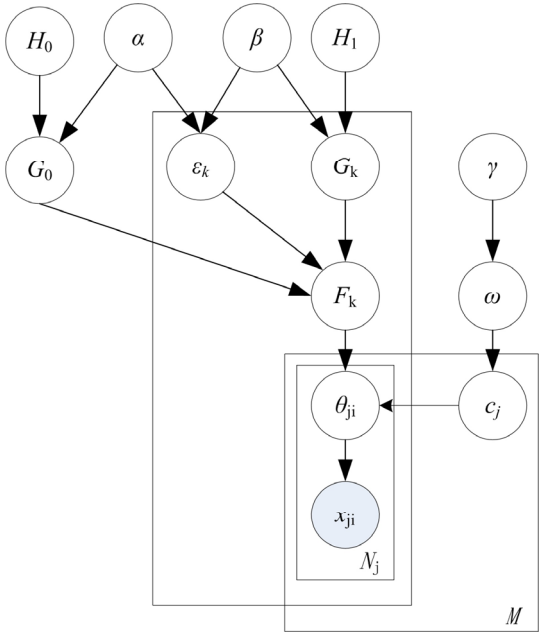
A Hybrid Nested/Hierarchical Dirichlet Process for Topic Modeling with Word Differentiation

Tengfei Ma, Issei Sato, Hiroshi Nakagawa

Introduction

- HDP does not work well in a corpus with several categories.
- Extending HDP by utilizing latent category information.
- Identifying discriminative words in topic modeling.

Figure 1. Graphical Model of hNHDP



Methods

1. Clustering Structure for data groups $F'_j \sim \sum_{k=1}^{\infty} \omega_k \delta_{F_k}$
2. Combination of global and local components in each cluster $F_k = \epsilon_k G_0 + (1 - \epsilon_k) G_k$.
3. Different base measures for local and global components: $G_0 \sim DP(\alpha, H_0)$ and $G_k \sim DP(\beta, H_1)$

Results

-Lowest perplexity on real datasets
-Ability to extract discriminative words
-Good clustering performance

Figure 2 Perplexity results