For each entity e, our model samples its context word distribution ξ_e from a V-dimensional Dirichlet distribution with hyperparameter δ.

Finally, the full entity-topic model is shown in Figure 3 using the plate representation.

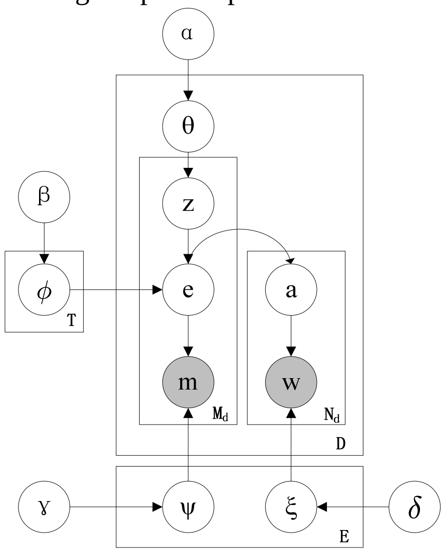

Figure 3. The plate representation of the entity-topic model

2.3 The Probability of a Corpus

Using the entity-topic model, the probability of generating a corpus $\mathbf{D}=\{d_1, d_2, ..., d_D\}$ given hyperparameters α, β, γ and δ can be expressed as:

$$\begin{aligned}P(\mathbf{D};\alpha,\beta,\gamma,\delta) &= \prod_d P(\mathbf{m_d},\mathbf{w_d};\alpha,\beta,\gamma,\delta) \\ &= \prod_d \sum_{\mathbf{e_d}} P(\mathbf{e_d}|\alpha,\beta)P(\mathbf{m_d}|\mathbf{e_d},\gamma)P(\mathbf{w_d}|\mathbf{e_d},\delta) \\ &= \int_\phi P(\phi|\beta)\int_\psi P(\psi|\gamma)\prod_d\sum_{\mathbf{e_d}} P(\mathbf{m_d}|\mathbf{e_d},\psi) \\ &\times \int_\xi P(\xi|\delta)\sum_{\mathbf{a_d}} P(\mathbf{a_d}|\mathbf{e_d})P(\mathbf{w_d}|\mathbf{a_d},\xi) \\ &\times \int_\theta P(\theta|\alpha)P(\mathbf{e_d}|\theta,\phi)d\theta d\xi d\psi d\phi \qquad (2.1)\end{aligned}$$

where $\mathbf{m_d}$ and $\mathbf{e_d}$ correspondingly the set of mentions and their entity assignments in document d, $\mathbf{w_d}$ and $\mathbf{a_d}$ correspondingly the set of words and their entity assignments in document d.

3 Inference using Gibbs Sampling

In this section, we describe how to resolve the entity linking problem using the entity-topic model. Overall, there were two inference tasks for EL:

1) ***The prediction task***. Given a document d, predicting its *entity assignments* ($\mathbf{e_d}$ for mentions and $\mathbf{a_d}$ for words) and *topic assignments* ($\mathbf{z_d}$). Notice that here the EL decisions are just the prediction of per-mention entity assignments ($\mathbf{e_d}$).

2) ***The knowledge discovery task***. Given a corpus $\mathbf{D}=\{d_1, d_2, ..., d_D\}$, estimating the global knowledge (including *the entity distribution of topics* ϕ, *the name distribution* ψ and *the context word distribution* ξ *of entities*) from data.

Unfortunately, due to the heaven correlation between *topics*, *entities*, *mentions* and *words* (the correlation is also demonstrated in Eq. (2.1), where the integral is intractable due to the coupling between θ, ϕ, ψ and ξ), the accurate inference of the above two tasks is intractable. For this reason, we propose an approximate inference algorithm – the *Gibbs sampling algorithm* for the entity-topic model by extending the well-known Gibbs sampling algorithm for LDA (Griffiths & Steyvers, 2004). In Gibbs sampling, we first construct the posterior distribution $P(\mathbf{z},\mathbf{e},\mathbf{a}|\mathbf{D})$, then this posterior distribution is used to: 1) estimate θ, ϕ, ψ and ξ; and 2) predict the entities and the topics of all documents in D. Specifically, we first derive the joint posterior distribution from Eq. (2.1) as:

$$P(\mathbf{z},\mathbf{e},\mathbf{a}|\mathbf{D}) \propto P(\mathbf{z})P(\mathbf{e}|\mathbf{z})P(\mathbf{m}|\mathbf{e})P(\mathbf{a}|\mathbf{e})P(\mathbf{w}|\mathbf{a})$$

where

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_t \Gamma(\alpha + C_{dt}^{DT})}{\Gamma(T\alpha + C_{d*}^{DT})} \qquad (3.1)$$

is the probability of the joint topic assignment $\mathbf{z}$ to all mentions $\mathbf{m}$ in corpus D, and

$$P(\mathbf{e}|\mathbf{z}) = \left(\frac{\Gamma(E\beta)}{\Gamma(\beta)^E}\right)^T \prod_{t=1}^{T} \frac{\prod_e \Gamma(\beta + C_{te}^{TE})}{\Gamma(E\beta + C_{t*}^{TE})} \qquad (3.2)$$

is the conditional probability of the joint entity assignments $\mathbf{e}$ to all mentions $\mathbf{m}$ in corpus D given all topic assignments $\mathbf{z}$, and

$$P(\mathbf{m}|\mathbf{e}) = \left(\frac{\Gamma(K\gamma)}{\Gamma(\gamma)^K}\right)^E \prod_{e=1}^{E} \frac{\prod_m \Gamma(\gamma + C_{em}^{EM})}{\Gamma(K\gamma + C_{e*}^{EM})} \qquad (3.3)$$

is the conditional probability of all mentions $\mathbf{m}$ given all per-mention entity assignments $\mathbf{e}$, and

$$P(\mathbf{a}|\mathbf{e}) = \prod_{d=1}^{D}\prod_{e\subset\mathbf{e_d}} \left(\frac{C_{de}^{DE}}{C_{d*}^{DE}}\right)^{C_{de}^{DA}} \qquad (3.4)$$

is the conditional probability of the joint entity assignments $\mathbf{a}$ to all words $\mathbf{w}$ in corpus D given all per-mention entity assignments $\mathbf{e}$, and