Which Matters Most? Comparing the Impact of Concept and Document Relationships in Topic Models

Silvia Terragni♣, Debora Nozza♠, Elisabetta Fersini♣, Enza Messina♣
♣University of Milano-Bicocca, Milan,
♠Bocconi University, Milan
♣ s.terragni4@campus.unimib.it, ♠debora.nozza@unibocconi.it,
♣{elisabetta.fersini, enza.messina}@unimib.it

A Preprocessing

We lowercased the text, removed English stopwords and words occurring less than 10 times, and filtered out documents composed of less than 2 words. Details on the vocabulary composition are reported in Table 1.

B Computing Infrastructure

Experiments were run on three common computers using CPUs. Models can be run with basic infrastructure. Two computers have 8GB of RAM and the other has 16GB of RAM.

C Hyperparameters

Each experiment, with a given set of parameters, is repeated for 100 times and the performance measures are averaged by the number of the samples. The hyperparameters α and β are set equal to $50/K$ and 0.1 respectively (as reported in (**?**)) for all the considered models. All the compared models are trained for 1,500 Gibbs iterations. In our evaluation, we consider only must-constraint relations that can be generated by entities and words. To select the most appropriate value for ϵ_m, we studied the performance of the topic coherence of our models by varying the value of the parameter. The values for the models with the potential functions EE and EW are, respectively, 0.8 and 0.7 for the dataset Cora, and 0.6 and 0.6 for WebKB.

D Joint Distributions of the Proposed Models

For the sake of completeness, we report the joint distribution of the proposed models. Entity Constrained Latent Dirichlet Allocation (EC-LDA) defines the following joint probability distribution:

$$P(\mathbf{u}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \alpha, \beta, L) \propto \tag{1a}$$

$$\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(u_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d) \tag{1b}$$

$$\prod_{k}^{K} p(\Phi_k|\beta) \cdot \xi(\mathbf{z}, L) \tag{1c}$$

where

- D denotes the set of documents
- N_d is the length of document d
- K denotes the fixed number of topics
- $\mathbf{u}$ denotes the set of word and named entity tokens
- $\mathbf{z}$ represents the set of topic assignments
- $\boldsymbol{\theta}$ represents the document-topic distribution
- $\boldsymbol{\Phi}$ denotes the topic-word distribution
- α and β are the Dirichlet hyperparameters related to θ and Φ
- $\xi(\mathbf{z}, L) = \prod_{z \in \mathbf{z}} \exp f_l(z, u)$.

Similarly, the joint probability distribution of Entity Constrained Relational Topic Models is defined as follows:

$$P(\mathbf{u}, \mathbf{z}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\Phi} | \alpha, \beta, \eta, \nu, L) \propto \tag{2a}$$

$$\prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(u_{nd}|\Phi_{z_{nd}}) p(z_{nd}|\theta_d) \tag{2b}$$

$$\prod_{k}^{K} p(\Phi_k|\beta) \prod_{\substack{d,d' \in D \\ d' \neq d}} \psi_\sigma(y_{d,d'}|z_d, z_{d'}, \eta, \nu) \cdot \xi(\mathbf{z}, L) \tag{2c}$$

where ψ_σ is the link probability function defined as $\psi_\sigma(y = 1) = \sigma(\eta^T(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}) + \nu)$, σ is the sigmoid function and $\overline{\mathbf{z}}_d = \frac{1}{N_d} \sum_n z_{nd}$. The link function models each per-pair binary variable related to links as a logistic regression (with hidden

	Processed corpus			Unprocessed corpus
	# unique entities	**# unique words**	**# unique entities and words**	**# unique words**
Cora	384	2,675	3,059	3,012
WebKB	355	1,874	2,229	2,247

Table 1: Summary of the vocabularies for the benchmark datasets before and after the preprocessing phase.

covariates), parameterized by coefficients η and intercept ν.