Modeling Knowledge Incorporation into Topic Models and their Evaluation

Silvia Terragni
s.terragni4@campus.unimib.it
silviatti.github.io
@TerragniSilvia
Outline

- Introduction and state of the art of topic models
- Incorporating knowledge into topic models
  - relationships between documents and words
  - pre-trained contextualized representations
- Evaluation of topic models
  - framework for comparing topic models
  - hyperparameter optimization
Topic Modeling
What is Topic Modeling

Corpus of documents → TOPIC MODEL → Distribution of topics in each document

**TOPIC 1**
- Supervised learning
- classify
- prediction
- class

**TOPIC 2**
- Learning reinforce
- reward
- agent
- q-learning

**TOPIC 3**
- clustering learning
- model
- similarity
- centroid

**Topic indicators**
The human genome is the complete set of nucleic acid sequences for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria…

We can express a document as a **multinomial distribution over the topics**: a document talks about different topics in different proportions.

- A document that talks about diseases only.
- A document that talks about human evolution.
- A document that talks about evolution, diseases, and a little about humans.
Topic Models as probabilistic models

This is not just a unordered list of words. We can expressed it as a **multinomial distribution over the vocabulary**.

It’s a probability distribution! It sums to 1.
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Topics are expressed by a multinomial distribution over the vocabulary
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Dirichlet hyperparameter that controls how the sparsity of the words characterizing a topic
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

A document is expressed as a multinominal distribution
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

The hyperparameter that controls the sparsity of the topics in a document
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

A topic is assigned to each word
Latent Dirichlet Allocation

- Most known topic model: LDA [Blei+ 03]
- Fully unsupervised (the only observations are the words in documents)

Words are sampled from the word distribution given the topic assignment.
State-of-the-art Topic models

- Usually based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003]

- Increase the capacity of the model by extending LDA:
  - relaxing some assumptions of the model [Wallach et al., 2006]
  - incorporating external knowledge [Nguyen et al., 2015]
  - changing the representation of words [Das et al., 2015]
State-of-the-art Topic models

- Neural Topic Models:
  - usually based on Variational Autoencoders (VAEs) [Miao et al., 2016]
  - the encoder discovers the latent topic document representation
  - the top-words of a topic are obtained by the weight matrix that reconstructs the BOW

![Diagram of neural topic models](image)
Research Questions

**RQ1:** How can we incorporate knowledge into topic models?

**RQ2:** How can we ensure fairer comparisons between the models?
Incorporating Knowledge in Topic Models: Relationships
Relational Topic Models

Most topic models assume that documents and its constituents (i.e. words) are independent from each other.

Word-level
[Yang et al, 2015; Nguyen et al, 2015]

Document-level
[Chang et al., 2009; Yang et al., 2016]

RQ1: How can we incorporate knowledge into topic models?
RQ1: How can we incorporate knowledge into topic models?

(Document) Relational Topic Models

RQ1: How can we incorporate knowledge into topic models?

**Document Constrained Relational Topic Models**

**Document labels in the form of relationships:** Two documents that share the same label are more likely to share the same topics.

RQ1: How can we incorporate knowledge into topic models?

Document Constrained Relational Topic Models

Document labels in the form of relationships: Two documents that share the same label are more likely to share the same topics.

RQ1: How can we incorporate knowledge into topic models?

**Entity Constrained Relational Topic Models**

- **Relationships between documents** (RTM)
- **Relationships between words and entities**: two named-entities or words that are related are more likely to share the same topics

RQ1: How can we incorporate knowledge into topic models?

Entity Constrained Relational Topic Models

- **Relationships between documents (RTM)**
- **Relationships between words and entities:** two named-entities or words that are related are more likely to share the same topics

![Graph showing coherence comparison between LDA, LDA-EE, RTM, and RTM-EE](https://github.com/MIND-Lab/EC-RTM)

Entity Constrained Relational Topic Models

- **Relationships between documents** (RTM)
- **Relationships between words and entities**: two named-entities or words that are related are more likely to share the same topics

RQ1: How can we incorporate knowledge into topic models?

Incorporating Knowledge in Topic Models:
Pre-trained Representations
RQ1: How can we incorporate knowledge into topic models?

Why using pre-trained representations

- capture syntactic and semantic information of the sentence
- can be multilingual
- handle out-of-vocabulary (OOV) words
RQ1: How can we incorporate knowledge into topic models?

Contextualized Topic Models: Combined CTM

- concatenation of BOW and Sentence BERT
- improve the coherence of the topics
- effective on short texts
- RoBERTa outperforms BERT

Combined CTM

Open-source python library: https://github.com/MilaNLProc/contextualized-topic-models
We reached over 32k downloads and 440 github stars :)

What if we replace the BOW representation with pre-trained multilingual representations?

We can **zero-shot predict the topics** of a document in an **unseen language**

*Open-source python library: [https://github.com/MilaNLProc/contextualized-topic-models](https://github.com/MilaNLProc/contextualized-topic-models)*

We reached over 32k downloads and 440 github stars :)

RQ1: How can we incorporate knowledge into topic models?

**Contextualized Topic Models: Zero-shot CTM**

also multilingual

- **text representation**
- **hidden layers**
- **sampling**
- **topic document representation**
- **reconstructed BOW representation**

### Table: Topic Prediction

<table>
<thead>
<tr>
<th>Text</th>
<th>Lang</th>
<th>Topic Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blackmore’s Night is a British/American traditional folk....</td>
<td>EN</td>
<td>rock, band, bass, formed, ....</td>
</tr>
<tr>
<td>I Blackmore’s Night sono la band fondatrice del renaissance rock...</td>
<td>IT</td>
<td>rock, band, bass, formed, ....</td>
</tr>
<tr>
<td>On nomme fourmi de Langton un automate cellulaire...</td>
<td>FR</td>
<td>mathematics, theory, space, numbers, ...</td>
</tr>
<tr>
<td>Die Ameise ist eine Turingmaschine mit einem zweidimensionalen...</td>
<td>DE</td>
<td>mathematics, theory, space, numbers, ...</td>
</tr>
</tbody>
</table>

Evaluating Topic Models
Evaluating a Topic Model

- Evaluating an unsupervised model is not trivial
- Recall that a topic model has two main outputs:

  **TOPIC 1**
  Supervised learning
classify prediction class

  **TOPIC 2**
  Learning reinforce reward agent q-learning

  **TOPIC 3**
  clustering learning model similarity centroid

RQ2: How can we ensure fairer comparisons between the models?
RQ2: How can we ensure fairer comparisons between the models?

Evaluation of the top words

Main aspects of the top words of the topics:

1) how **coherent** are the topics?
2) how **diverse** are the topics?
RQ2: How can we ensure fairer comparisons between the models?

Evaluation of the top words

Main aspects of the top words of the topics:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

- **GOOD TOPICS**
  - Evolution
  - Evolutionary
  - Human
  - Organisms
  - Life
  - Dna
  - Genetic
  - Genes
  - Sequence

- **JUNK TOPIC**
  - Disease
  - Pizza
  - Music
  - Diseases
  - Sport
  - Bacterial

Some words are not related to others!
RQ2: How can we ensure fairer comparisons between the models?

Evaluation of the top words

Main aspects of the top words of the topics:

1) how **coherent** are the topics?
2) how **diverse** are the topics?

We’d like that topics express separate ideas or semantic areas.
Evaluation of the document-topic distribution

- intrinsic evaluation:
  - **perplexity**: what is the likelihood that the words of the test document $x$ have been generated by the trained topic model?

- extrinsic evaluation:
  - evaluate the **classification** performance
  - any other external task

RQ2: How can we ensure fairer comparisons between the models?
Why evaluating topic models is hard

- No benchmark datasets and non-standard pre-processing
- Stochasticity of the results
- Which topic model? Few releases in different programming languages, need to adapt data to each different implementation

A first solution: ToModAPI

Why evaluating topic models is hard

- Hyperparameters setting:
  - Comparing the models by fixing their hyperparameters is not fair
  - Finding the best hyperparameter configuration is time-consuming

RQ2: How can we ensure fairer comparisons between the models?
RQ2: How can we ensure fairer comparisons between the models?

Optimizing and Comparing Topic Models is Simple!

Pre-processing:
- Most common pre-processing tools
- Ready-to-use pre-processed datasets

Evaluation metrics
- Topic coherence
- Topic diversity
- Topic significance
- Document classification

Topic models:
- 4 classical topic models
- 4 neural topic models

Hyperparameter search
- Bayesian optimization for optimizing the hyperparameters

Open-source python library & local web dashboard: https://github.com/mind-lab/octis
We reached over 8k downloads and 170 github stars :)
RQ2: How can we ensure fairer comparisons between the models?

Bayesian Optimization

- **probabilistic surrogate model**: approximates the objective function
- **acquisition function**: select the next configuration using the mean and the confidence of the surrogate model
Bayesian Optimization for Topic Modeling

RQ2: How can we ensure fairer comparisons between the models?

Bayesian Optimization algorithm

Hyperparameter configuration

Evaluation metric

Repeat for M runs & average

Score

Repeat for N iterations

TOPIC MODEL

apple
banana
kiwi
fruit

wood
lake
water
nature

lion
dog

animal
cat

It's a black box!
RQ2: How can we ensure fairer comparisons between the models?

Optimizing the Hyperparameters

- We optimize the performance of relational topic models with respect to the classification metric F1-score
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs
RQ2: How can we ensure fairer comparisons between the models?

Optimizing the Hyperparameters

- We optimize the performance of relational topic models with respect to the classification metric F1-score
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs

The configuration identified by BO leads to a better performance with respect to its initial configuration.
Optimizing the Hyperparameters

- We optimize the performance of relational topic models with respect to the classification metric F1-score.
- We also evaluate other qualitative metrics to investigate different aspects of the RTMs.

The configuration identified by BO leads to a better performance with respect to its initial configuration.

Optimizing for classification purposes can be detrimental to different qualitative metrics.

RQ2: How can we ensure fairer comparisons between the models?
What’s next?

- If we optimize for a metric, what happens to the others?

- BO can be expensive:
  - Which hyperparameters are important to optimize?
  - Can we reduce the space of the hyperparameters?
  - Hyperparameter transfer

RQ2: How can we ensure fairer comparisons between the models?
Thank you :)
References