Integrating Topics and Syntax

Paper Presentation

CSC 665 – Advanced Topics in Probabilistic Graphical Models

Marium Yousuf

Outline

- Introduction
- Background
- Model
- Inference
- Results

Introduction

- Word dependencies
  - Short-range (syntax)
  - Long-range (context)
- Generative Model
  - Both kinds of dependencies
  - Syntactic classes and semantic topics
    - no representation beyond statistical dependency

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Background

- Syntactic Class
  - Short-range
  - Syntax
  - Span words within the limit of a sentence
  - Function Words
  - Handled by Hidden Markov Model
- Semantic Topic
  - Long-range
  - Context
  - Span words throughout the document (similar words)
  - Content Words
  - Handled by topic model

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Model

- Captures the interaction between the two components
  - Modularity
- Identify the role that words play in a document
  - Organizes words into syntactic and semantic classes
- Combination of two models
  - Each sensitive to one kind of dependency
- Mixture: either short- or long-range dependencies
- Product: both short- and long-range dependencies
- Asymmetry captured in a **composite** model

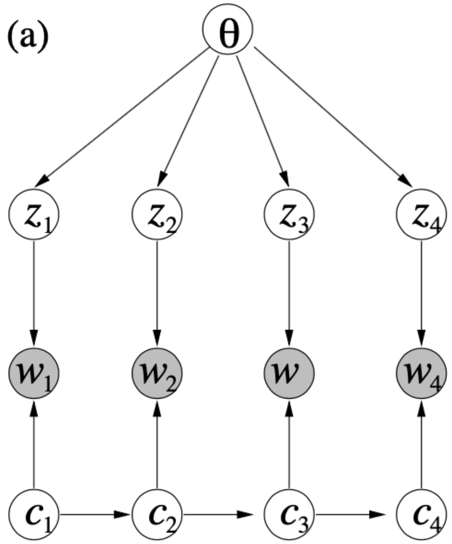
Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Composite Model

- The syntactic model
  - HMM
  - when to emit a content word, and
- The semantic model to choose
  - Topic model
  - which word to emit.
- Three sets of variables:
  - A sequence of words $w = \{w_1, \ldots, w_n\}$
  - A sequence of topic assignments $z = \{z_1, \ldots, z_n\}$
  - A sequence of classes $c = \{c_1, \ldots, c_n\}$
    - One class $c_i = 1$ is designated the "semantic" class.

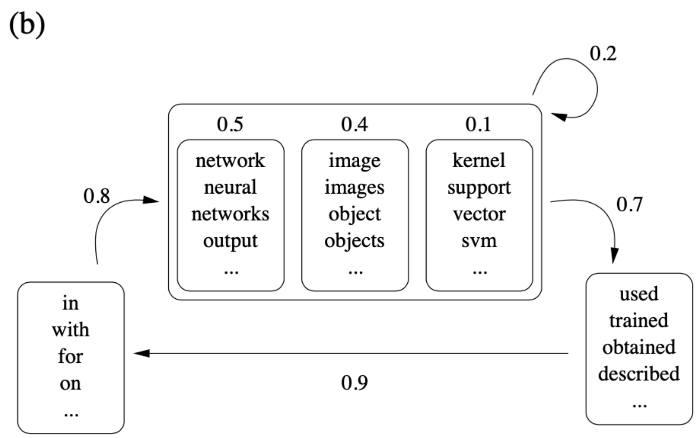
Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Composite Model



- a distribution over words $\phi^{(z)}$
- each class $c \neq 1$ is associated with a distribution over words $\phi^{(c)}$
- d has a distribution over topics $\theta^{(d)}$
- transitions between classes c_{i-1} and c_i follow a distribution $\pi^{(c_{i-1})}$
- Document generation:
  1) Sample $\theta^{(d)}$ from a Dirichlet (α) prior
  2) For each word w_i in document d
     - Draw z_i from $\theta^{(d)}$
     - Draw c_i from $\pi^{(c_{i-1})}$
     - If $c_i = 1$, then draw w_i from $\phi^{(z_i)}$ (***semantic***),
       else draw w_i from $\phi^{(c_i)}$ (***syntactic***)

Composite Model



- Phrase generation
- Three-class HMM
  - Multinomial distributions over words ($c_i \neq 1$)
  - Topic model containing three topics ($c_i = 1$)
- Probabilities in semantic class
  - to choose a topic when the HMM transitions to the semantic class
  - generate sentences with the same syntax but different content

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Inference

- θ: Dirichlet(α) distribution
- $\phi^{(z)}$: Dirichlet(β) distribution
- rows of the transition matrix: HMM Dirichlet(γ) distribution
- $\phi^{(c)}$: Dirichlet(δ) distribution
- All Dirichlet distributions are symmetric (uniform vector of reals)
- Gibbs Sampling
  - Draw topic and class assignment
  - Collapsed Gibbs Sampling (HMM)

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).

Results

- Syntactic classes and semantic topics
  - HMM allocates content words into semantic class
  - Assigned to topics
- Identifying function and content words
  - Factorization of words between the two components
- Marginal Probabilities
  - LDA outperforms on smaller corpora compared to HMM model
- **Part-of-speech tagging**
  - Focus on identifying the syntactic class of a word
- **Document Classification**
  - Grouping documents according to context

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Advances in neural information processing systems (pp. 537-544).