1 Methods

1.1 The Bayesian Probabilistic Tensor Factorization Model

	joy	gladden	sorrow	sadden	anger
joyfulness	1	1	-1		
gladden	1	1		-1	
sad	-1		1	1	

(a) T_1: Semantic similarity

	joy	gladden	sorrow	sadden	anger
joyfulness	.3	.1	$-.1$	.1	$-.2$
gladden	.2	1	.2	.7	$-.3$
sad	.6	0	.4	.5	$-.1$

(b) T_2: Distributional similarity

Table 1: Similarity tensor

We assume the similarity R_{ij}^k can be expressed as the inner-product of D-dimensional latent vectors:

$$p(R_{ij}^k|U_i, U_j, T_k) \sim \mathcal{N}(< U_i, U_j, T_k >, \alpha^{-1}),$$

$$< U_i, U_j, T_k > \equiv \sum_{d=1}^{D} U_i^{(d)} U_j^{(d)} T_k^{(d)},$$

Add Gaussian priors to the variables:

$$p(U_i|\mu_U, \Lambda_U) \sim \mathcal{N}(\mu_U, \Lambda_U^{-1})$$

$$p(T_i|\mu_T, \Lambda_T) \sim \mathcal{N}(\mu_T, \Lambda_T^{-1})$$

Model the hyper-parameters as conjugate priors:

$$p(\alpha) = \mathcal{W}(\alpha|\hat{W}_0, \nu_0),$$

$$p(\mu_U, \Lambda_U) = \mathcal{N}(\mu_U|\mu_0, (\beta_0\Lambda_U)^{-1})\mathcal{W}(\Lambda_U|W_0, \nu_0),$$
$$p(\mu_T, \Lambda_T) = \mathcal{N}(\mu_T|\mu_0, (\beta_0\Lambda_T)^{-1})\mathcal{W}(\Lambda_T|W_0, \nu_0),$$

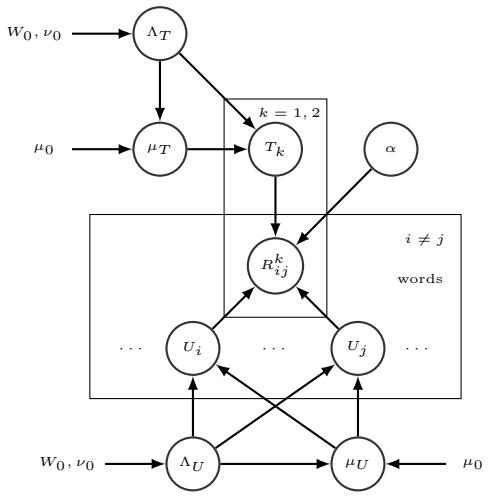

Figure 1: The graphical model for Bayesian PTF.

1.2 Gibbs Sampling

With conjugate priors, and assume $i \neq j$ doesn't exist for every word pair entries, the posterior distributions become to:

- $p(\alpha|\mathbf{R}, \mathbf{U}, \mathbf{T}) = \mathcal{W}(\hat{W_0}^*, \hat{\nu_0}^*)$,

  $\hat{\nu}_0^* = \hat{\nu}_0 + \sum_{k=1}^{2} \sum_{i,j=1}^{N} I_{ij}^k$,

  $(\hat{W}_0^*)^{-1} = \hat{W}_0^{-1} + \sum_{k=1}^{2} \sum_{i,j=1}^{N} I_{ij}^k (R_{ij}^k - < U_i, U_j, T_k >)^2$,

- $p(\mu_U, \Lambda_U|\mathbf{U}) = \mathcal{N}(\mu_U|\mu_0^*, (\beta_0^* \Lambda_U)^{-1}) \mathcal{W}(\Lambda_U|W_0^*, \nu_0^*)$,

  $\mu_0^* = \frac{\beta_0 \mu_0 + N\bar{U}}{\beta_0 + N}, \beta_0^* = \beta_0 + N, \nu_0^* = \nu_0 + N$,

  $(W_0^*)^{-1} = W_0^{-1} + N\bar{S} + \frac{\beta_0 N}{\beta_0 + N} (\mu_0 - \bar{U})(\mu_0 - \bar{U})^T$,

  $\bar{U} = \frac{1}{N} \sum_{i=1}^{N} U_i, \bar{S} = \frac{1}{N} \sum_{i=1}^{N} (U_i - \bar{U})(U_i - \bar{U})^T$.

- $p(\mu_T, \Lambda_T|\mathbf{T}) = \mathcal{N}(\mu_T|\mu_0^*, (\beta_0^* \Lambda_T)^{-1}) \mathcal{W}(\Lambda_T|W_0^*, \nu_0^*)$,
  which has the same form of $p(\mu_U, \Lambda_U|\mathbf{U})$.

- $p(U_i|\mathbf{R}, \mathbf{U}_{\neg i}, \mathbf{T}, \mu_U, \Lambda_U, \alpha) = \mathcal{N}(\mu_i^*, (\Lambda_i^*)^{-1})$,

  $\mu_i^* = (\Lambda_i^*)^{-1}(\Lambda_U \mu_U + \alpha \sum_{k=1}^{2} \sum_{j=1}^{N} I_{ij}^k R_{ij}^k Q_{jk})$,

  $\Lambda_i^* = \Lambda_U + \alpha \sum_{k=1}^{2} \sum_{j=1}^{N} I_{ij}^k Q_{jk} Q_{jk}^T$,

  where $Q_{jk} = U_j \cdot T_k, \cdot$ is the element-wise product.

- $p(T_i|\mathbf{R}, \mathbf{U}, \mathbf{T}_{\neg i}, \mu_T, \Lambda_T, \alpha) = \mathcal{N}(\mu_i^*, (\Lambda_i^*)^{-1})$,

  $\mu_k^* = (\Lambda_k^*)^{-1}(\Lambda_T \mu_T + \alpha \sum_{i,j=1}^{N} I_{ij}^k R_{ij}^k X_{ij})$,

  $\Lambda_k^* = \Lambda_T + \alpha \sum_{i,j=1}^{N} I_{ij}^k X_{ij} X_{ij}^T$,

  where $X_{ij} = U_i \cdot U_j, \cdot$ is the element-wise product.

1.3 Out-of-Vocabulary Embedding

When updating U_i, one only needs the existing similarities along with other sampled parameters. This implies that we can use the distributional similarities, which is easily to get, to sample the corresponding vector of an out-of-vocabulary word (like an weighted-averaging over vectors):

$$\mu_i^* = (\Lambda_i^*)^{-1}(\Lambda_U \mu_U + \alpha T_2 \cdot \sum_{j=1}^{N} I_{ij}^2 R_{ij}^2 U_j).$$

2 Experiments

Table 2: The test results of GRE questions.

	Dev. Set			Test Set		
	Prec	Rec	F_1	Prec	Rec	F_1
Encarta lookup	0.65	0.61	0.63	0.61	0.56	0.59
Yih, PILSA	0.86	0.81	0.84	0.81	0.74	0.77
Chang, MRLSA	0.87	0.82	0.84	0.82	0.74	0.78
Roget lookup	0.35	0.35	0.35	0.31	0.31	0.31
Roget w/o second slice	0.63	0.61	0.61	0.52	0.51	0.51
Roget subset[1] BPTF	0.83	0.79	0.81	0.80	0.77	0.79
Roget BPTF	0.81	0.79	0.80	0.75	0.74	0.74

[1] $\sim$5k words, $\sim$29k words in total.