$$P(\mathbf{w}|\mathbf{a}) = (\frac{\Gamma(V\delta)}{\Gamma(\delta)^V})^E \prod_{e=1}^{E} \frac{\prod_w \Gamma(\delta + C_{ew}^{EW})}{\Gamma(V\delta + C_{e*}^{EW})} \quad (3.5)$$

is the conditional probability of all words $\mathbf{w}$ given all per-word entity assignments $\mathbf{a}$. In all above formulas, $\Gamma(.)$ is the Gamma function, C_{dt}^{DT} is the times topic t has been assigned for all mentions in document d, $C_{d*}^{DT} = \sum_t C_{dt}^{DT}$ is the topic number in document d, and C_{te}^{TE}, C_{em}^{EM}, C_{de}^{DE}, C_{de}^{DA}, C_{ew}^{EW} have similar explanation.

Based on the above joint probability, we construct a Markov chain that converges to the posterior distribution $P(\mathbf{z}, \mathbf{e}, \mathbf{a}|\mathbf{D})$ and then draw samples from this Markov chain for inference. For entity-topic model, each state in the Markov chain is an assignment (including *topic assignment to a mention, entity assignment to a mention* and *entity assignment to a word*). In Gibbs sampling, all assignments are sequentially sampled conditioned on all the current other assignments. So here we only need to derive the following three fully conditional assignment distributions:

1) $P(z_i = t|\mathbf{z}_{-i}, \mathbf{e}, \mathbf{a}, \mathbf{D})$: the topic assignment distribution to a mention given the current other topic assignments $\mathbf{z_{-i}}$, the current entity assignments $\mathbf{e}$ and $\mathbf{a}$;
2) $P(e_i = e|\mathbf{z}, \mathbf{e}_{-i}, \mathbf{a}, \mathbf{D})$: the entity assignment distribution to a mention given the current entity assignments of all other mentions $\mathbf{e_{-i}}$, the current topic assignments $\mathbf{z}$ and the current entity assignments of context words $\mathbf{a}$;
3) $P(a_i = e|\mathbf{z}, \mathbf{e}, \mathbf{a}_{-i}, \mathbf{D})$: the entity assignment distribution to a context word given the current entity assignments of all other context words $\mathbf{a_{-i}}$, the current topic assignments $\mathbf{z}$ and the current entity assignments $\mathbf{e}$ of mentions.

Using the Formula 3.1-3.5, we can derive the above three conditional distributions as (where m_i is contained in doc d):

$$P(z_i = t|\mathbf{z}_{-i}, \mathbf{e}, \mathbf{a}, \mathbf{D}) \propto \frac{C_{(-i)dt}^{DT} + \alpha}{C_{(-i)d*}^{DT} + T\alpha} \times \frac{C_{(-i)te}^{TE} + \beta}{C_{(-i)t*}^{TE} + E\beta}$$

where the topic assignment to a mention is determined by the probability this topic appearing in doc d (the 1st term) and the probability the referent entity appearing in this topic (the 2nd term);

$$P(e_i = e|\mathbf{z}, \mathbf{e}_{-i}, \mathbf{a}, \mathbf{D}) \propto \frac{C_{(-i)te}^{TE} + \beta}{C_{(-i)t*}^{TE} + E\beta} \times \frac{C_{(-i)em}^{EM} + \gamma}{C_{(-i)e*}^{EM} + K\gamma} \times (\frac{C_{(-i)de}^{DE} + 1}{C_{(-i)de}^{DE}})^{C_{de}^{DA}}$$

where the entity assignment to a mention is determined by the probability this entity extracted from the assigned topic (the 1st term), the probability this entity is referred by the name m (the 2nd term) and the contextual words describing this entity in doc d (the 3rd term);

$$P(a_i = e|\mathbf{z}, \mathbf{e}, \mathbf{a}_{-i}, \mathbf{D}) \propto \frac{C_{de}^{DE}}{C_{d*}^{DE}} \times \frac{C_{(-i)ew}^{EW} + \delta}{C_{(-i)e*}^{EW} + V\delta}$$

where the entity assignment to a word is determined by the number of times this entity has been assigned to mentions in doc d (the 1st term) and the probability the word appearing in the context of this entity (the 2nd term).

Finally, using the above three conditional distributions, we iteratively update all assignments of corpus $\mathbf{D}$ until coverage, then the global knowledge is estimated using the final assignments, and the final entity assignments are used as the referents of their corresponding mentions.

Inference on Unseen Documents. When unseen documents are given, we predict its entities and topics using the incremental Gibbs sampling algorithm described in (Kataria et al., 2011), i.e., we iteratively update the entity assignments and the topic assignments of an unseen document as the same as the above inference process, but with the previously learned global knowledge fixed.

Hyperparameter setting. One still problem here is the setting of the hyperparameters α, β, γ and δ. For α and β, this paper empirically set the value of them to $\alpha = 50/T$ and $\beta = 0.1$ as in Griffiths & Steyvers(2004). For γ, we notice that $K\gamma$ is the number of pseudo names added to each entity, when $\gamma = 0$ our model only mentions an entity using its previously used names. Observed that an entity typically has a fixed set of names, we set γ to a small value by setting $K\gamma = 1.0$. For δ, we notice that $V\delta$ is the number of pseudo words added to each entity, playing the role of smoothing its context word distribution. As there is typically a relatively loose correlation between an entity and its context words, we set δ to a relatively large value by fixing the total smoothing words added to each entity, a typical value is $V\delta$ = 2000.