Grounding Topic Models with Knowledge Bases

Zhiting Hu\(^1\)*, Gang Luo\(^2\), Mrinmaya Sachan\(^1\), Eric Xing\(^1\), Zaiqing Nie\(^3\)

\(^1\)Carnegie Mellon University
\(^2\)Microsoft, California, US
\(^3\)Microsoft Research, Beijing, China

*This work was done when the first two authors were at Microsoft Research, Beijing
Topic Modeling

• Represents latent topics as probability distributions over words
Topic Modeling

- Represents latent topics as probability distributions over words.

LDA (latent Dirichlet process)
Topic Modeling

• Represents latent topics as probability distributions over words

LDA (latent Dirichlet process)

[Background]

Blei et al., 2003
Topic Modeling

- Represents latent topics as probability distributions over words
  - hard to interpret due to incoherence
  - lack of background context
  - no grounded semantics

Background

[Blei et al., 2003]
Topic Modeling

• Represents latent topics as probability distributions over words
  • hard to interpret due to incoherence
  • lack of background context
  • no grounded semantics

• Previous work combines external knowledge
  • improves coherence, but topics = word distributions
  • imposes one-to-one binding of topics to pre-defined knowledge base (KB) entities
  • Sacrifices flexibility

[Background]

[Blei et al., 2003]
This work

- A structured topic representation based on *entity taxonomy* from KBs
This work

• A structured topic representation based on *entity taxonomy* from KBs

Topic “Death of Whitney Houston”
This work

• A structured topic representation based on entity taxonomy from KBs
  • grounded semantics
  • improved coherenceness: captures entity correlations encoded in the taxonomy
This work

• A structured topic representation based on *entity taxonomy* from KBs
  • grounded semantics
  • improved coherenceness: captures entity correlations encoded in the taxonomy
• A probabilistic model to infer both hidden *topics* and *entities* from text corpora
Document Modeling

• Augments bag-of-word documents with *entity mentions*
  • mentions carry salient semantics of a document

• \{co-founder, wealthiest, man, ...\}
• \{Gates, Microsoft, ...\}
Document Modeling

• Generative process:
  • each mention <- an entity and a topic
  • each word <- an index indicating which mention to describe
Topic: Random Walk on Taxonomy

- Entity taxonomy
  - leaf: entity
  - internal nodes: category

- Each topic as a root-to-leaf random walk
  - a set of parent-to-child transition probabilities
  - -> entity/category weights
Topic: Random Walk on Taxonomy

• Entity taxonomy
  • leaf: entity
  • internal nodes: category

• Each topic as a root-to-leaf random walk
  • a set of parent-to-child transition probabilities
  • \( \rightarrow \) entity/category weights

• Path-sharing:
  • encourages clustering correlated entities into the same topic
Entity Modeling

• A distribution over mentions
  • captures relatedness between the entity and mentions
  • *Microsoft Inc.* – MS, Gates

• A distribution over words
  • characterizes the entity attributes
  • *Bill Gates* - wealthiest

![Diagram showing entity modeling with a tree structure and mentions like Gates, co-founder of Microsoft, was the wealthiest man in the world.](image)
Graphical Model Representation

Method
Graphical Model Representation

Latent Grounded Semantic Analysis (LGSA)

Gates, the co-founder of Microsoft, was the wealthiest man in the world.
Experiments

  • Entity Wikipedia pages
  • Entity category hierarchy

• Datasets
  • TMZ (tmz.com): celebrity gossip news
    • celebrity labels
    • #doc ~ 30K
  • New York Times news (LDC)
    • #doc ~ 330K

• Baselines
Topic Perplexity

Experiments

On the TMZ dataset

On the NYT dataset
Key Entity Identification

• Key entity of a document
  • E.g., the persons a news article is mainly about
• TMZ dataset: ground truth (celebrity label) available
• LGSA: $\theta'_d$ - distribution over entities for document $d$
Key Entity Identification

- Key entity of a document
  - E.g., the persons a news article is mainly about
- TMZ dataset: ground truth
- LGSA: $\theta_d'$ - distribution over
Example Topics: Sports

- Kobe Bryant Absolved in Church Assault Case
- San Diego Church: Kobe's Innocent!
- Kobe Bryant in the Gym with Manny Pacquiao
- LeBron James Just Jumped over a Guy!
- LeBron Alleged Mishugas at Jewish Basketball Game
Example Topics: Kardashian and Humphries’ Divorce

- Minnesota
- Kardashian family
- Marriage
- Practice of law
- Fraud
- Annulment
- Marriage
- Lawyer
- Divorce
- Lawsuit
- Kim Kardashian
- Kris Humphries
- Kris Has lawyered up for Divorce
- The Annulment Documents
- Kim: Kris' Parents Hated Me
- Kim: No Reconciliation

Experiments
Conclusion

• Traditional word-based topic representation lacks interpretability and grounded semantics
• A structured topic representation based on entity taxonomy from KBs
• A probabilistic model (LGSA) to infer latent grounded topics
• Improved performance on topic perplexity and key entity identification
Thanks..