SentenceLDA

Discriminative and Robust Document Representation with Sentence Level Topic Model

Taehun Cha and Donghun Lee

Contents

Topic Modeling

- What are the "Topics" of the documents?
  - e.g. News - Sports, Politics, Business ...
- How can we discover it?
  - Supervised learning (classification)
  - Unsupervised way?

Topic Modeling

- In statistics and natural language processing, a topic model is a type of **statistical model for discovering the abstract "topics"** that occur in a collection of documents.
- Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently.
  - "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear approximately equally in both.

https://en.wikipedia.org/wiki/Topic_model

Topic Modeling - LDA

- Consider a data generating process:
  - Choose N ~ $Poisson(\xi)$
  - Choose θ ~ $Dirichlet(\alpha)$
  - For each of the N words w_n:
    - Choose a topic z_n ~ $Multinomial_T(\theta)$
    - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Topic Modeling - LDA

- For example, for topics (sports, business, politics):
  - Choose N = 100
  - Choose θ = (0.3, 0.5, 0.2)
  - For each of the N words w_n :
    - Choose a topic z_n = business
    - Choose a word w_n from $p(w_n|z_n, \beta)$,

Topic / Word	Soccer	Stock	Democracy	…
Sports	0.5	0.05	0.01	…
Business	0.1	0.4	0.2	…
Politics	0.05	0.1	0.5	…

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

SentenceLDA

- About Word-level Topic Models
  - Same words contain different meaning depending on contexts (Discriminative)
  - Different words contain same meaning depending on contexts (Robust)
  - List of words = Topic? (Interpretable)

Topic / Word	Soccer	Stock	Democracy	…
Sports	0.5	0.05	0.01	…
Business	0.1	0.4	0.2	…
Politics	0.05	0.1	0.5	…

SentenceLDA

- Latent Dirichlet Allocation (LDA)
  - Choose N ~ $Poisson(\xi)$
  - Choose θ ~ $Dirichlet(\alpha)$
  - For each of the N words w_n:
    - Choose a topic z_n ~ $Multinomial_T(\theta)$
    - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

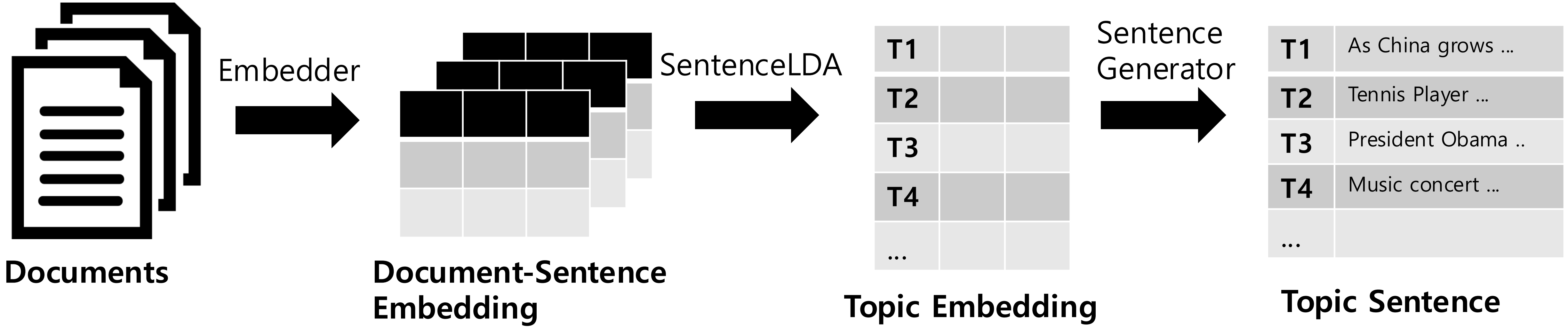
SentenceLDA

- GaussianLDA
  - Choose N ~ $Poisson(\xi)$
  - Choose θ ~ $Dirichlet(\alpha)$
  - For each of the N words w_n:
    - Choose a topic z_n ~ $Multinomial_T(\theta)$
    - Choose a word embedding w_n from $p(w_n|z_n, \beta)$, a Gaussian probability conditioned on the topic z_n

SentenceLDA

- SentenceLDA
  - Choose N ~ $Poisson(\xi)$
  - Choose θ ~ $Dirichlet(\alpha)$
  - For each of the N sentences s_n:
    - Choose a topic z_n ~ $Multinomial_T(\theta)$
    - Choose a sentence embedding s_n from $p(s_n|z_n, \beta)$, a Gaussian probability conditioned on the topic z_n

SentenceLDA



Experiment 1 – Discriminative

- Hypothesis
  - Does the sentence-level topic model improve discriminative (or classification) power of document representation?
- Dataset
  - 20News: 17.3K documents, 6 coarse, 20 fine grained classes
  - NYT: 11.6K documents, 5 coarse, 26 fine grained classes
- Baselines
  - LDA, GaussianLDA
  - Contextual TM: Word-level neural topic model utilizing contextual information
  - SenClu: Sentence-level topic model depending on similarity metric

Experiment 1 – Discriminative

Dataset	Topics	Class	LDA	GLDA	CTM	SenClu	SLDA (Ours)
20News	10	Computer (5)	44.99% (2.68)	24.03% (1.47)	36.34% (4.05)	**45.66% (3.35)**	42.25% (0.72)
		Ride (2)	64.05% (5.13)	51.92% (0.74)	75.40% (4.91)	73.13% (3.50)	**82.19% (0.84)**
		Sports (2)	76.61% (7.09)	63.29% (2.02)	84.29% (3.12)	77.33% (13.75)	**88.70% (1.53)**
		Science (4)	65.03% (2.32)	30.56% (1.75)	64.11% (3.16)	76.01% (2.01)	**78.38% (0.85)**
		Religion (3)	49.30% (3.05)	40.89% (0.08)	47.10% (2.60)	54.45% (3.55)	**58.64% (0.94)**
		Politics (3)	60.90% (3.21)	37.94% (1.31)	60.80% (3.23)	62.57% (4.14)	**68.28% (0.67)**
	20	Computer (5)	43.73% (1.98)	26.65% (1.39)	37.40% (3.76)	47.69% (3.60)	**52.20% (2.25)**
		Ride (2)	64.39% (2.56)	55.13% (0.80)	78.42% (3.95)	74.34% (2.50)	**80.66% (1.04)**
		Sports (2)	72.57% (5.19)	63.26% (2.34)	87.40% (2.28)	83.86% (2.61)	**88.43% (0.55)**
		Science (4)	66.67% (2.12)	34.75% (2.92)	69.06% (2.57)	76.35% (1.99)	**78.64% (0.76)**
		Religion (3)	46.89% (1.22)	40.85% (0.00)	51.74% (1.76)	57.02% (1.93)	**59.96% (1.15)**
		Politics (3)	63.06% (2.91)	39.57% (1.93)	60.91% (5.09)	68.14% (1.71)	**71.22% (0.96)**
		All (20)	37.99% (2.39)	8.72% (0.35)	34.55% (1.66)	40.91% (2.51)	**42.73% (3.51)**
NYT	10	Arts (4)	65.24% (5.41)	39.52% (0.00)	73.90% (4.65)	78.47% (6.97)	**93.14% (0.83)**
		Business (4)	72.63% (2.72)	46.97% (0.00)	74.24% (3.03)	62.83% (5.93)	**74.85% (2.22)**
		Politics (9)	60.20% (2.23)	41.79% (0.00)	61.09% (2.30)	60.40% (5.94)	**66.86% (0.67)**
		Science (2)	85.26% (6.14)	55.79% (2.58)	78.95% (7.44)	90.52% (2.11)	**91.58% (2.58)**
		Sports (7)	**91.91% (3.77)**	25.72% (0.00)	75.04% (3.75)	71.64% (6.33)	69.33% (3.99)
	20	Arts (4)	71.05% (4.22)	39.52% (0.00)	75.71% (6.10)	85.91% (2.82)	**95.81% (0.76)**
		Business (4)	72.42% (5.19)	46.97% (0.00)	77.48% (3.52)	75.96% (4.31)	**78.48% (1.45)**
		Politics (9)	65.67% (2.16)	41.79% (0.00)	63.48% (1.85)	69.55% (3.14)	**73.13% (1.75)**
		Science (2)	78.95% (11.04)	54.73% (2.58)	66.31% (5.37)	80.00% (2.10)	**89.47% (3.33)**
		Sports (7)	**96.59% (0.35)**	25.86% (0.20)	86.24% (1.12)	85.84% (5.88)	89.50% (2.03)
		All (26)	**82.98% (2.84)**	19.48% (0.25)	70.80% (1.78)	64.86% (6.24)	65.65% (1.12)

Experiment 1 – Discriminative

- Better performance of SenClu and SentenceLDA shows the sentence-level topic model improves the discriminative power
- GLDA returns almost same distribution for any document
- Superior performance of SentenceLDA is not just because of the sentence embedding

Dataset	Class	SBERT	SLDA
20News	Computer	34.96%	**52.20%**
	Ride	70.79%	**80.66%**
	Sports	80.60%	**88.43%**
	Science	52.07%	**78.64%**
	Religion	57.34%	**59.96%**
	Politics	63.68%	**71.22%**
NYT	Arts	56.19%	**95.81%**
	Business	65.15%	**78.48%**
	Politics	**75.62%**	73.13%
	Science	84.21%	**89.47%**
	Sports	**94.57%**	89.50%

Experiment 2 – Robust

- Hypothesis
  - Does sentence-level topic model improve robustness of document representation for paraphrasing?
- Paraphrasing Method
  - Lexical: Substitue words with synonyms
  - Syntactic: Parrot paraphraser (tends to change word order while maintaining words)
- Metric
  - $D_{sum}(P, Q) = \frac{1}{2}\Sigma_{i=1}^{K}|P_i - Q_i|$
  - Kendall's Tau: Compute rank correlation from –1 to 1

Experiment 2 – Robust

			Lexical				Syntactic			
Metric	Corpus	Topics	LDA	GLDA*	CTM	SLDA	LDA	GLDA*	CTM	SLDA
D_{sum}	**20News**	10	0.1868	0.0109	**0.1475**	0.1689	0.0765	0.0096	0.1319	**0.0743**
		20	0.2214	0.0210	**0.1636**	0.2223	**0.0909**	0.0156	0.1471	0.1020
	NYT	10	0.1814	0.0122	0.1645	**0.0808**	**0.0342**	0.0048	0.1222	0.0346
		20	0.1823	0.0133	0.1747	**0.1352**	**0.0434**	0.0090	0.1386	0.0594
τ	**20News**	10	0.7460	0.9360	0.5487	**0.8286**	0.8971	0.9500	0.5872	**0.9259**
		20	0.7319	0.9014	0.5765	**0.7748**	0.8875	0.9329	0.6096	**0.8973**
	NYT	10	0.7626	0.7790	0.5506	**0.9145**	0.9237	0.9548	0.5973	**0.9587**
		20	0.7647	0.8757	0.5149	**0.8624**	0.9194	0.9406	0.5454	**0.9326**

- GaussianLDA returns almost same distribution for any document
- LDA performs better for Syntactic than Lexical
- SentenceLDA is robust to both Lexical and Syntactic paraphrasing

Corpus-level Key Opinion Mining

- Dataset
  - DebateSum - "Impact Defense Core"
  - 762 debate documents with 12,957 sentences

- Model
  - 10 topics
  - Train GPT2-XL on DebateSum corpus (embedding to sentence)

Corpus-level Key Opinion Mining

Model	Extracted Topics
SLDA	1. With the war in Ukraine, Russia has not been able to count on the United States and Europe to keep Moscows feet firmly to the fire, much less to revive the stalled SinoRussian economic cooperation. 6. As China grows more powerful, it is increasingly at odds with Japan, which has a strong economic stake in the success of SinoAmerican relations and is understandably nervous about Beijings intentions in the South China Sea. 10. The worlds oceans have been shown to be less able to absorb and store carbon dioxide and other greenhouse gases, and the number of species known to be experiencing reduced populations has been rising since the 1950s.
LDA	1. nuclear, would, weapons, iran, war, states, could, us, one, attack 3. fish, ocean, one, species, global, change, said, also, data, warming 5. energy, oil, gas, us, said, prices, also, years, new, industry 6. china, us, military, russia, trade, said, japan, security, new, would 8. states, war, world, power, china, conflict, economic, political, united, global

Conclusion

- Semantic unit extension from word to sentence improves
  - Discriminative
  - Robust power of topic models

- SentenceLDA returns more interpretable topic sentences in sentence form