A Research on Text Vector Representations and Modelling based on Neural Networks

L.-Q. NIU

20/4/2016

Outline

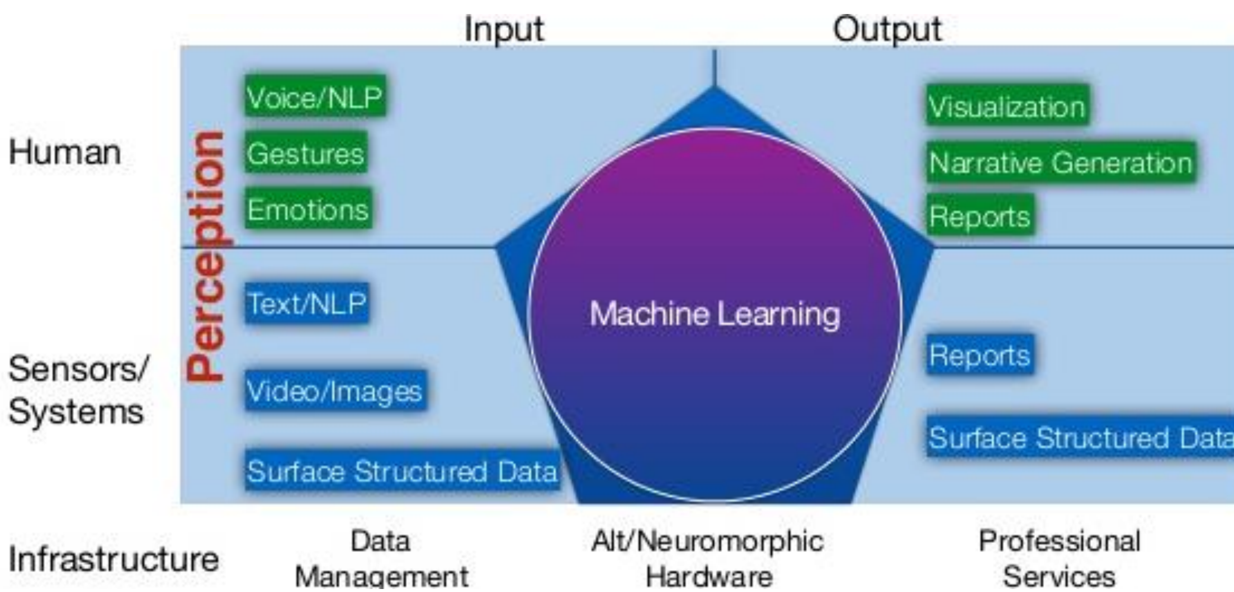
- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

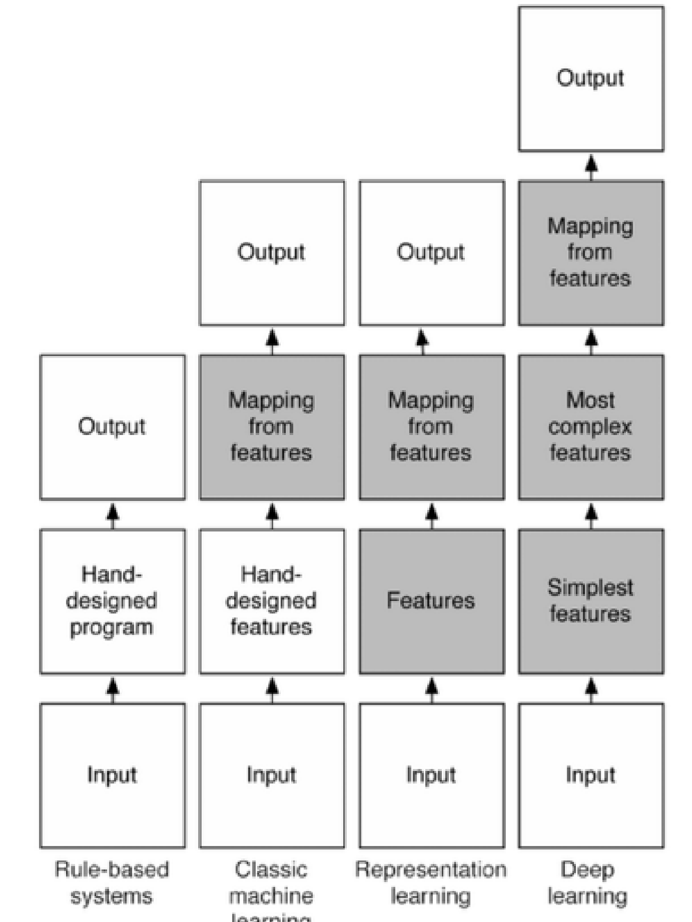
Background

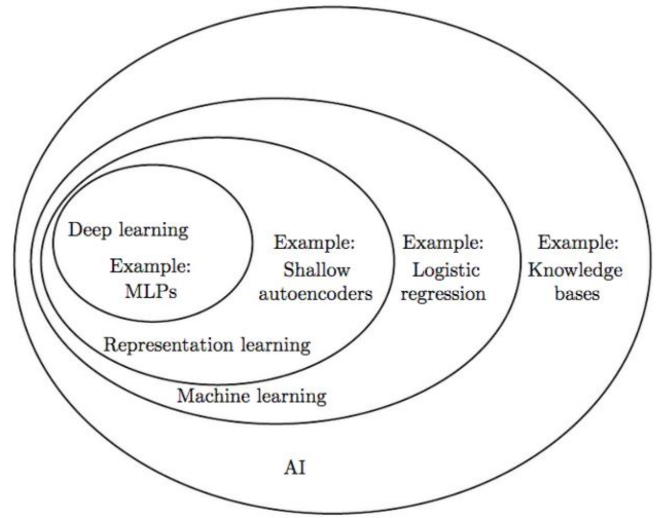
- Modern AI Systems
  - Perception: image/speech recognition, text understanding, etc.
  - Cognition: inference, reasoning, decision-making, etc.

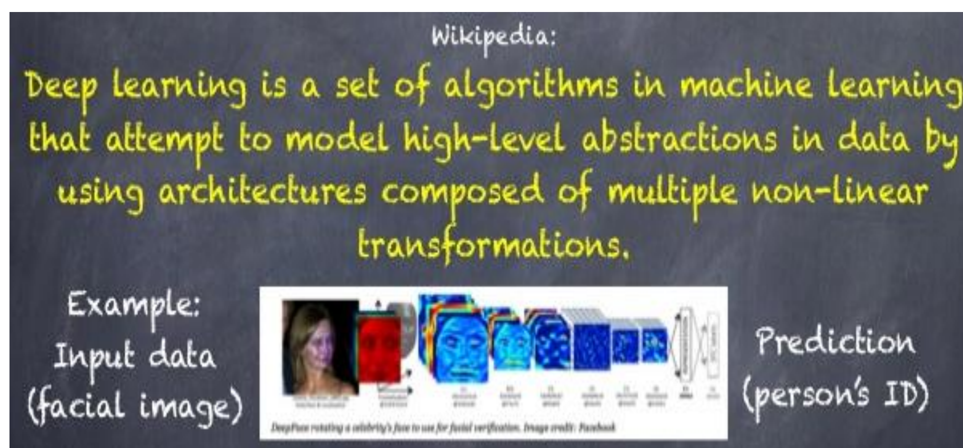


Background

- Machine Learning and Deep Learning







Background

- Deep Learning for Natural Language Processing (NLP)
  - The need for distributed representations
  - Distributed representations deal with the curse of dimensionality
  - Unsupervised feature and weight learning
  - Learning multiple levels of representation
  - Handling the recursivity of human language
- Deep Learning models have already achieved impressive results
  - LM, NER, POS-Tagging, Chunking, SA, etc.

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Traditional text representations

- The standard word representation

  The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: hotel, conference, walk

  In vector space terms, this is a vector with one 1 and a lot of zeroes

  [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

  Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

  We call this a "one-hot" representation. Its problem:

  motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
  hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

- Bag-of-Words (BOW)

Traditional text representations

- Distributional similarity based representations

  "You shall know a word by the company it keeps"

  (J. R. Firth 1957: 11)

  One of the most successful ideas of modern statistical NLP

  government debt problems turning into banking crises as has happened in

  saying that Europe needs unified banking regulation to replace the hodgepodge

  ↖ These words will represent *banking* ↗

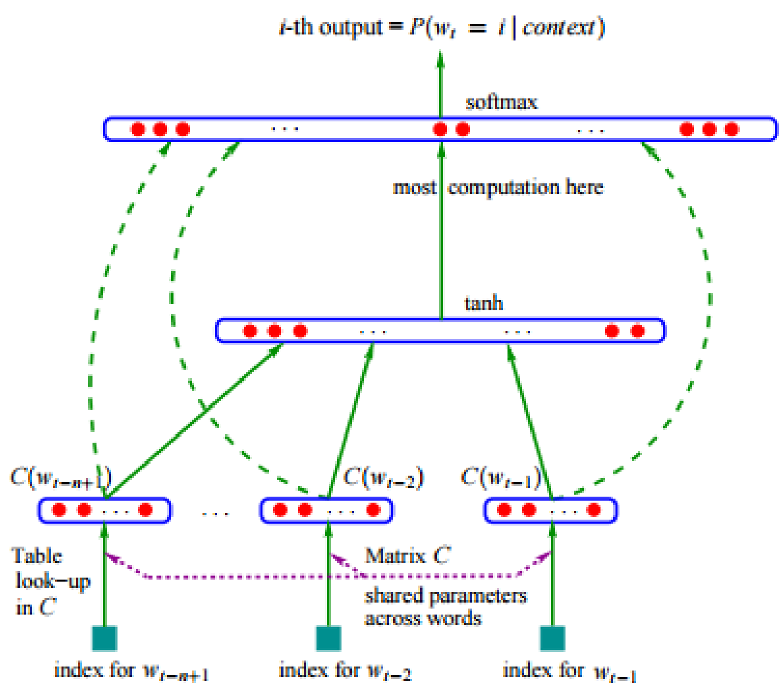
- Distributional representations
  - Latent Semantic Analysis (LSA), LSI, PLSA
  - Latent Dirichlet Allocation (LDA)
  - Hyperspace Analogue to Language (HAL)
- Clustering-based word representations

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Distributed representations

- Neural Probabilistic Language Models (NPLMs)
  - learns simultaneously (1) a distributed representation for each word along with (2) the probability function for word sequences, expressed in terms of these representations.



Distributed representations

- Neural word embeddings as a distributed representation
  - Word2Vec
    - CBOW
    - Skip-gram
  - Optimization
    - Hierarchical softmax
    - Negative sampling
    - SGD

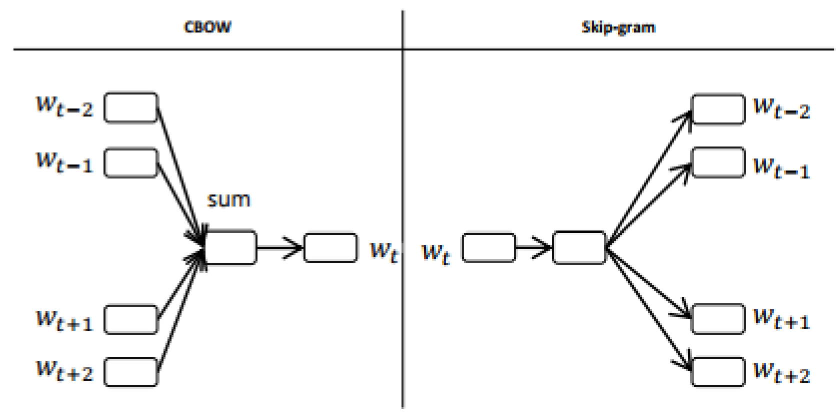


图 2-2: Word2Vec 结构图

$$L_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M} \log p(w_i|w_{cxt})$$

$$L_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M} \sum_{-k\leq c\leq k, c\neq 0} \log p(w_{i+c}|w_i)$$

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Motivations

- Perception tasks: image/speech recognition, text understanding, etc.
  - Deep learning: RBM, CNN, RNN, AE, etc.
- Cognitive tasks: inference, reasoning, decision-making, etc.
  - Bayesian graphical models: LDA, PMF, etc.
- Naturally, to integrate deep learnings and Bayesian models

I am convinced that the crux of the problem of learning is recognizing relationships and being able to use them.

Christopher Strachey in a letter to Alan Turing, 1954

Motivations

- Extending Word2Vec and LDA
  - Topic2Vec: Learning Distributed Representations of Topics, *IALP 2015*
  - A Unified Framework for Jointly Learning Distributed Representations of Word and Attributes, *ACML 2015*
- Integrating Word2Vec and LDA
  - Word Embedding Enhanced Topic Models

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Learning Distributed Representations of Topics

- Topic2Vec
  - CBOW
  - Skip-gram

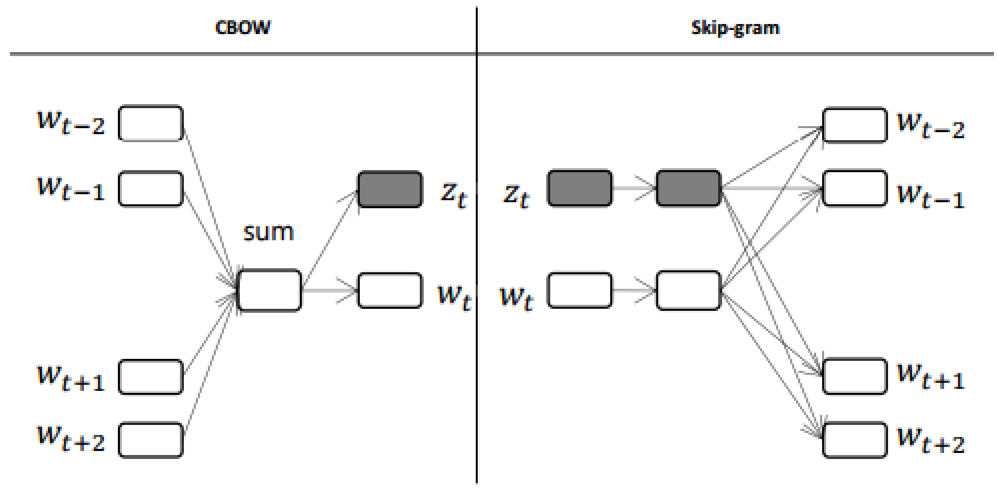


图 3-1: Topic2Vec 结构图

  - Optimization
    - Negative sampling
    - SGD

$$L_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M}(\log p(w_i|w_{cxt}) + \log p(z_i|w_{cxt}))$$

$$L_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\le c\le k, c\ne 0}(\log p(w_{i+c}|w_i) + \log p(w_{i+c}|z_i))$$

Learning Distributed Representations of Topics

- Experiment
  - Topic words

	Topic_6		Topic_19		Topic_27		Topic_47	
	word	prob.	word	prob.	word	prob.	word	prob.
LDA	food	0.027	drug	0.031	medical	0.033	dog	0.011
	restaurant	0.008	drugs	0.019	hospital	0.024	garden	0.009
	eat	0.008	cancer	0.019	care	0.019	tree	0.009
	more	0.005	study	0.011	patients	0.018	dogs	0.009
	chicken	0.005	patients	0.011	doctors	0.016	plants	0.008
	cooking	0.005	treatment	0.009	health	0.013	trees	0.008
	eating	0.005	fda	0.009	doctor	0.009	animal	0.007
	one	0.005	heart	0.008	patient	0.009	plant	0.007
	good	0.005	risk	0.008	surgery	0.008	animals	0.006
	foods	0.005	more	0.007	center	0.008	zoo	0.006
	word/topic	cos.	word/topic	cos.	word/topic	cos.	word/topic	cos.
Topic2Vec	cheeseburgers	0.564	topic_62	0.618	topic_19	0.519	dogwood	0.498
	meatless	0.535	aricept	0.531	topic_62	0.478	dogwoods	0.494
	smoothies	0.534	topic_27	0.519	neonatal	0.466	topic_33	0.485
	topic_95	0.533	memantine	0.514	topic_13	0.457	bark	0.484
	meatloaf	0.530	enbrel	0.512	anesthesiologists	0.445	fescue	0.483
	tastier	0.530	gabapentin	0.511	anesthesia	0.439	aphids	0.478
	topic_52	0.527	colorectal	0.509	reconstructive	0.437	mulched	0.478
	cheeseburger	0.525	prilosec	0.507	comatose	0.437	azaleas	0.477
	concoctions	0.522	placebos	0.507	hysterectomy	0.433	shrub	0.475
	vegetarians	0.515	intravenously	0.504	ventilator	0.432	camellias	0.472
	Topic_53		Topic_67		Topic_79		Topic_93	
	word	prob.	word	prob.	word	prob.	word	prob.
LDA	government	0.022	www	0.028	computer	0.016	russia	0.028
	africa	0.015	com	0.023	technology	0.010	russian	0.027
	people	0.015	hotel	0.018	phone	0.009	putin	0.017
	african	0.011	travel	0.015	software	0.009	soviet	0.013
	country	0.009	trip	0.011	digital	0.008	moscow	0.012
	international	0.008	night	0.010	apple	0.008	president	0.010
	darfur	0.007	per	0.009	use	0.007	country	0.007
	sudan	0.007	day	0.008	system	0.006	former	0.007
	south	0.007	tour	0.008	microsoft	0.006	state	0.007
	human	0.007	cruise	0.007	up	0.006	union	0.006
	word/topic	cos.	word/topic	cos.	word/topic	cos.	word/topic	cos.
Topic2Vec	mozambique	0.428	fairmont	0.569	wirelessly	0.584	topic_88	0.469
	uganda	0.423	motorcoach	0.553	handhelds	0.573	boris	0.435
	ghana	0.419	stateroom	0.547	desktops	0.572	leonid	0.411
	addis	0.417	uniworld	0.540	pda	0.566	dmitry	0.404
	darfur	0.412	maarten	0.533	smartphone	0.566	vladimir	0.397
	burundi	0.408	tourcrafters	0.529	megabyte	0.562	mikhail	0.397
	lanka	0.407	wyndham	0.528	macbook	0.556	dmitri	0.396
	congo	0.406	cunard	0.527	handheld	0.549	alexei	0.394
	ababa	0.403	safaris	0.522	treo	0.549	eduard	0.392
	darfurians	0.402	trafalgar	0.518	modems	0.548	kasparov	0.391

图 3-2: 对比 LDA 和 Topic2Vec 模型列举出给定主题所包含的主题词

Learning Distributed Representations of Topics

- Experiment: t-SNE 2D embedding

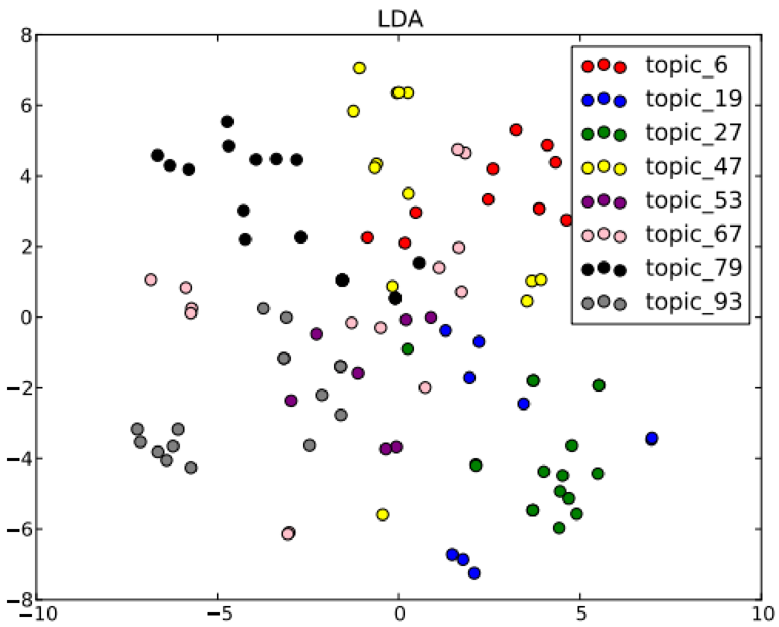


图 3-3: LDA 结果中每个主题所包含主题词基于 t-SNE 的在 2 维空间的映射

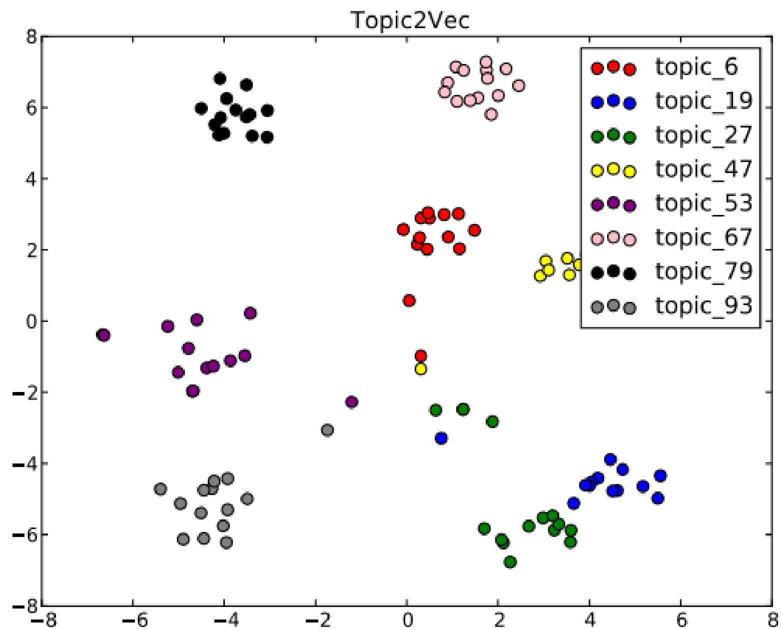


图 3-4: Topic2Vec 结果中每个主题所包含主题和词基于 t-SNE 的在 2 维空间的映射

Learning Distributed Representations of Topics

- Summary
  - Topic2Vec: learning distributed topic representations
  - LDA: representing topic as probability distribution over words
  - Distributed topic representations perform better than distributional topic representations

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

A Unified Learning Framework for Words and Attributes

- Embeddings beyond word level
  - Phrases and sentences
  - Entities and relations
  - Social and citation networks
- Words occur with attributes
  - POS-Tag, lemma
  - Phrase, sentence
  - Language
  - Sentiment
  - Name

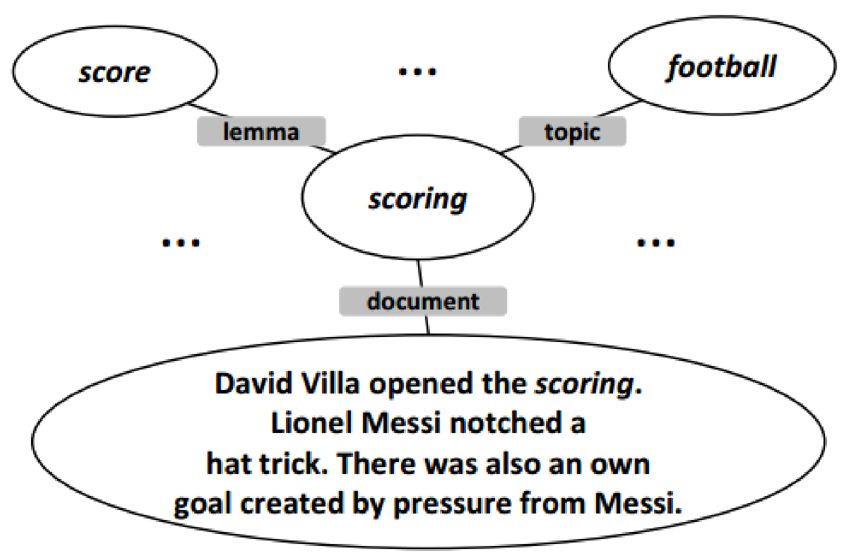


图 4-1: 词 "*scoring*" 及其节点属性图例

A Unified Learning Framework for Words and Attributes

- The need for a unified framework for jointly learning distributed representations of word and attributes

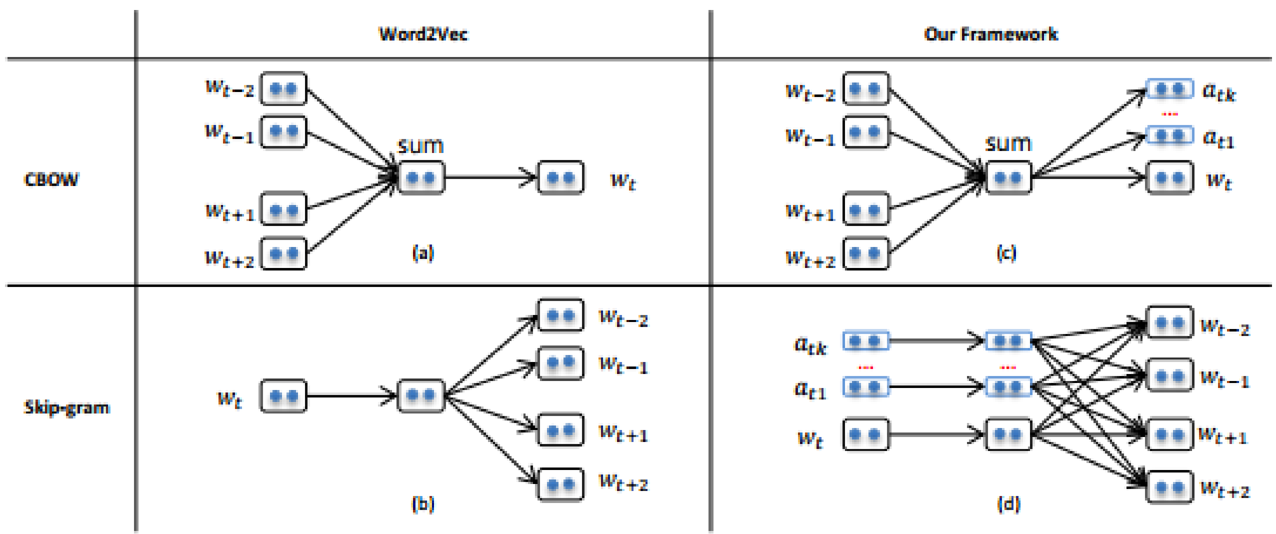


图 4-2: Word2Vec 和统一学习框架中的 CBOW 和 Skip-gram 模型结构对比图

A Unified Learning Framework for Words and Attributes

- Models

Models	Word and Attributes	Learning Targets
Word2Vec	word	word representations
TW	word:topic	topic representations and improved word representations
DW	word:document	document representations
LW	word:lemma	improved word representations
TLW	word:topic:lemma	improved word representations

Table 1: Pairs of word and attributes and learning targets used in Word2Vec Mikolov et al. (2013) and our models (TW, DW, LW and TLW).

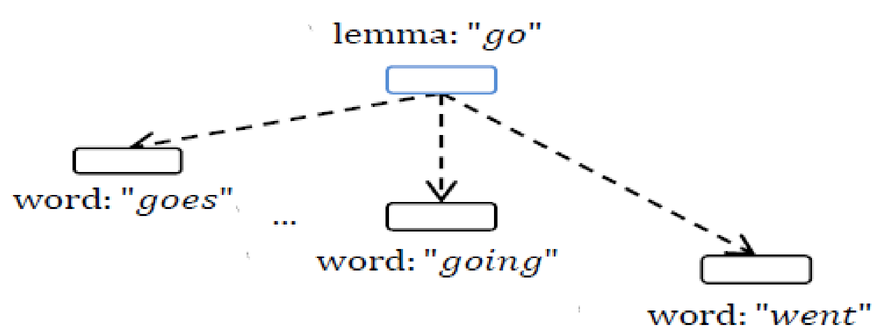


Figure 3: An example of lemma and variational words in morphology.

A Unified Learning Framework for Words and Attributes

- TW: Learning Topic Representations

$$\mathcal{L}_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M}(\log p(w_i|w_{cxt}) + \log p(z_i|w_{cxt})),$$

$$\mathcal{L}_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\leq c\leq k, c\neq 0}(\log p(w_{i+c}|w_i) + \log p(w_{i+c}|z_i)).$$

- DW: Learning Document Representations

$$\mathcal{L}_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M}(\log p(w_i|w_{cxt}) + \log p(d_i|w_{cxt})),$$

$$\mathcal{L}_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\leq c\leq k, c\neq 0}(\log p(w_{i+c}|w_i) + \log p(w_{i+c}|d_i)).$$

A Unified Learning Framework for Words and Attributes

- Improving Word Representations
  - LW

$$\mathcal{L}_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M}(\log p(w_i|w_{cxt}) + \log p(l_i|w_{cxt})),$$

$$\mathcal{L}_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\leq c\leq k, c\neq 0}(\log p(w_{i+c}|w_i) + \log p(w_{i+c}|l_i)).$$

  - TLW

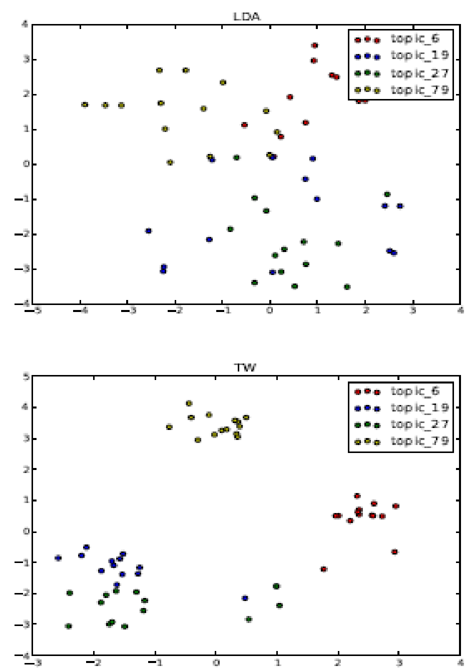
$$\mathcal{L}_{CBOW}(D) = \frac{1}{M}\sum_{i=1}^{M}(\log p(w_i|w_{cxt}) + \log p(z_i|w_{cxt}) + \log p(l_i|w_{cxt})),$$

$$\mathcal{L}_{Skip-gram}(D) = \frac{1}{M}\sum_{i=1}^{M}\sum_{-k\leq c\leq k, c\neq 0}(\log p(w_{i+c}|w_i) + \log p(w_{i+c}|z_i) + \log p(w_{i+c}|l_i)).$$

A Unified Learning Framework for Words and Attributes

- Evaluation for Topic Representations

	Topic_6		Topic_19		Topic_27		Topic_79	
LDA	word	prob.	word	prob.	word	prob.	word	prob.
	food	0.027	drug	0.031	medical	0.033	computer	0.016
	restaurant	0.008	drugs	0.019	hospital	0.024	technology	0.010
	eat	0.008	cancer	0.019	care	0.019	phone	0.009
	more	0.005	study	0.011	patients	0.018	software	0.009
	chicken	0.005	patients	0.011	doctors	0.016	digital	0.008
	cooking	0.005	treatment	0.009	health	0.013	apple	0.008
	eating	0.005	fda	0.009	doctor	0.009	use	0.007
	one	0.005	heart	0.008	patient	0.009	system	0.006
	good	0.005	risk	0.008	surgery	0.008	microsoft	0.006
	foods	0.005	more	0.007	center	0.008	up	0.006
	dinner	0.004	use	0.007	treatment	0.007	music	0.006
	make	0.004	blood	0.007	hospitals	0.007	video	0.006
	fresh	0.004	women	0.006	heart	0.006	one	0.006
	chef	0.004	disease	0.006	dr	0.006	more	0.005
	made	0.004	percent	0.005	one	0.005	computers	0.005
TW	word/topic	cos.	word/topic	cos.	word/topic	cos.	word/topic	cos.
	cheeseburgers	0.564	topic_62	0.618	topic_19	0.519	wirelessly	0.584
	meatless	0.535	aricept	0.531	topic_62	0.478	handhelds	0.573
	smoothies	0.534	topic_27	0.519	neonatal	0.466	desktops	0.572
	topic_95	0.533	memantine	0.514	topic_13	0.457	pda	0.566
	meatloaf	0.530	enbrel	0.512	anesthesiologists	0.445	smartphone	0.566
	tastier	0.530	gabapentin	0.511	anesthesia	0.439	megabyte	0.562
	topic_52	0.527	colorectal	0.509	reconstructive	0.437	macbook	0.556
	cheeseburger	0.525	prilosec	0.507	comatose	0.437	handheld	0.549
	concoctions	0.522	placebos	0.507	hysterectomy	0.433	treo	0.549
	vegetarians	0.515	intravenously	0.504	ventilator	0.432	modems	0.548
	twinkies	0.514	adderall	0.502	checkup	0.429	camcorders	0.547
	veggie	0.513	inhibitor	0.502	pacemaker	0.428	toshiba	0.545
	panera	0.513	opioid	0.501	aneurysms	0.423	peripherals	0.545
	pepperoni	0.507	oncologists	0.501	respirator	0.423	android	0.544
	condiments	0.504	precancerous	0.501	caesarean	0.422	centrino	0.543



(*a*) Nearest words and topics

(*b*) t-SNE 2D embedding

Figure 4: (a): Nearest words and topics for each topic. Words are listed with corresponding probabilities in LDA while words and topics are listed with calculated cosine similarity in TW. (b): t-SNE 2D embedding of the nearest word representation for each topic in LDA (above) and TW (below).

A Unified Learning Framework for Words and Attributes

- Evaluation for Document Representations
  - Text Classification

Models		Dim	Accuracy	Precision	Recall	F1-Measure
LDA		80	72.2	70.8	70.7	70.0
PV-DM		400	72.4	72.1	71.5	71.5
PV-DBOW		400	75.4	74.9	74.3	74.3
DW	CBOW	300	74.4	73.9	73.5	73.4
		400	**75.8**	**75.4**	**74.9**	**74.8**
	Skip-gram	300	72.1	71.5	71.2	71.1
		400	72.9	72.4	72.1	72.2

Table 2: The performance of DW compared to other approaches on 20NewsGroup. The results of other methods are reported in Liu et al. (2015). Bold scores are the best overall related models.

A Unified Learning Framework for Words and Attributes

- Evaluation for Improved Word Representations
  - Word analogy

Models (dim=300)		Dataset	Google			MSR	Time
			semantic	syntactic	total	syntactic	hours
CBOW	W2V	DS-100k	19.08	33.73	27.69	32.36	0.1
	TW	DS-100k	20.42	31.42	26.88	31.47	0.2
	LW	DS-100k	28.64	25.71	26.92	29.35	0.2
	TLW	DS-100k	28.15	27.32	27.67	30.21	0.2
Skip-gram	W2V	DS-100k	27.56	35.63	32.31	29.85	1.1
	TW	DS-100k	31.26	35.13	33.53	29.03	1.2
	LW	DS-100k	33.94	**37.13**(+1.50)	36.16	**35.42**(+5.57)	1.2
	TLW	DS-100k	**36.04**(+8.48)	36.60	**36.37**(+4.06)	34.65	1.3
Glove:iter=5		DS-100k	43.64	40.83	41.99	39.47	1.1
CBOW	W2V	DS-500k	30.57	50.57	41.74	44.97	2.1
	TW	DS-500k	28.12	49.60	40.12	43.93	2.2
	LW	DS-500k	41.80	46.11	44.21	42.43	2.2
	TLW	DS-500k	41.76	47.63	45.04	44.44	2.2
Skip-gram	W2V	DS-500k	41.77	50.63	46.89	43.38	6.8
	TW	DS-500k	41.46	49.46	45.93	41.39	7.4
	LW	DS-500k	**45.72**(+3.95)	**50.86**(+0.23)	**48.59**(+1.7)	**46.10**(+2.72)	7.2
	TLW	DS-500k	44.85	50.58	48.05	45.62	7.7
Glove:iter=5		DS-500k	51.32	49.12	50.09	46.36	6.3
Glove:iter=15		DS-500k	51.88	53.41	52.74	48.32	17.2

Table 3: Accuracy (%) in word analogy tasks, higher values are better. We compare our models (TW, LW and TLW) with baseline model W2V (Word2Vec) and state-of-the-art Glove. Bold scores are the best of our models for each dataset. Time is roughly estimated on a single machine with 8GB RAM.

A Unified Learning Framework for Words and Attributes

- Evaluation for Improved Word Representations
  - Word similarity

Model (dim=300)		**Corpus**	$\rho \times 100$
Glove:iter=5		DS-100k	51.9
CBOW	Word2Vec	DS-100k	55.6
	TW	DS-100k	62.6
	LW	DS-100k	63.9
	TLW	DS-100k	65.0
Skip-gram	Word2Vec	DS-100k	61.5
	TW	DS-100k	63.7
	LW	DS-100k	**65.4**
	TLW	DS-100k	63.5
Glove:iter=5		DS-500k	50.8
Glove:iter=15		DS-500k	50.9
CBOW	Word2Vec	DS-500k	63.7
	TW	DS-500k	62.2
	LW	DS-500k	65.9
	TLW	DS-500k	**67.5**
Skip-gram	Word2Vec	DS-500k	65.8
	TW	DS-500k	63.7
	LW	DS-500k	64.6
	TLW	DS-500k	63.9

Table 4: Comparing Spearman rank correlation coefficient of our models (TW, LW and TLW) with Word2Vec and Glove on WordSim-353. Bold scores are the best overall for each dataset.

A Unified Learning Framework for Words and Attributes

- Summary
  - We propose a unified framework for learning distributed representations of word and attributes.
  - Our models not only learn topic and document representations which achieve distinct and competitive results in corresponding tasks, but also improve original word representations significantly.
  - Our proposed framework is flexible and scalable.

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Embedding Enhanced Topic Models

- Word embeddings
  - Unsupervised learning
  - Large-scale datasets
  - Syntactic and semantic information
- Latent topic models
  - LDA models structure of words, topics, and documents
  - Gibbs sampling

Embedding Enhanced Topic Models

- Integrating word2vec and LDA
  - Word embedding clustering prior LDA
    - Using external large-scale dataset
  - Context-aware LDA
    - Adding a latent variable
  - Word embedding enhanced LDA
    - Using word embedding during inference

Embedding Enhanced Topic Models

- Word embedding clustering prior LDA (wecpLDA)
  - Using external large-scale dataset

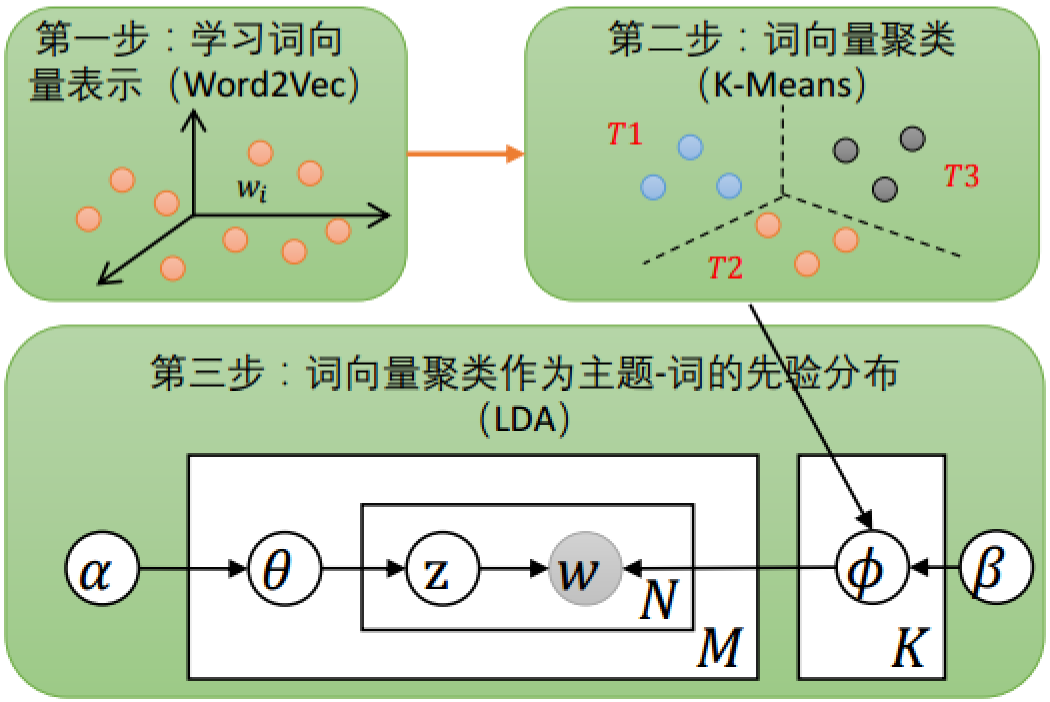


图 5-2: 词向量聚类先验潜在狄利克雷分布

Embedding Enhanced Topic Models

- Evaluation of Topic Coherence (noise data)

主题ID	LDA 主题词（top 30）	wecpLDA 主题词（top 30）
topic_0	the, of, to, and, in, that, is, are, for, or, have, it, they, be, as, their, not, can, at, more, but, with, people, by, an, than, you, on, who, said	you, as, it's, like, new, can, its, up, all, them, such, also, even, way, much, him, too, little, where, world, when, good, few, something, own, day, best, being, life, than
topic_1	the, to, of, that, in, and, said, was, he, on, for, is, by, from, not, his, court, case, with, had, who, an, has, have, law, this, at, as, Simpson, be	of, that, said, they, were, there, than, people, do, say, made, make, get, year, did, take, under, among, called, told, even, come, go ,including, work, going, three, use, know, put
topic_2	the, to, of, and, said, in, that, us, on, officials, for, have, it, was, by, at, be, were, an, from, has, they, would, been, military, with, not, but, is	percent, its, new, million, year, which, will, billion, company, more, market, up, money, companies, such, that, business, pay, industry, program, federal, workers, cost, stock, service, fund, government, funds, rates
topic_3	the, to, of, in, and, that, for, is, percent, said, it, will, on, its, at, as, has, are, with, by, be, new, year, have, million, but, more, which, company, market	the, to, and, in, for, is, on, with, it, as, at, have, by, but, from, has, are, an, was, this, their, would, had, who, one, will, about, been, more, we
topic_4	the, and, of, in, to, for, on, is, from, at, by, are, with, new, as, its, will, an, which, two, more, through, city, air, most, national, including, where, or, be	in, was, and, were, officials, police, city, which, where, up, army, them, air, international, here, un, two, day, national, world, war, who, south, near, miles, area, town, into, building, official
topic_5	the, and, to, of, is, in, that, it, her, she, with, on, as, you, for, but, this, it's, be, who, has, says, an, about, have, like, what, so, not, all	his, and, her, she, was, who, with, not, be, says, on, my, about, family, show, me, which, new, night, man, film, when, movie, life, TV, woman, wife, he's, first
topic_6	the, of, he, was, and, his, to, had, said, who, were, at, they, as, that, for, their, with, him, after, on, when, but, been, one, people, from, it, an	he, be, not, his, Clinton, house, that, president, or, white, also, by, federal, officials, which, congress, case, administration, law, campaign, republication, state, senate, court, bill, him, committee, dole, republications, public
topic_7	the, and, to, of, in, with, for, or, is, it, from, on, are, you, as, food, can, but, this, be, into, about, water, one, when, at, until, that, if, add	us, his, united, government, its, states, political, war, military, president, new, peace, foreign, party, north, or, minister, country, Russian, troops, china, Israel, leaders, power, group, Bosnian, against, economic, nations, Israeli
topic_8	the, to, of, and, in, that, for, on, Clinton, said, would, is, house, by, as, be, he, it, with, but, has, president, have, not, will, his, congress, who, republican, this	or, are, children, women, health, which, study, research, medical, test, group, also, university, found, percent, nuclear, Angeles, school, parents, may, blood, drug, care, report, safety, system, problem, public, problems, they
topic_9	the, to, of, in, and, that, is, for, with, by, has, as, on, have, be, but, it, government, will, from, are, united, us, its, said, an, states, this, war	he, be, or, not, you, is, your, it, so, says, just, my, into, each, get, which, up, water, this, there, don't, minutes, food, from, then, make, until, video, place, hot

图 5-3: 迭代次数为 100 时列举 wecpLDA 和 LDA 的主题词

Embedding Enhanced Topic Models

- Comparing wecpLDA with LDA
  - wecpLDA uses external knowledge

表 5-1: LDA 和 wecpLDA 主题模型评估结果对比

主题模型评估	LDA	wecpLDA
初始化方法	狄利克雷先验随机初始化	词向量 K-Means 聚类先验初始化
收敛速度	慢	快
收敛结果	差	好
主题一致性	差	好
主题多样性	差	好
处理稀疏和噪音数据	差	好

Embedding Enhanced Topic Models

- Context-aware LDA (caLDA)
  - Adding a latent variable *c*

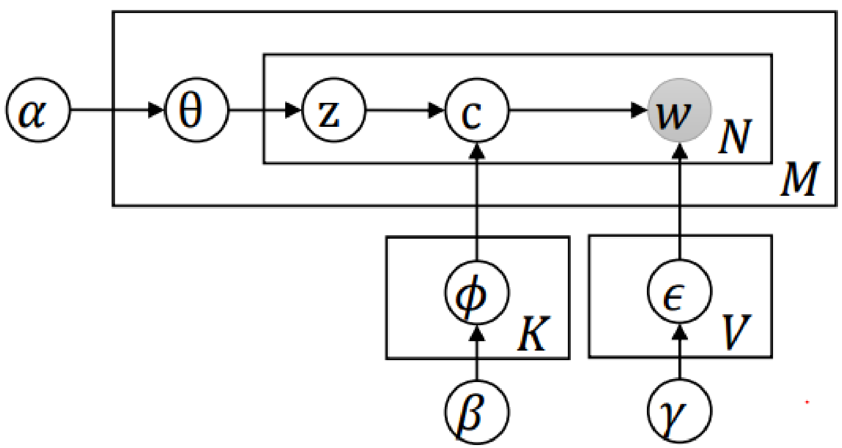


图 5-6: 上下文感知 LDA 的图模型表示

1. 对每一个文档 $For\ d = 1, ..., M : \theta_d \sim Dir(\alpha)$
2. 对每一个主题 $For\ k = 1, ..., K : \phi_k \sim Dir(\beta)$
3. 对每一个上下文词 $For\ v = 1, ..., V : \epsilon_v \sim Dir(\gamma)$
4. 对文档中出现的每一个词 $For\ n = 1, ..., N :$
   - 当前主题 $z_n \sim Mult(\theta_d)$
   - 当前上下文词 $c_n \sim Mult(\phi_{z_n})$
   - 当前词 $w_n \sim Mult(\epsilon_{c_n})$

Embedding Enhanced Topic Models

- Context-aware LDA (caLDA)
  - Gibbs sampling

$$P(c_i = v|\mathbf{c_{-i}}, \mathbf{w}, \mathbf{z}) \propto \frac{n_{-i,v}^{(w_i)} + \gamma}{n_{-i,v}^{(.)} + V\gamma} \cdot \frac{n_{-i,v}^{(z_i)} + \beta}{n_{-i,.}^{(z_i)} + V\beta}$$

$$P(z_i = k|\mathbf{z_{-i}}, \mathbf{w}, \mathbf{c}) \propto \frac{n_{-i,k}^{(c_i)} + \beta}{n_{-i,k}^{(.)} + V\beta} \cdot \frac{n_{-i,k}^{(d_i)} + \alpha}{n_{-i,.}^{(d_i)} + K\alpha}$$

  - Inference

$$\hat{\epsilon}_c^{(w)} = \frac{n_c^{(w)} + \gamma}{n_c^{(.)} + V\gamma} \qquad \hat{\phi}_j^{(c)} = \frac{n_j^{(c)} + \beta}{n_j^{(.)} + V\beta} \qquad \hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_.^{(d)} + K\alpha}$$

Embedding Enhanced Topic Models

- Word embedding enhanced LDA (weeLDA)
  - Using word embedding during inference

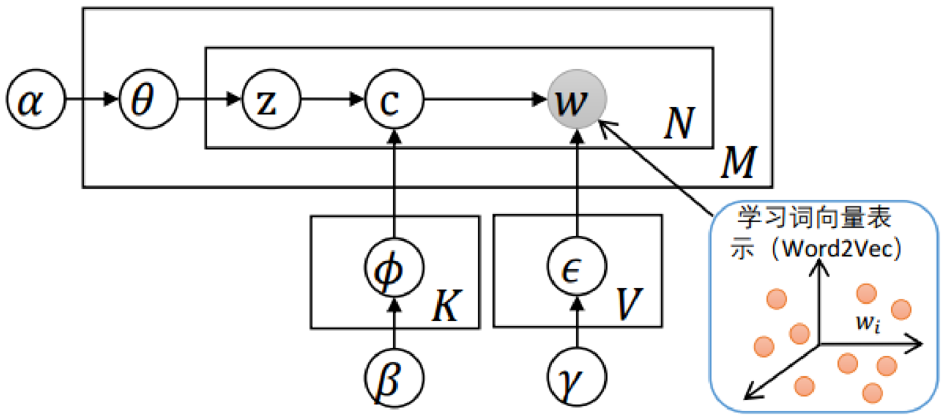


图 5-7: 词向量加强 LDA 的图模型表示

1. 对每一个文档 $For\ d = 1, ..., M : \theta_d \sim Dir(\alpha)$
2. 对每一个主题 $For\ k = 1, ..., K : \phi_k \sim Dir(\beta)$
3. 对每一个上下文词 $For\ v = 1, ..., V : \epsilon_v \sim Dir(\gamma)$
4. 对每一个上下文词 $For\ v = 1, ..., V : \epsilon'_v \sim Dis()$
5. 对文档中出现的每一个词 $For\ n = 1, ..., N :$
   - 当前主题 $z_n \sim Mult(\theta_d)$
   - 当前上下文词 $c_n \sim Mult(\phi_{z_n})$
   - 当前词 $w_n \sim Mult(\epsilon_{c_n}) \cdot Dis((\epsilon'_{c_n})$

Embedding Enhanced Topic Models

- Word embedding enhanced LDA (weeLDA)
  - Gibbs sampling

$$P(c_i = v|\mathbf{c_{-i}}, \mathbf{w}, \mathbf{z}) \propto \frac{n_{-i,v}^{(w_i)} + \gamma}{n_{-i,v}^{(\cdot)} + V\gamma} \cdot \frac{\exp(\mathbf{v} \cdot \mathbf{w_i})}{\sum_{c \in C} \exp(\mathbf{c} \cdot \mathbf{w_i})} \cdot \frac{n_{-i,v}^{(z_i)} + \beta}{n_{-i,\cdot}^{(z_i)} + V\beta}$$

$$P(z_i = k|\mathbf{z_{-i}}, \mathbf{w}, \mathbf{c}) \propto \frac{n_{-i,k}^{(c_i)} + \beta}{n_{-i,k}^{(\cdot)} + V\beta} \cdot \frac{n_{-i,k}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}$$

  - Inference

$$\hat{\epsilon}_c^{(w)} = \frac{n_c^{(w)} + \gamma}{n_c^{(\cdot)} + V\gamma} \qquad \hat{\phi}_j^{(c)} = \frac{n_j^{(c)} + \beta}{n_j^{(\cdot)} + V\beta} \qquad \hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + K\alpha}$$

Embedding Enhanced Topic Models

- Summary
  - Word embedding clustering prior LDA
    - wecpLDA performed better than LDA
  - Context-aware LDA
    - Implementation and experiments
  - Word embedding enhanced LDA (weeLDA)
    - Implementation and experiments
- Explore more
  - Bayesian deep learning
  - Denoising auto-encoders

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Conclusions

- Focus on text representations and modeling in NLP
  - Importance of representations
  - Related methods
  - Our methods
    - Learning Distributed Representations of Topics
    - A Unified Learning Framework for Words and Attributes
    - Embedding Enhanced Topic Models
- Future work
  - Exploring integration of deep learning and Bayesian models for AI systems

Outline

- Background
- Related Work
  - Traditional text representations
  - Distributed representations
- My Work
  - Motivations
  - Learning Distributed Representations of Topics
  - A Unified Learning Framework for Words and Attributes
  - Embedding Enhanced Topic Models
- Conclusions
- Reference

Reference

- **LeCun, Y., Bengio, Y. and Hinton, G. E. (2015) Deep Learning, Nature, Vol. 521, pp 436-444. http://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf**
- **Artificial Intelligence A Modern Approach http://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf**
- **Deep Learning for NLP http://nlp.stanford.edu/courses/NAACL2013/NAACL2013-Socher-Manning-DeepLearning.pdf**
- **Word representations: A simple and general method for semi-supervised learning http://www.aclweb.org/anthology/P10-1040**
- **Towards Bayesian Deep Learning: A Survey http://arxiv.org/pdf/1604.01662v2.pdf**
- ***An Introduction to MCMC for Machine Learning*** **http://www.cs.ubc.ca/~arnaud/andrieu_defreitas_doucet_jordan_intromontecarlomachinelearning.pdf**
- ***Information Theory, Inference, and Learning Algorithms http://www.inference.phy.cam.ac.uk/itprnn/book.pdf***
- **Yittoo http://139.129.37.204/niuwp/index.php/category/resources/**
- **Machine learning a probabilistic perspective**
- **A New Look at the System, Algorithm and Theory Foundations of Distributed Machine Learning http://petuum.github.io/papers/SysAlgTheoryKDD2015.pdf**
- **Pattern Recognition and Machine Learning**

Thank you!