & \sim & \mathrm{DP}(\gamma) & (10)\\ Z^N|\gamma & \sim & \mathrm{DP}(\gamma) & (11)\\ \eta(t^A,t^N)|\beta & \sim & \mathrm{Beta}(\eta(t^A,t^N);\beta,\beta) & (12)\\ (a,n)|z_a^A,z_n^N,\eta & \sim & \mathrm{Bern}(1;\eta(z_a^A,z_n^N)). & (13) \end{array}$$

See Table 1 for the notation. If Bernoulli in (13) returns one, the model generates a co-occurrence (a,n), otherwise, it generates a negative relation. One may notice that we can not observe negative relations. In other words, corpora include only positive relations. So, in our model, we regard missing relations as negative ones as the case of the IRM. Finally, the model is described as,

$$p(D|Z^A,Z^N,\eta) = \prod_{a\in A}\prod_{n\in N} \boldsymbol{u}_a^A \boldsymbol{u}_n^N \eta(z_a^A,z_n^N)^{f(a,n)} \\ (1-\eta(z_a^A,z_n^N))^{I(f(a,n)=0)}, \quad (14)$$

where $I(\cdot)$ is the indicator function.

As we will see later, the word co-occurrence dataset is too huge to apply MCMC. Therefore, we utilize variational inference for the FIRM. Blei proposed variational inference for Dirichlet process (VDP), and showed that a variational inference algorithm is much more efficient than a DP sampler [Blei 05]. Although the FIRM is not a simple DP mixture, we can derive a variational inference algorithm for the FIRM as we explain in Section 5・3.

FIRM

	n1	n2	n3
a1	○:8	○:3	○:5
a2	×	○:6	×
a3	×	×	○:2

IRM

	n1	n2	n3
a1	○	○	○
a2	×	○	×
a3	×	×	○

Fig. 4 An example of inputs of the IRM and the FIRM. For example, "○:8" at $(a1, n1)$ means that co-occurrence $(a1, n1)$ occurs eight times in the dataset. The FIRM allows multiple observations as in the left figure. On the other hand, the IRM accepts only one observation for one relation as in the right figure.

5・2 Variational Dirichlet Process

Variational inference is an alternative to sampling methods for Bayesian learning [?] especially in the context of large-scale problems. Blei has derived variational inference for the Dirichlet process [Blei 05] in the stick–breaking (SB) representation [Sethuraman 94]. The SB representation introduces random parameters $\boldsymbol{v} = (v_1, v_2, ...)$, which give a multinomial parameter for mixtures, $\boldsymbol{\pi}(\boldsymbol{v}) = (\pi_1(\boldsymbol{v}), ...)^T$ where $\pi_t(\boldsymbol{v}) = v_t \prod_{s=1}^{t-1}(1-v_s)$. The Dirichlet process in the SB representation is represented as,

$$\begin{array}{lcll} v_t|\gamma & \sim & \mathrm{Beta}(v_t;1,\gamma), & \text{for } t=1,...\\ \theta_t|G_0 & \sim & G_0, & \text{for } t=1,...\\ z_i|\boldsymbol{v} & \sim & \mathrm{Mult}(z_i;\boldsymbol{\pi}(\boldsymbol{v})), & \text{for } i=1,...,m\\ x_i|z_i,\theta & \sim & p(x_i|\theta_{z_i}), & \text{for } i=1,...,m \end{array}$$

where θ_t is the parameter of the tth component, G_0 is a base distribution for the Dirichlet process, and z_i and x_i are the ith assignment and observation, respectively.

VDP infers $q(Z,\theta,\boldsymbol{v})$ as an approximate posterior, $p(Z,\theta,\boldsymbol{v}|D)$, assuming the following factorization,

$$q(Z,\theta,\boldsymbol{v}) = \prod_{t=1}^{T-1} q(v_t) \prod_{t=1}^{T} q(\theta_t) \prod_{i=1}^{m} q(z_i), \quad (15)$$

where T is a truncation level. At truncation level T, VDP assumes $p(v_T=1)=1$ and $q(v_T=1)=1$. This assumption leads to $\pi_t = 0$ for all $t > T$. Therefore,

```
(1) Input: data $D$, hyperparameters $(\alpha, \beta, \gamma, T^A, T^N)$
    and initial $q(Z^A)$ and $q(Z^N)$
(2) Until $\mathcal{B}(D)$ converges
      a  Update $q(\boldsymbol{v}^A)$, $q(\boldsymbol{v}^N)$, $q(\eta)$, $q(\boldsymbol{u}^A)$ and $q(\boldsymbol{u}^N)$
      b  Update $q(Z^A)$ and $q(Z^N)$
(3) End
(4) Output: $q(Z^A)$ and $q(Z^N)$
```

Fig. 5 Algorithm: Variational Inference for FIRM. See Appendix A for the equations of q.

the infinite mixture boils down to a finite mixture. Note that if we set T large enough, the approximation is quite good in practice. This is called the truncated Dirichlet process [Ishwaran 01].

Using Jensen's inequality, we find a lower bound of $\log p(D)$,

$$\log p(D) \geq E\left[\log \frac{p(D, Z, \theta, \boldsymbol{v})}{q(Z, \theta, \boldsymbol{v})}\right]_{q(Z,\theta,\boldsymbol{v})} \tag{16}$$

where $E[f(x)]_{g(x)} = \int dx\, g(x) f(x)$. The approximate posterior, q, is derived by taking the variation of (16).

5·3 Variational Inference for FIRM

Let $W = (Z^A, Z^N, \boldsymbol{v}^A, \boldsymbol{v}^N, \eta, \boldsymbol{u}^A, \boldsymbol{u}^N)$, which is a set of hidden variables. We are interested in inferring the posterior distribution, $p(W|D)$. Using variational inference, we approximate the posterior as $q(W)$. First, we make the following bound of $\log p(D)$,

$$\log p(D) \geq E\left[\log \frac{p(D, W)}{q(W)}\right]_{q(W)} \equiv \mathcal{B}(D). \tag{17}$$

We update the approximate posterior, q, iteratively regarding $\mathcal{B}(D)$ as the objective function.

To make the approximate posterior, q, tractable, we assume the following factorization,

$$q(W) = q(Z^A)q(Z^N)q(\boldsymbol{v}^A)q(\boldsymbol{v}^N)q(\eta)q(\boldsymbol{u}^A)q(\boldsymbol{u}^N) \tag{18}$$

As (8)–(13) show, the joint distribution of the FIRM is factorized as follows,

$$\begin{aligned} p(D, W) =& p(D|Z^A, Z^N, \eta)p(Z^A|\boldsymbol{v}^A)p(Z^N|\boldsymbol{v}^N) \\ & \times p(\boldsymbol{v}^A)p(\boldsymbol{v}^N)p(\eta)p(\boldsymbol{u}^A)p(\boldsymbol{u}^N). \end{aligned} \tag{19}$$

We put the following priors into (19).

$$\eta(t^A, t^N) \sim \mathrm{Beta}(\eta(t^A, t^N); \beta, \beta) \tag{20}$$

$$v^A_{t^A}|\gamma \sim \mathrm{Beta}(v^A_{t^A}; 1, \gamma), \quad v^N_{t^N}|\gamma \sim \mathrm{Beta}(v^N_{t^N}; 1, \gamma) \tag{21}$$

$$Z^A|\boldsymbol{\pi}(\boldsymbol{v}^A) \sim \mathrm{Mult}(Z^A; \boldsymbol{\pi}(\boldsymbol{v}^A)) \tag{22}$$

$$Z^N|\boldsymbol{\pi}(\boldsymbol{v}^N) \sim \mathrm{Mult}(Z^N; \boldsymbol{\pi}(\boldsymbol{v}^N)) \tag{23}$$

$$\boldsymbol{u}^A \sim \mathrm{Dir}(\boldsymbol{u}^A; \alpha), \qquad \boldsymbol{u}^N \sim \mathrm{Dir}(\boldsymbol{u}^A; \alpha) \tag{24}$$

Note that (10) and (11) lead to (21)–(23) in stick–breaking representation. Taking the variation of (17), we will find the approximate posterior, q. See Appendix A for the derivation. We summarize the algorithm in Figure 5.

We have derived a variational inference algorithm for the FIRM. We also apply variational inference to the IRM in experiments while it is quite similar to derive the variational IRM.

6. Experimental Results

In this section, we see how the FIRM can partition co-occurrences. We use Mainichi newspaper 1993–2002 as a dataset. We extract adjective–noun co-occurrences by CaboCha, a Japanese dependency structure analyzer [Kudo 03]. The dataset has more than one million co-occurrences consisting of 210,605 distinct co-occurrences, 1,291 adjectives and 3,705 nouns. First of all, we apply the semantic aggregate model (SAM) [Mochihashi 02]. We compare the frequency-based infinite relational model (FIRM) with the infinite relational model (IRM) [Kemp 06] using the results of the SAM.

We first conducted an experiment using the SAM. The model was trained by a variational Bayesian algorithm [Nakagawa 06]. We set the number of clusters, K, to 50. Although results are affected by K, Nakagawa et al. showed by psychological experiments that $K = 50$ gives well-organized clusters on this dataset. The experiment was repeated 30 times, then we chose the best result in terms of the free energy. Each trial of the experiment took less than 15 minutes[*4]. One discovered cluster is shown in Table 2.

Next, we applied relational models. For both of the IRM and the FIRM, we set truncation levels T^A and T^N to 80 and 120, and set hyperparameters (α, β, γ) to $(1, 0.1, 1)$. This experiment was repeated 30 times. Each trial of the experiment took less than 5 minutes. The best results of 30 trials in terms of the free energy are depicted in Figure 6 and Figure 7 with the most likely Z^A and Z^N, which maximize $q(Z^A)$ and $q(Z^N)$. Each row is one adjective, and each column is one noun. Clusters are ordered in descending order, i.e. the top–most and left–most clusters are the largest clusters. Each black dot represents the existence of an adjective-noun co-occurrence. For example, dense cells show the strength of the relations

[*4] We conducted all experiments on Opteron 254 and SuSE Linux 10.

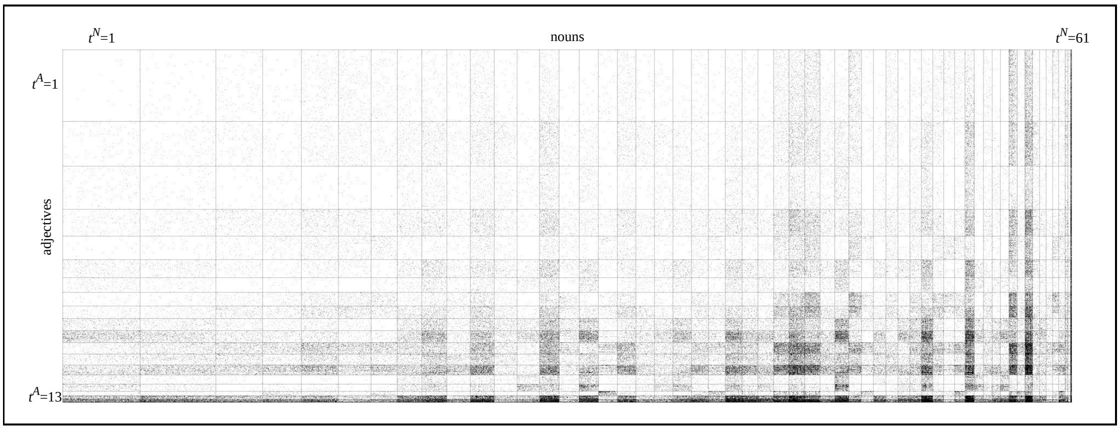


Fig. 6 Clustering results by the IRM. 13 adjective clusters and 61 noun clusters were discovered.

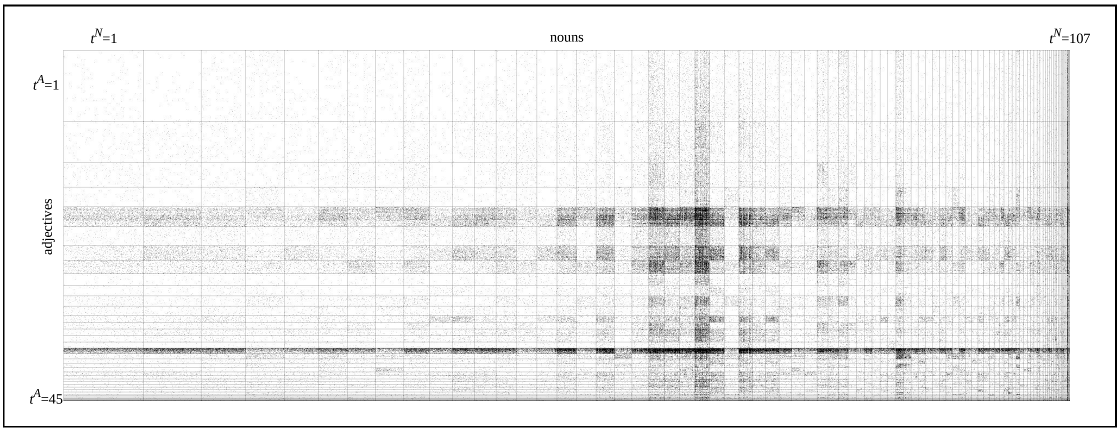


Fig. 7 Clustering results by the FIRM. 45 adjective clusters and 107 noun clusters were discovered.

Table 2 One cluster discovered by the semantic aggregate model.

c=11			
adjective		noun	
$p(c\|w)$	w	$p(c\|w)$	w
0.91	blackish	0.98	powder
0.85	red	0.95	t-shirt
0.83	cardinal red	0.95	shirt
0.82	whitish	0.93	ribbon
0.81	yellow	0.93	plastic
0.81	white	0.91	pants
0.80	smooth	0.90	jacket
0.79	blue	0.89	yellow
0.73	black	0.89	car
0.71	halting	0.89	rose

between adjective clusters and noun clusters.

7. Discussion

First of all, we discuss the results of the FIRM and the IRM. To compare them, we use the SAM as the gold standard, which has been verified by psychological experiments [Nakagawa 06]. Nakagawa et al. report that $p(a|n)$ inferred from the SAM and the results of psychological experiments are highly correlated (the correlation coefficient is 0.555). Next, we show that the FIRM has an advantage over the SAM in analyzing clusters.

To see to what extent the IRM and the FIRM capture the SAM, we evaluate them in terms of *coverage* and *purity* defined with the SAM.

Coverage shows how many SAM clusters are discovered, whose definition is (#covered SAM clusters) / (#SAM clusters). Let's say we are looking at a cell specified by adjective cluster t^A and noun cluster t^N, cell(t^A, t^N). cell(t^A, t^N) contains adjective–noun co-occurrences, $\{(a,n)|a \in t^A \text{ and } n \in t^N\}$. We can predict the most likely SAM cluster for each co-occurrences by $p(c|a,n)$. Let $d(c, t^A, t^N)$ be the number of co-occurrences in cell(t^A, t^N) whose most likely SAM cluster is c. A SAM cluster c is covered if and only if $\exists t^A, t^N \; c = \arg\max_j d(j, t^A, t^N)$.

We show the distribution of covered SAM clusters in Figure 8. The FIRM discovered clusters which cover all of the SAM clusters, i.e. 100% *coverage*, although clusters discovered by the IRM cover 86% of the SAM clusters, i.e. 86% *coverage*.

Purity is defined for each cell. Cells which have high *purity* consist of co-occurrences that have the same

Table 3 Clusters that are not discovered by the SAM but discovered by relational models.

adjective cluster	new, good, different,...
noun cluster	thing, object, place,...

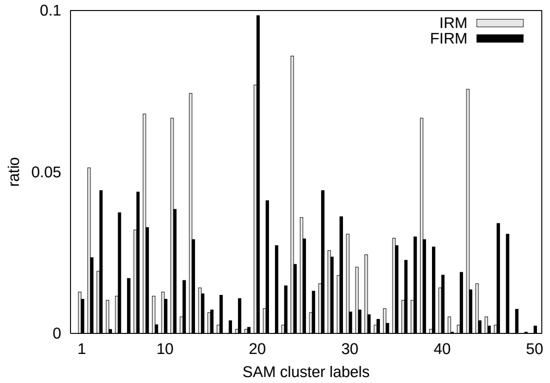


Fig. 8 Distributions of cells over the SAM clusters. Y axis shows the ratio of the number of cells that are covered by each SAM cluster. Taking a look at X axis, the IRM and the FIRM covered 43 SAM clusters and 50 SAM clusters, respectively. *Coverage* of the IRM and the FIRM is 86% (= 43/50) and 100% (= 50/50).

most likely SAM cluster. For example, when a cell has *purity* 0.9, 90% of co-occurrences in the call has the same most likely SAM cluster (see [Zhao 01] for a more general definition of *purity*). More formally, *purity* $S_i(t^A, t^N)$ is defined as

$$S_i(t^A, t^N) = \frac{\sum_{j=1}^{i} \tilde{d}(j, t^A, t^N)}{\sum_j \tilde{d}(j, t^A, t^N)} \quad \text{for } i = 1, ..., K. \quad (25)$$

where K is the number of clusters of the SAM and $\tilde{d}$ is sorted d in descending order, i.e.

$$\tilde{d}(1, t^A, t^N) > \tilde{d}(2, t^A, t^N) > ... \quad (26)$$

For example, $S_1(t^A, t^N)$ is the ratio of co-occurrences that belong to the SAM cluster covering cell(t^A, t^N), and $S_2(t^A, t^N)$ is the ratio of co-occurrences that belong to the covering cluster or the second largest cluster. We also define *purity* of assignments,

$$S_i(Z^A, Z^N) = \sum_{t^A=1}^{T^A} \sum_{t^N=1}^{T^N} \frac{\left\{\sum_j d(j, t^A, t^N)\right\} S_i(t^A, t^N)}{|A||N|} \quad (27)$$

We plot *purity* of assignments in Figure 9 varying a hyperparameter , $\beta = 1, 0.1$ and 0.01. For every β, the FIRM achieved higher *purity* over $i = 1$ to 10.

We have seen that the FIRM gives better clustering results than the IRM in terms of *coverage* and *purity*. One may suspect that better *coverage* and *purity* is simply because the FIRM has discovered more clusters than the IRM. However, please note that the number of clusters was given by inference.

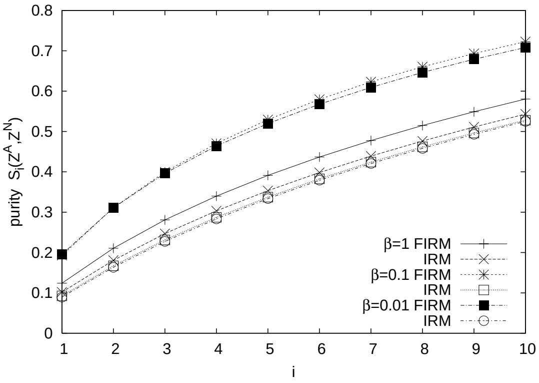


Fig. 9 *Purity* $S_i(Z^A, Z^N)$ varying hyperparameter β.

For both of the IRM and the FIRM, *purity* of whole assignments was not very high because *purity* of some cells are quite small. The reason is that relational models also find clusters that are not discovered by the SAM like Table 3. It is easy to imagine that these words make co-occurrences with many words. Therefore, these clusters are uninformative for the SAM, and the SAM does not find them as clusters.

Figure 10 shows a concrete example of the FIRM. Each cell covers SAM cluster 11 shown in Table 2. It seems that the adjectives in Table 2 are similar to cluster B in Figure 10. However, we notice that cells a–A and a–C also belong to SAM cluster 11. As this example shows, we can independently interpret adjectives and nouns with the FIRM whereas the SAM only gives an interpretation of “co-occurrences.” Therefore, the FIRM seems more helpful to analyze adjectives and nouns independently.

In summary, the FIRM has achieved better *purity* and *coverage* than the IRM, i.e. the FIRM clusters have captured the SAM clusters better than the IRM clusters. The FIRM has also found clusters which are not discovered by the SAM. The clusters of adjectives and nouns discovered by the FIRM seem more helpful to analyze adjectives and nouns independently than the SAM. We strongly believe that these advantages of the FIRM improve the discovery of latent concepts in applications, for example metaphor understanding proposed in [Nakagawa 06].

8. Related Work

Biclustering has been applied to gene expression data [Cheng 00]. It partitions both of genes and con-

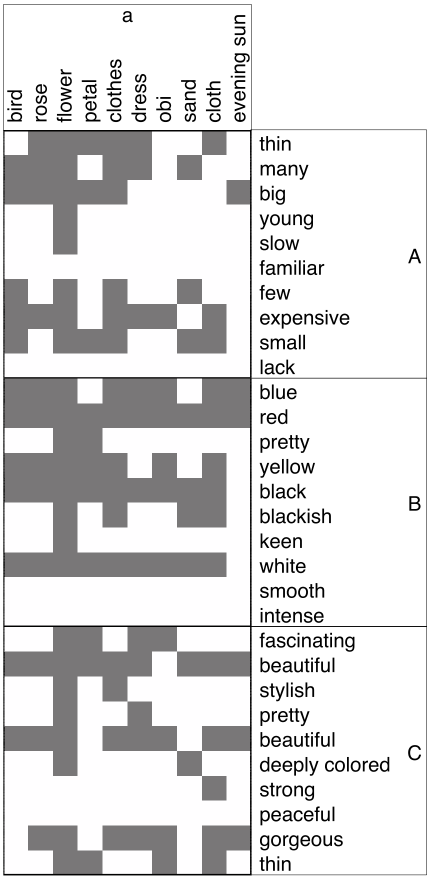


Fig. 10 Clusters discovered by the FIRM in Figure 7. Each cell covers SAM cluster 11. $S_1(a,A) = 0.81$, $S_1(a,B) = 0.98$ and $S_1(a,C) = 0.77$.

ditions simultaneously as the FIRM partitions adjectives and nouns. The objective function of biclustering in [Cheng 00] is similar to the least square error, but it is modified for biclustering. Since the FIRM models discrete data, we can not apply the FIRM to continuous data like gene expression data. However, it is possible to derive a stochastic blockmodel that handles continuous data by using a continuous distribution instead of Bernoulli in (13).

We can find other related work in document clustering. Slonim and Tishby applied the information bottleneck method [Tishby 99] to find document clustering using word clusters [Slonim 00]. Their algorithm first finds word clusters, then partitions document using the word clusters. The latent Dirichlet allocation (LDA) also partitions documents and words simultaneously [Blei 03]. The LDA samples a document in the following procedure, 1. Generate a parameter of multinomial for word topics, 2. Sample topics according to the parameter, 3. Sample words according the topics. We notice that we can also apply the FIRM to document–word clustering. However, one difference is that the FIRM treats document clusters and word clusters equally, i.e. no order of sampling documents and words, but the other methods do not treat them equally. Therefore, it should be interesting to compare results of these algorithms.

9. Conclusion and Future Work

We have proposed to make clusters or each part of speech from word co-occurrence data. To take this approach, we have proposed a model called frequency-based infinite relational model (FIRM). It takes into account the frequency of observations unlike traditional stochastic blockmodels which ignore it. We have also derived a variational inference algorithm to apply the FIRM to a large dataset.

The FIRM discovers adjective clusters and noun clusters. As a result, the FIRM is more useful to analyze adjectives and nouns independently than the SAM. We also experimentally showed that the FIRM clusters capture the SAM clusters better than the IRM clusters. Moreover, the FIRM found clusters which were not discovered by the SAM.

It is straight forward to apply the FIRM to feature-rich datasets e.g. co-occurrences of subject–verb, verb–objective and adjective–noun. It should be interesting to see clustering results of such a dataset.

In this study, we treated missing relations as negative samples. However, it is desirable to model missing relations in a more suitable way as [Cussens 01]. This modeling remains to be investigated.

Acknowledgments

This research was funded by the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources."

♢ References ♢

[Antoniak 74] Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, Vol. 2, pp. 1152–1174 (1974)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Blei 05] Blei, D. M. and Jordan, M. I.: Variational Inference for Dirichlet Process Mixtures, *Bayesian Analysis*, Vol. 1, No. 1, pp. 121–144 (2005)

[Cheng 00] Cheng, Y. and Church, G. M.: Biclustering of Expression Data, in *ISMB*, pp. 93–103 (2000)

[Cussens 01] Cussens, J.: Parameter Estimation in Stochastic Logic Programs, *Machine Learning*, Vol. 44, No. 3, pp. 245–271 (2001)

[Ferguson 73] Ferguson, T.: A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, Vol. 1, pp. 209–230 (1973)

[Ishwaran 01] Ishwaran, H. and James, L. F.: Gibbs Sampling Methods for Stick Breaking Priors, *Journal of the American Statistical Association*, Vol. 96, No. 453, pp. 161–173 (2001)

[Kemp 06] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N.: Learning Systems of Concepts with an Infinite Relational Model, in *AAAI* (2006)

[Kudo 03] Kudo, T. and Matsumoto, Y.: Fast Methods for Kernel-Based Text Analysis, in Hinrichs, E. and Roth, D. eds., *the 41st Annual Meeting of the ACL*, pp. 24–31 (2003)

[Mochihashi 02] Mochihashi, D. and Matsumoto, Y.: Probabilistic Representatin of Meaning, in *IPSJ-NL*, Vol. 4, pp. 77–84 (2002)

[Nakagawa 06] Nakagawa, M., Terai, A., and Sato, T.: A Computational Model of Metaphor Understanding Using a Statitical Analysis of Japanese Corpora Based on Soft Clustering – Toward a Metaphorical Search Engine –, in *Framework for Systematization and Application of Large-scale Knowledge Resources* (2006)

[Pereira 93] Pereira, F., Tishby, N., and Lee, L.: Distributional Clustering of English Words, in *31st Annual Meeting of the ACL*, pp. 183–190 (1993)

[Sethuraman 94] Sethuraman, J.: A constructive definition of Dirichlet priors, *Statistica Sinica*, Vol. 4, pp. 639–650 (1994)

[Slonim 00] Slonim, N. and Tishby, N.: Document clustering using word clusters via the information bottleneck method, in *ACM SIGIR*, pp. 208–215 (2000)

[Tishby 99] Tishby, N., Pereira, F. C., and Bialek, W.: The Information Bottleneck Method, in *The 37th annual Allerton Conference on Communication, Control, and Computing* (1999)

[Zhao 01] Zhao, Y. and Karypis, G.: Criterion functions for document clustering: Experiments and analysis, Technical Report Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis (2001)

〔担当委員：伊藤 公人〕

Received August 15, 2006.

♢ Appendix ♢

A. Derivation of Variational Inference

In this section, we derive the approximate posterior q. Taking the variation of (17), we find q to have the following form.

$$q(v_t^A) \propto \exp E[\log p(D,W)]_{q(Z^A)} \tag{A.1}$$

$$q(v_t^N) \propto \exp E[\log p(D,W)]_{q(Z^N)} \tag{A.2}$$

$$q(\eta(t^A,t^N)) \propto \exp E[\log p(D,W)]_{q(Z^A,Z^N)} \tag{A.3}$$

$$q(\boldsymbol{u}^A) = p(\boldsymbol{u}^A|D) \tag{A.4}$$

$$q(\boldsymbol{u}^N) = p(\boldsymbol{u}^N|D) \tag{A.5}$$

$$q(z_a^A) \propto \exp E[\log p(D,W)]_{q(Z^N,\eta,v^A)} \tag{A.6}$$

$$q(z_n^N) \propto \exp E[\log p(D,W)]_{q(Z^A,\eta,v^N)}. \tag{A.7}$$

After some algebra, (A.1)–(A.5) lead to,

$$q(v_t^A) = \mathrm{Beta}(v_t^A;\gamma_{1,t}^A,\gamma_{2,t}^A) \qquad q(v_t^N) = \mathrm{Beta}(v_t^N;\gamma_{1,t}^N,\gamma_{2,t}^N)$$

$$q(\eta(t^A,t^N)) = \mathrm{Beta}(\eta(t^A,t^N);\tau_{1,t^A,t^N},\tau_{2,t^A,t^N})$$

$$q(\boldsymbol{u}^A) = \mathrm{Dir}(\boldsymbol{u}^A;\boldsymbol{\sigma}^A) \qquad q(\boldsymbol{u}^N) = \mathrm{Dir}(\boldsymbol{u}^N;\boldsymbol{\sigma}^N),$$

where

$$\gamma_{1,t^A}^A = 1 + m_{t^A}^A, \qquad \gamma_{2,t^A}^A = \gamma + \sum_{j=t^A+1}^{T^A} m_j^A$$

$$\gamma_{1,t^N}^N = 1 + m_{t^N}^N, \qquad \gamma_{2,t^N}^N = \gamma + \sum_{j=t^N+1}^{T^N} m_j^N$$

$$m_{t^A}^A = \sum_{a\in A} q(z_a^A = t^A), \qquad m_{t^N}^N = \sum_{n\in N} q(z_n^N = t^N)$$

$$\tau_{1,t^A,t^N} = \beta + \sum_{a\in A}\sum_{n\in N} q(z_a^A = t^A) q(z_n^N = t^N) f(a,n)$$

$$\tau_{2,t^A,t^N} = \beta + \sum_{a\in A}\sum_{n\in N} q(z_a^A = t^A) q(z_n^N = t^N) I(f(a,n) = 0)$$

$$\boldsymbol{\sigma}_a^A = \alpha + \sum_{n\in N} f(a,n), \qquad \boldsymbol{\sigma}_n^N = \alpha + \sum_{a\in A} f(a,n).$$

Note that we set the truncation level of adjective clusters and noun clusters to T^A and T^N, respectively.

Author's Profile

Kurihara, Kenichi (Member)

Mr. Kurihara is a graduate student at Department of Computer Science, Tokyo Institute of Technology. He is also a JSPS research fellow. His research interests are machine learning and natural language processing. He is a member of JSAI.

Kameya, Yoshitaka (Member)

Dr. Kameya is a research associate of Department of Computer Science, Tokyo Institute of Technology. He received Ph.D. in computer science from Tokyo Institute of Technology in 2000. He worked for NS solutions in 2001–2003. His research interests are machine learning, natural language processing, probabilistic inference system etc. He is a member of JSAI.

Sato, Taisuke (Member)

Dr. Sato is a professor of Department of Computer Science, Tokyo Institute of Technology. He received B.E., M.E. and Ph.D. from Tokyo Institute of Technology. He worked for Electrotechnical Laboratory in 1975–1995. His research interests are logic programming, AI, etc. He is a member of JSAI, IPSJ and EATCS.