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 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 arisen from one latent concept. This assumption seems moderately natural to make clusters of co-occurrences. However, to analyze latent concepts more deeply, the assumption may be too restrictive. We propose to make clusters independently for each part of speech. For example, we make adjective clusters and noun clusters form adjective–noun co-occurrences while the SAM builds clusters of adjective–noun pairs. This would lead to more specific clusters than the SAM.

In this paper, we propose a frequency-based infinite relational model (FIRM) for word co-occurrences. The FIRM is similar to the infinite relational model (IRM) proposed in statistical relational learning. Our model differs in the use of frequency. Since the IRM is proposed for clustering relation, it ignores multiple observations. We also derive variational inference methods for these models to apply to a large dataset. Experimental results show that the FIRM gives better clustering results than the IRM in terms of the high resolution compared to the SAM.

1. Introduction

Clustering word co-occurrences has been studied to discover latent concepts. Pereira et al. write discovered clusters seem semantically significant [Pereira 93]. Mochihashi and Matsumoto experimentally discover 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 is from one latent concept. This assumption seems moderately natural to make clusters of co-occurrences. However, to analyze latent concepts more deeply, the assumption may be too restrictive.

The goal of this study is to discover more specific clusters as concepts than the semantic aggregate model (SAM). We propose to make clusters independently for each part of speech (POS). For example, we make adjective clusters and noun clusters form adjective–noun co-occurrences while the SAM builds clusters of adjective–noun pairs. This would lead to more specific clusters than the SAM.

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

Word co-occurrences can also be regarded as re-

*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/

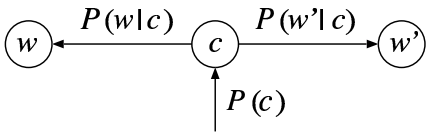

Fig. 1 Graphical representation of Semantic Aggregate Model.

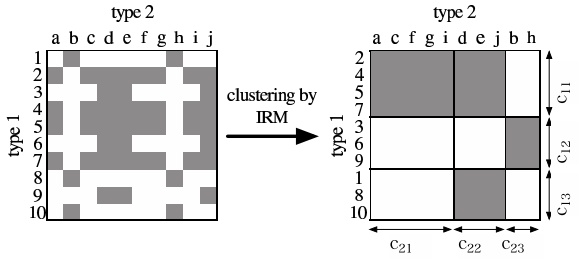

Fig. 2 A toy example of the infinite relational model (IRM). The IRM partitions type 1 (1–10) into clusters $c_{11} - c_{13}$ and type 2 (a–j) into $c_{21} - c_{23}$.

lation. Suppose adjective a and noun n makes 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. Although the SAM models latent concepts of pairs of words, relational models directly discover latent concepts of POS, which are more specific latent concepts than that of the SAM.

In this paper, we propose a frequency-based infinite relational model (FIRM) for word co-occurrences. The FIRM is similar to the IRM. Our model differs in the use of frequency. Since the IRM is proposed for clustering relation, it does not take into account multiple observations, which are statistically informative. We also derive variational inference methods for these models to apply to one million datacases. Since it has experimentally been shown that the SAM makes clusters that are consistent with psychological experiments [Nakagawa 06], we evaluate the FIRM and the IRM using the SAM as the gold standard. Experimental results show that the FIRM gives better clustering results than the IRM in terms of the high resolution compared to the SAM.

2. Semantic Aggregate Model

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

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

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 context 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 the following similarity measure between word w and word w',

$$\delta(w, w') = e^{-KL(w||w')}, \tag{3}$$

where $KL(w||w')$ is the Kullback–Liebler divergence,

$$KL(w||w') = \sum_{c=1}^{K} p(c|w) \log \frac{p(c|w)}{p(c|w')}. \tag{4}$$

Nakagawa et al. conducted psychological experiments on the clustering results by the SAM, and showed that the clusters are consistent with the of psychological experiments [Nakagawa 06]. Therefore, we evaluate co-occurrence models by comparing 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 types into clusters. Figure 2 is a toy example. The left hand side matrix shows relation R : (type 1) $\times$ (type 2) $\to \{0, 1\}$ (each black and white dot shows $R(\cdot,\cdot) = 1$ and $R(\cdot,\cdot) = 0$, respectively). 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 cluster c_{11}–c_{13}, and type 2 is also partitioned into cluster c_{21}–c_{23}. For example, type 1, type 2 and R can be a set of people, a set of

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

animals and predicate "like", i.e. $R(i,j)=1$ means person i likes animal j. Although Figure 2 uses only one relation over type 1 and type 2, the IRM can handles the arbitrary number of relations over the arbitrary number of types.

We apply the IRM to word co-occurrence clustering. The dataset we use consists of adjective–noun pairs. Therefore, we consider a relation between adjectives and nouns, $R:\{\text{adjectives}\}\times\{\text{nouns}\}\to\{0,1\}$. If adjective–noun pair (a,n) exists in a dataset, $R(a,n)=1$, otherwise $R(a,n)=0$. Kemp et al. writes that they ignore missing relation whereas we regard them as negative relations, i.e. $R(a,n)=0$. This is because the number of observed adjective–noun pairs is quite smaller than that of possible pairs*3

The IRM for adjective–noun co-occurrence is modeled by,

$$Z^A|\gamma \sim \mathrm{DP}(\gamma) \quad (5)$$

$$Z^N|\gamma \sim \mathrm{DP}(\gamma) \quad (6)$$

$$\eta(t^A,t^N)|\beta \sim \mathrm{Beta}(\beta,\beta) \quad (7)$$

$$R(a,n)|z_a^A,z_n^N,\eta \sim \mathrm{Bernoulli}(\eta(z_a^A,z_n^N)), \quad (8)$$

where Z^A and Z^N are assignments*4 of adjectives and nouns, respectively,

$$Z^A=\{z_a^A|a\in A\},\qquad 1\le z_a^A\le T^A$$
$$Z^N=\{z_n^N|n\in N\},\qquad 1\le z_n^N\le T^N$$

and DP is the Dirichlet process [Ferguson 73, Antoniak 74]. Using the Dirichlet process, the IRM does not need to specify the number of clusters whereas the traditional stochastic blockmodel requires the number of clusters. Note that we can express $p(R|Z^A,Z^N,\eta)$ as

$$p(R|Z^A,Z^N,\eta)=\prod_{a\in A}\prod_{n\in B}\eta(z_a^A,z_n^N)^{R(a,n)}(1-\eta(z_a^A,z_n^N))^{1-R(a,n)}. \quad (9)$$

The inference of Z^A and Z^N can be carried out using Markov chain Monte Carlo (MCMC) methods to sample from the posterior on cluster assignments $p(Z^A,Z^N|R)$. Kemp et al. simply infer the best partition Z^A and Z^N by hill climbing [Kemp 06].

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

*4 In this paper, the Dirichlet process is represented in the stick-breaking representation [Sethuraman 94]. Therefore, we use labels in stead of partitions. See [Griffiths 05] their definitions.

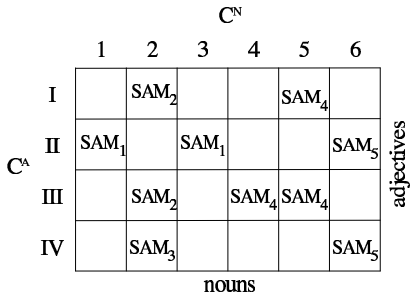

Fig. 3 A conceptual image of the relational clustering and clustering by the SAM. $C^{\mathrm{SAM}}=\{\mathrm{SAM}_1,...,\mathrm{SAM}_5\}$ is the clusters by the SAM. $C^A=\{\mathrm{I},...,\mathrm{IV}\}$ and $C^N=\{1,...,6\}$ are clusters of adjective and nouns by the a relational model. C^{SAM} is divided into the cells of C^A and C^N.

4. Semantic Aggregate Model and Relational Models

The semantic aggregate model (SAM) models latent clusters of word co-occurrences as concepts. The SAM implicitly makes an assumption that words in one co-occurrence have the same latent concept. For example, when a pair consists of an adjective and a noun, the assumption means that adjectives and nouns have the same concepts, see Figure 1. However, adjectives and nouns should have the different sets of concepts, intuitively.

Relational models like the IRM discovers latent clusters in each types. Therefore, in the case of word co-occurrences, it discovers latent clusters of part of speech (POS), e.g. adjective clusters and noun clusters from adjective–noun co-occurrences.

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. We expect that clusters by a relational model achieves higher resolution on C^{SAM} as an example in Figure 3 shows. In Figure 3, C^{SAM} is a subset of the product set of C^A and C^N, $C^{\mathrm{SAM}}\subset C^A\times C^N$. Using relational clusters C^A and C^N, cluster SAM_1 is described as a co-occurrence cluster consisting of noun cluster II and adjective clusters 1 and 3, $\mathrm{SAM}_1=(\{\mathrm{II}\},\{1,3\})$. Although this is just an example, we can say that a relational method discovered more specific clusters than the SAM in Figure 3. To achieve this, we improve 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

Although co-occurrence data has frequencies of co-occurrences, the infinite relational model (IRM) ignores frequency. This is because the IRM is proposed to find clusters over binary relations. However, word co-occurrence should not be a binary relation but should allow multiple observations.

We propose a frequency-based infinite relational model (FIRM). The FIRM allows multiple observations of the same co-occurrence. Its generative model is described as,

$$
\begin{aligned}
a|\boldsymbol{u}^A &\sim \text{Multinomial}(\boldsymbol{u}^A) && (10)\\
n|\boldsymbol{u}^N &\sim \text{Multinomial}(\boldsymbol{u}^N) && (11)\\
Z^A|\gamma &\sim \text{DP}(\gamma) && (12)\\
Z^N|\gamma &\sim \text{DP}(\gamma) && (13)\\
\eta(t^A,t^N)|\beta &\sim \text{Beta}(\beta,\beta) && (14)\\
(a,n)|z_a^A,z_n^N,\eta &\sim \text{Bernoulli}(\eta(z_a^A,z_n^N)). && (15)
\end{aligned}
$$

(a,n) denotes a co-occurrence. Since co-occurrences are generated by Bernoulli, co-occurrences may fail. Generative models with failure have been studied in [Cussens 01, Sato 04]. Instead of the exact inference with failure, we simplify this model like,

$$
p(D|Z^A,Z^N,\eta)=\prod_{a\in A}\prod_{n\in N}\boldsymbol{u}^A_{e_i^A}\boldsymbol{u}^N_{e_i^N}\eta(z_a^A,z_n^N)^{f(a,n)}(1-\eta(z_a^A,z_n^N))^{I(f(a,n)=0)}, \tag{16}
$$

where D is a training dataset, $f(a,n)$ is the number of observations of co-occurrence (a,n) and $I(\cdot)$ is the indicator function.

As we will see later, the word co-occurrence dataset is too huge to apply MCMC to. Therefore, we utilize a variational inference for the FIRM. A variational inference for Dirichlet process (VDP) has been proposed in [Blei 06]. They empirically showed that the variational inference was much more efficient than a DP sampler. Although the FIRM is not a simple DP mixture, we can derive a variational inference for the FIRM as we explain in the next section.

5•2 Variational Dirichlet Process

Variational inference methods are alternatives to sampling methods for Bayesian learning [Attias 00, Ghahramani 00] especially in the context of large-scale problems. Blei and Jordan have applied a variational inference for Dirichlet process [Blei 06] in the stick-breaking (SB) representation [Sethuraman 94]. The SB representation introduces random parameters $\boldsymbol{v}=(v_1,...)$. The Dirichlet process in the SB representation is represented as,

$$
\begin{aligned}
v_t|\alpha &\sim \text{Beta}(1,\alpha), && \text{for } t=1,...\\
\eta_t|G_0 &\sim G_0, && \text{for } t=1,...\\
z_i|\boldsymbol{v} &\sim \text{Multinomial}(\pi(\boldsymbol{v})), && \text{for } n=1,...,m\\
x_i|z_i,\eta &\sim p(x_i|\eta_{z_i}), && \text{for } n=1,...,m
\end{aligned}
$$

where η_t is the parameter of the tth component, $\pi_t(\boldsymbol{v})=v_t\prod_{s=1}^{t-1}(1-v_s)$, and z_i and x_i are the ith assignment and observation, respectively.

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

$$
q(Z,\eta,\boldsymbol{v})=\prod_{t=1}^{T-1}q(v_t)\prod_{t=1}^{T}q(\eta_t)\prod_{i=1}^{m}q(z_i), \tag{17}
$$

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

$$
p(X)\geq E\left[\log\frac{p(X,Z,\eta,\boldsymbol{v})}{q(Z,\eta,\boldsymbol{v})}\right]_{q(Z,\eta,\boldsymbol{v})} \tag{18}
$$

where $E[f(x)]_{g(x)}=\int dx\ g(x)f(x)$.

The approximate posterior, q, is derived by taking variation of (18) and setting to zero.

5•3 Variational Inference for Frequency-based IRM

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

$$
p(D)\geq E\left[\log\frac{p(D,W)}{q(W)}\right]_{q(W)}. \tag{19}
$$

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

$$
q(W)=q(Z^A)q(Z^N)q(\boldsymbol{v}^A)q(\boldsymbol{v}^N)q(\eta)q(\boldsymbol{u}^A)q(\boldsymbol{u}^N). \tag{20}
$$

As (10)–(15) show, the joint distribution of the FIRM is factorized as follows,

$$p(D,W) = p(D|Z^A,Z^N,\eta)p(Z^A|\boldsymbol{v}^A)p(Z^N|\boldsymbol{v}^N) \times p(\boldsymbol{v}^A)p(\boldsymbol{v}^N)p(\eta)p(\boldsymbol{u}^A)p(\boldsymbol{u}^N). \quad (21)$$

We put the following priors into (21).

$$p(\eta(t^A,t^N)) = \mathrm{Beta}(\eta(t^A,t^N);\beta,\beta)$$

$$p(Z^A|\boldsymbol{v}) = \prod_{a\in A}\pi^A_{z^A_a}, \qquad p(Z^N|\boldsymbol{v}) = \prod_{n\in N}\pi^N_{z^N_n}$$

$$p(v^A_{t^A}) = \mathrm{Beta}(v^A_{t^A};1,\gamma)$$

$$p(v^N_{t^N}) = \mathrm{Beta}(v^N_{t^N};1,\gamma)$$

$$p(\boldsymbol{u}^A) = \mathrm{Dirichlet}(\boldsymbol{u}^A;\alpha^A_0)$$

$$p(\boldsymbol{u}^N) = \mathrm{Dirichlet}(\boldsymbol{u}^N;\alpha^N_0)$$

Taking the variation of (19), we will find the approximate posterior, q,

$$q(v^A_t) = \mathrm{Beta}(v^A_t;\gamma^A_{1,t},\gamma^A_{2,t})$$

$$q(v^N_t) = \mathrm{Beta}(v^N_t;\gamma^N_{1,t},\gamma^N_{2,t})$$

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

$$q(z^A_a) \propto \exp E[\log p(R,Z^A,Z^N,\eta,\boldsymbol{v}^A,\boldsymbol{v}^N)]_{q(Z^A_{\neg a},Z^N,\eta,\boldsymbol{v}^A)}$$

$$q(z^N_n) \propto \exp E[\log p(R,Z^A,Z^N,\eta,\boldsymbol{v}^A,\boldsymbol{v}^N)]_{q(Z^A,Z^N_{\neg n},\eta,\boldsymbol{v}^N)}$$

$$q(\boldsymbol{u}^A) = \mathrm{Dirichlet}(\boldsymbol{u}^A;\boldsymbol{\alpha}^A)$$

$$q(\boldsymbol{u}^N) = \mathrm{Dirichlet}(\boldsymbol{u}^N;\boldsymbol{\alpha}^N),$$

where

$$\gamma^A_{1,t^A} = 1 + m^A_{t^A}, \qquad \gamma^A_{2,t^A} = \gamma + \sum_{j=t^A+1}^{T^A} m^A_j$$

$$\gamma^N_{1,t^N} = 1 + m^N_{t^N}, \qquad \gamma^N_{2,t^N} = \gamma + \sum_{j=t^N+1}^{T^N} m^N_j$$

$$m^A_{t^A} = \sum_{a\in A} q(z^A_a = t^A), \quad m^N_{t^N} = \sum_{n\in N} q(z^N_n = t^N)$$

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

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

$$\alpha^A_a = \alpha^A_0 + \sum_{n\in N} f(a,n), \quad \alpha^N_n = \alpha^N_0 + \sum_{a\in A} f(a,n).$$

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

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

Table 1 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

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 pairs by CaboCha, a Japanese dependency structure analyzer. The dataset has more than one million pairs consisting of 210,605 distinct pairs, 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 is trained by the variational Bayes [Nakagawa 06]. We set the number of clusters, K, to 50. Although results are affected by K, it was verified by psychological experiments that $K = 50$ gives well-organized clusters on this dataset in [Nakagawa 06]. 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*5. One discovered cluster is shown in Table 1.

Next, we applied relational models. For both of the IRM and the FIRM, we set truncation level T^A and T^N to 80 and 120, and set β to 0.1. 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 4 and Figure 5 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

*5 We conducted all experiments on Opteron 254 and Linux SuSE 10.

dot represents the existence of an adjective-noun co-occurrence. Therefore, dense cells show the strength of the relations between adjective clusters and noun clusters.

We discuss the results of the IRM and the FIRM in the next section. Note the results of the SAM is verified by psychological experiments [Nakagawa 06]. We compare the FIRM with the IRM using the SAM as the gold standard.

7. Discussion

The goal of this study is to discover more specific clusters as concepts than the semantic aggregate model (SAM). Relational models build clusters independently for POS. For example, we make adjective clusters and noun clusters form adjective–noun co-occurrences while the SAM builds clusters of adjective–noun pairs. This would lead to more specific clusters than the SAM.

To see how much the IRM and the FIRM achieve this goal, we evaluate them by comparing with the the SAM in terms of *coverage* and *purity*.

The *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 pairs, $\{(a,n)|a \in t^A \text{ and } n \in t^N\}$. We can predict the most likely SAM cluster for each pair by $p(c|a,n)$. Let $d(c, t^A, t^N)$ be the number of pairs in cell(t^A, t^N) whose most likely SAM cluster is c. A SAM cluster c is covered if and only if there exist t^A and t^N such that $c = \arg\max_j d(j, t^A, t^N)$.

We show the distribution of covered SAM clusters in Figure 6. The FIRM discovered clusters which cover all of the SAM clusters although clusters discovered by the IRM cover 86% of the SAM clusters.

Purity is defined for each cell. Cells which have high *purity* consist of pairs that have the same most likely SAM cluster. For example, when a cell has *purity* 0.9, 90% of pairs in the call has the same most likely SAM cluster (see [Zhao 01] for more general definition of the *purity*). More formally,

Purity $S_i(t^A, t^N)$ is defined as

$$S_i(t^A, t^N) = \frac{\sum_{j=1}^{i} \tilde{d}(j, t^A, t^N)}{\sum_j d(j, t^A, t^N)} \quad \text{for } i = 1, ..., K. \quad (22)$$

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

$$\tilde{d}(1, t^A, t^N) > \tilde{d}(2, t^A, t^N) > ... \quad (23)$$

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

$$S_i(Z^A, Z^N) = \sum_{t^A=1}^{T^A} \sum_{t^N=1}^{T^N} \frac{\left\{\sum_j d(j, t^A, t^N)\right\} S_i(t^A, t^N)}{|A||N|} \quad (24)$$

We plot the *purity* in Figure 7 varying a hyperparameter , $\beta = 1, 0.1$ and 0.01. For every β, the FIRM achieved higher *purity* over $i = 1$ to 10.

One may worry that the better *coverage* and the *purity* is because the FIRM discovered more clusters than the IRM. However, the number of clusters was given by inference. In other words, taking into account frequency enabled us to discover more clusters.

For both of the IRM and the FIRM, the *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 2. 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.

We have seen that the FIRM gave better clustering results than the IRM in terms of the *coverage* and the *purity*. Figure 8 shows a concrete example of the FIRM. Each cell covers SAM cluster 11 shown in Figure 1. It seems that the adjectives in Figure 1 are similar to cluster B in Figure 8. However, we notice that cells a–A and a–C cluster also belong to SAM cluster 11. Clearly, the FIRM discovered more specific clusters than the SAM.

Figure 8 is just one example of results. However, it is clear that the clustering results give higher resolution on clusters than the SAM. Moreover, we found that relational models discover clusters that are not discovered by the SAM, e.g. Table 2. We strongly believe that this higher resolution improves inference of concepts in applications, for example metaphor understanding proposed in [Nakagawa 06].

Table 2 Clusters that are not discovered by the SAM but discovered by relational models.

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

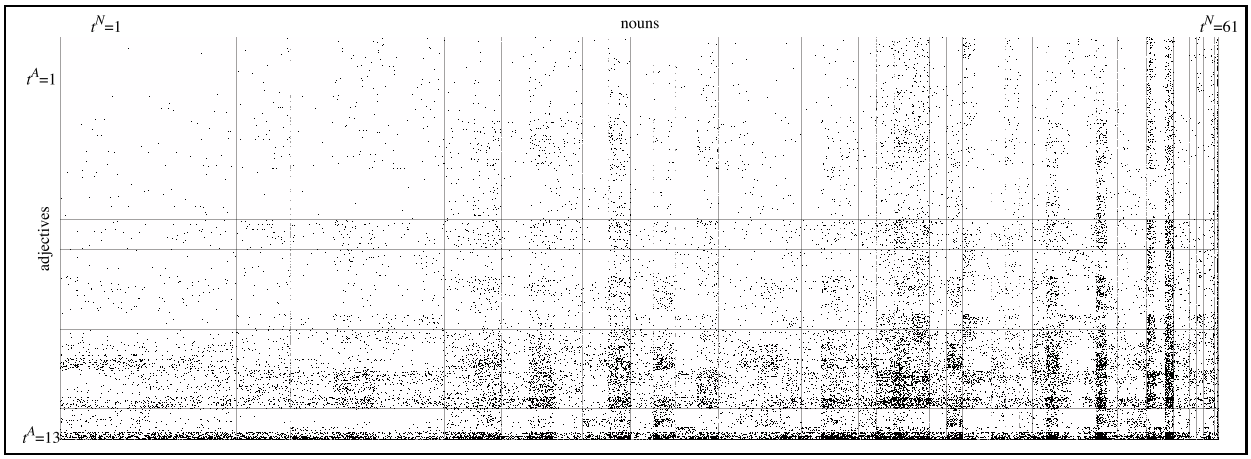

Fig. 4 Clustering results by the IRM. 13 adjective clusters and 61 noun clusters were discovered.

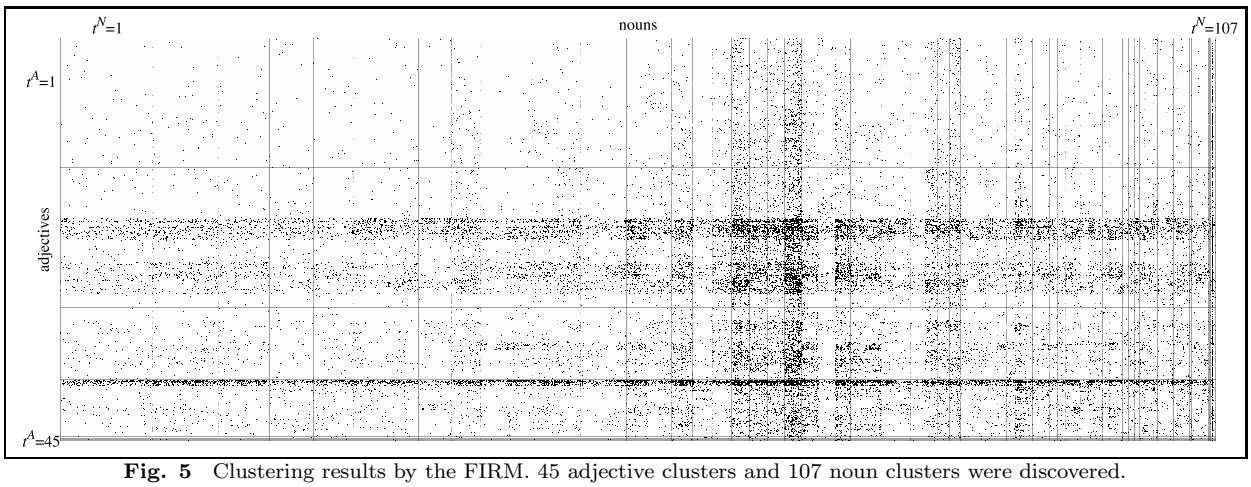

Fig. 5 Clustering results by the FIRM. 45 adjective clusters and 107 noun clusters were discovered.

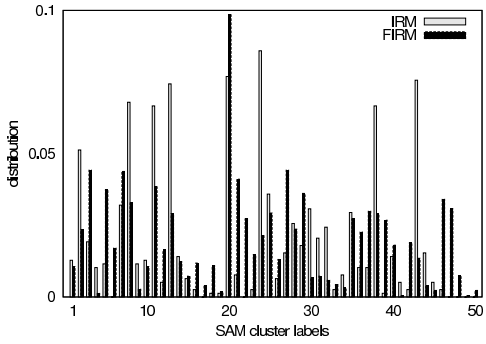

Fig. 6 Distributions of cells over the SAM clusters. Y axis shows the ratio of the number of cells that are covered by each SAM cluster. The IRM and the FIRM covered 43 SAM clusters and 50 SAM clusters, respectively. The *coverage* of the IRM and the FIRM is 86% (= 43/50) and 100% (= 50/50).

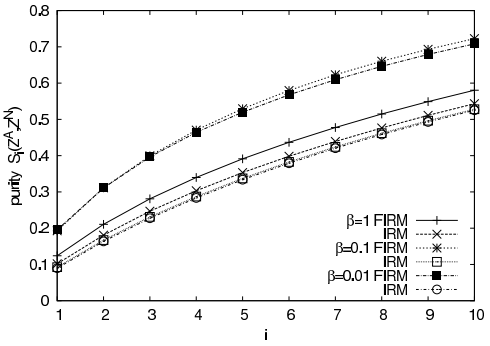

Fig. 7 *Purity* $S_i(Z^A, Z^N)$ varying β.

8. Conclusion and Future Work

We experimentally showed that clustering word co-occurrences with relational models gives higher resolution on word clusters as concepts than a previously proposed model, the semantic aggregate model (SAM). To achieve better results, we proposed a frequency-based infinite relational model (FIRM). we also derived a variational inference methods to apply the models to a large dataset. Since it has experimentally been shown that the SAM makes clusters that are consistent with psychological experiments, we evaluate the FIRM and the IRM using the SAM as the gold standard. Experimental results show that the FIRM

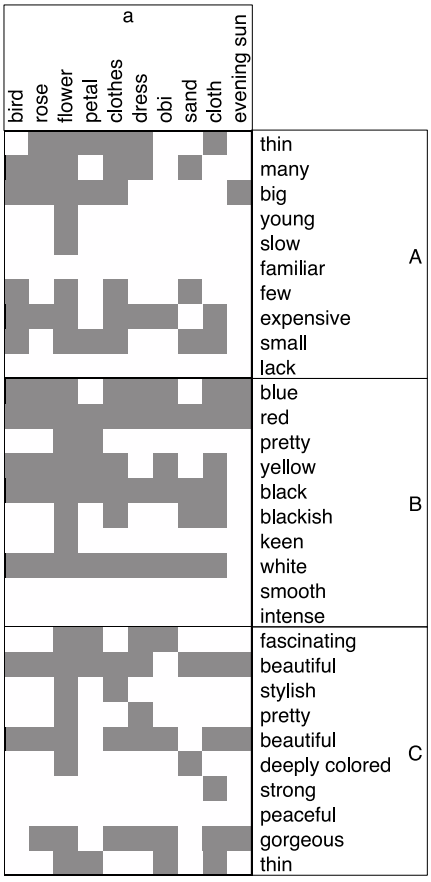

Fig. 8 Clusters discovered by the FIRM in Figure 5. Each cell covers SAM cluster 11. $S_1(a, A) = 0.81$, $S_1(a, B) = 0.98$ and $S_1(a, C) = 0.77$.

gives better clustering results than the IRM in terms of the high resolution compared to the SAM. We also found that relational models also find clusters that are not discovered by the SAM.

The IRM is initially proposed for relational clustering with the arbitrary number of types. Therefore, it is straight forward to apply the FIRM to feature-rich datasets like 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 failure in an ad hoc way. Exact inference with failure 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)

[Attias 00] Attias, H.: A variational Bayesian framework for graphical models, in *Advances in Neural Information Processing Systems*, Vol. 12 (2000)

[Blei 06] Blei, D. M. and Jordan, M. I.: Variational Inference for Dirichlet Process Mixtures, *Bayesian Analysis*, Vol. 1, No. 1, pp. 121–144 (2006)

[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)

[Ghahramani 00] Ghahramani, Z. and Beal, M. J.: Variational inference for Bayesian mixtures of factor analysers, in *Advances in Neural Information Processing Systems*, Vol. 12 (2000)

[Griffiths 05] Griffiths, T. L. and Ghahramani, Z.: Infinite Latent Feature Models and the Indian Buffet Process, Technical report, Gatsby Computational Neuroscience Unit, University College London (2005)

[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)

[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)

[Sato 04] Sato, T. and Kameya, Y.: A Dynamic Programming Approach to Parameter Learning of Generative Models with Failure, in *Proceedings of ICML Workshop on Statistical Relational Learning and its Connection to the Other Fields* (2004)

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

[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)