

Toward Joint Learning of Topic Models and Word Embeddings

Patrick Ding, Kristyn Pantoja



December 3, 2018

Overview

1 Topic Models

- Introduction
- Topic Models
- AEVB
- Autoencoded Variational Inference for Topic Models

2 Word Embeddings

- Introduction
- Word2Vec Skip-Gram
- Bayesian Skip-Gram

3 Word Embeddings and Topic Models

- Gaussian Soft-Max
- Experiments
- Work on Jointly Learning Word Embeddings for Topic Models
- Proposed Model

Section 1

1 Topic Models

- Introduction
- Topic Models
- AEVB
- Autoencoded Variational Inference for Topic Models

2 Word Embeddings

- Introduction
- Word2Vec Skip-Gram
- Bayesian Skip-Gram

3 Word Embeddings and Topic Models

- Gaussian Soft-Max
- Experiments
- Work on Jointly Learning Word Embeddings for Topic Models
- Proposed Model

Topic Models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

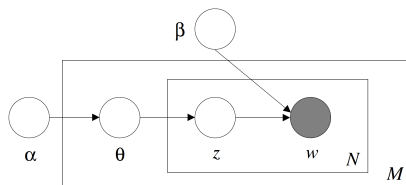
The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Figure: [Blei, Ng, and Jordan 2003]

Latent Dirichlet Allocation

[Blei, Ng, and Jordan 2003]

$$\begin{aligned}\theta_d &\sim \text{Dirichlet}(\alpha) \\ z_n | \theta_d &\sim \text{Categorical}(\theta_d) \\ w_n | z_n, \beta &\sim \text{Categorical}(\beta_{z_n})\end{aligned}$$



- ➊ Mean field VI: approximate $p(\theta, z | d, \alpha, \beta)$ using $q_\gamma(\theta)q_\phi(z)$

Auto-Encoding Variational Bayes

[Kingma and Welling, 2014]

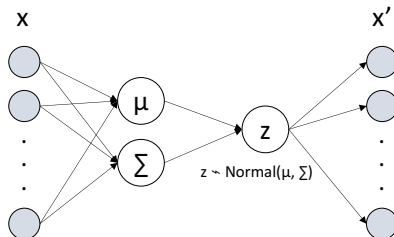
1 Inference network for amortized variational inference

Mean field approximation

$$q(z) = \prod_j q_{\lambda_j}(z_j)$$

Inference network

$$q_{\lambda}(z|x)$$



2 Reparameterization gradients

Autoencoded VI for Topic Models

[Srivastava and Sutton, 2017]

- 1 Sum out z and use Laplace approximation for Dirichlet prior on θ
- 2 Approximate $p(\theta|d, \alpha, \beta)$ using $q(\theta|d, \alpha, \beta)$

$$L(\Theta) = \sum_{i=1}^D \left[-D_{KL}(q(\theta|d, \alpha, \beta) || \hat{p}(\theta|\mu_1, \Sigma_1)) \right. \\ \left. + \mathbb{E}_{\epsilon \sim N(0,1)} \left[d_i^T \log(\sigma(\beta)\sigma(\mu_0 + \Sigma_0^{1/2}\epsilon)) \right] \right]$$

- 3 Faster training, easy to implement new models (ProdLDA)

Section 2

1 Topic Models

- Introduction
- Topic Models
- AEVB
- Autoencoded Variational Inference for Topic Models

2 Word Embeddings

- Introduction
- Word2Vec Skip-Gram
- Bayesian Skip-Gram

3 Word Embeddings and Topic Models

- Gaussian Soft-Max
- Experiments
- Work on Jointly Learning Word Embeddings for Topic Models
- Proposed Model

Word Embeddings

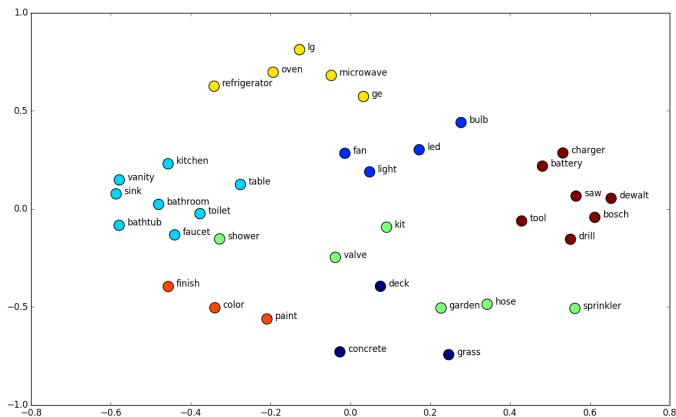
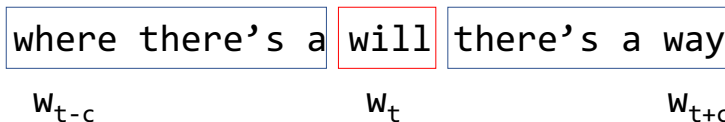


Figure: Lynn, Shane

<https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/>

Word2Vec Skip-Gram



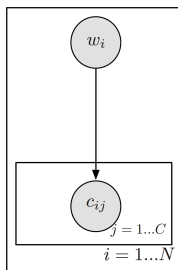
Word2Vec Skip-Gram Model [Mikolov et al, 2013] : Predict context given center word

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

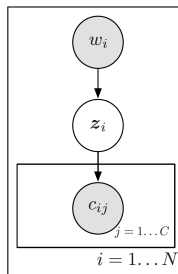
$$p(w_{t+j} | w_t) = \frac{\exp(v'_{w_{t+j}} v_{w_t})}{\sum_{w=1}^W \exp(v'_w v_{w_t})}$$

Bayesian Skip-Gram

[Brazinskas et al, 2018]



(a) Skip-gram



(b) Bayesian Skip-gram

- ① Token vectors drawn from type normal distributions
- ② polysemous words' vectors are drawn from Normals with higher variance
- ③ trained with inference networks

Section 3

1 Topic Models

- Introduction
- Topic Models
- AEVB
- Autoencoded Variational Inference for Topic Models

2 Word Embeddings

- Introduction
- Word2Vec Skip-Gram
- Bayesian Skip-Gram

3 Word Embeddings and Topic Models

- Gaussian Soft-Max
- Experiments
- Work on Jointly Learning Word Embeddings for Topic Models
- Proposed Model

Gaussian Soft-Max

In Gaussian Soft-Max topic model [Miao et al, 2017],

$$x_d \sim N(\mu_0, \Sigma_0)$$

$$\theta_d = \sigma(W_1^T x_d)$$

$$z_n | \theta_d \sim \text{Categorical}(\theta_d) \quad \forall n \in [1, N_d]$$

$$w_n | z_n, \beta \sim \text{Categorical}(\beta_{z_n}) \quad \forall n \in [1, N_d]$$

❶ μ_0, Σ_0 are learned using an inference network

❷ $\beta = \sigma(vt^T)$ incorporates word vectors

Experiment Setup

- ➊ Goal: To combine topic models with word embeddings to see if they improve the quality of the discovered topics
- ➋ 20 News Groups:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

- ➌ We compared the results of the Gaussian Soft-Max model (GSM) with different word embeddings, including:
 - ➊ word vectors pre-trained on 20News
 - ➋ word vectors pre-trained on wikipedia articles (a larger corpus)
 - ➌ un-initialized word vectors

And we compared these to the Neural Variational LDA (NVLDA).

Example Output

Top 10 words of topics from a Gaussian Soft-Max model with 50 topics, using word embeddings that were pre-trained on 20 News Groups dataset:

- ① god jesus people christian christians believe bible faith say church
- ② israel jews people turkish israeli armenian greek government turks world
- ③ gun control government states american state firearms new united guns
- ④ drive scsi hard disk card windows controller drives problem dos
- ⑤ game year team play win period games players university hockey
- ⑥ space nasa research information launch internet center program new national

Perplexities

1 Perplexity

Table: Comparing models' perplexities. Smaller is better.

Model	Mean Perplexity	Standard Error
GSM pre-trained 20NewsGroups	868	1.6
GSM pre-trained Wikipedia	2296132600	1257643168
GSM	893	1.8
NVLDA	911	1.6

Coherences

1 Coherence

Table: Comparing models' coherences. Larger is better.

Model	Mean Coherence	Standard Error
GSM pre-trained 20NewsGroups	117	3.8
GSM pre-trained Wikipedia	100	7.4
GSM	118	3.3
NVLDA	117	4.2

Work on Jointly Learning Word Embeddings for Topic Models

1 Ida2vec [Moody 2016]

Table: most similar words to the topic vector

"Space"	"Encryption"	"X Windows"	"Middle East"
astronomical	encryption	mydisplay	Armenian
Astronomy	wiretap	xlib	Lebanes
satellite	encrypt	window	Muslime
planetary	escrow	cursor	Turk
telescope	Clipper	pixmap	sy

2 STE [Shi et al 2017]

Table: highest probability bigrams

Topic 1	Topic 4	Topic 5	Topic 8
mcdonnell douglas	deir yassin	mucus membrane	saint aloysius
remote sensing	ottoman empire	amino acids	empty tomb
southern hemisphere	bedouin negev	kidney stones	respiratory papillomatosis
northern hemisphere	negev bedouin	anti fungal	zaurak kamsarakan
ozone layer	ermeniz mezalimi	candida albicans	biblical contradictions

Proposed Model

- 1 Extend Bayesian Skip-Gram model [Brazinskas et al, 2018]
- 2 Each context is assigned a topic

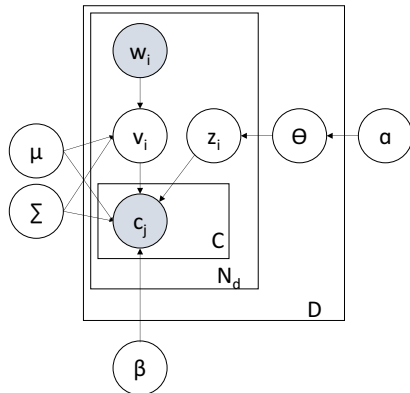
$$\theta_d \sim \text{Dirichlet}(\alpha)$$

$$z_i | \theta_d \sim \text{Categorical}(\theta_d)$$

$$v_i | w_i \sim \text{Normal}(\mu_w, \Sigma_w)$$

$$w_n | z_n, \beta \sim \text{Categorical}(\beta_{z_n})$$

$$c_j | v_i, z_i \sim \frac{\text{Normal}(v_i; \mu_{c_j i}, \Sigma_{c_j i}) \beta_{c_j i, z_i}}{\sum_{k=1}^V \text{Normal}(v_i; \mu_k, \Sigma_k) \beta_{k, z_i}}$$




Proposed Model

$$\begin{aligned} L(\Theta) &= \sum_{j=1}^c E_{q_{\phi}(v, \theta | c, w)} [\log p(c_{ji} | v_i, \theta)] \\ &\quad - KL(q(v | c, w) q(\theta | c, w) || p(v_i | w_i) p(\theta)) \\ p(c_{ji} | v_i, \theta) &= [\tilde{\beta} \cdot \theta]_{c_{ji}} \\ \tilde{\beta}_{c_{ji}, z_i} &= \frac{\text{Normal}(v_i; \mu_{c_{ji}}, \Sigma_{c_{ji}}) \beta_{c_{ji}, z_i}}{\sum_{k=1}^V \text{Normal}(v_i; \mu_k, \Sigma_k) \beta_{k, z_i}} \end{aligned}$$

References

Topic Modeling

 David Blei, Andrew Ng, and Michael Jordan (2003)

Latent Dirichlet Allocation

JMLR

 Durk Kingma and Max Welling (2014)

Auto-encoding Variational Bayes

ICLR 2014

 Akash Srivastava and Charles Sutton (2017)

Autoencoding Variational Inference For Topic Models

ICLR 2017

 Yishu Miao, Edward Grefenstette, Phil Blunsom (2017)

Discovering Discrete Latent Topics with Neural Variational Inference

ICML 2017

References

Word embedding



Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013)
Distributed Representations of Words and Phrases and their Compositionality
NIPS 2013



Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017)
Enriching Word Vectors with Subword Information
ACL 2017



Arthur Brazinskas, Serhii Havrylov, and Ivan Titov (2018)
Embedding Words as Distributions with a Bayesian Skip-gram Model
COLING 2018

References

Combining topic models and word embedding



Christopher E. Moody (2016)

Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec

arXiv



Bei Shi, Wai Lam, Shoaib Jameel, Steven Shockaert, and Kwun Ping Lai (2017)

Jointly Learning Word Embeddings and Latent Topics

SIGIR 2017

Thank You!