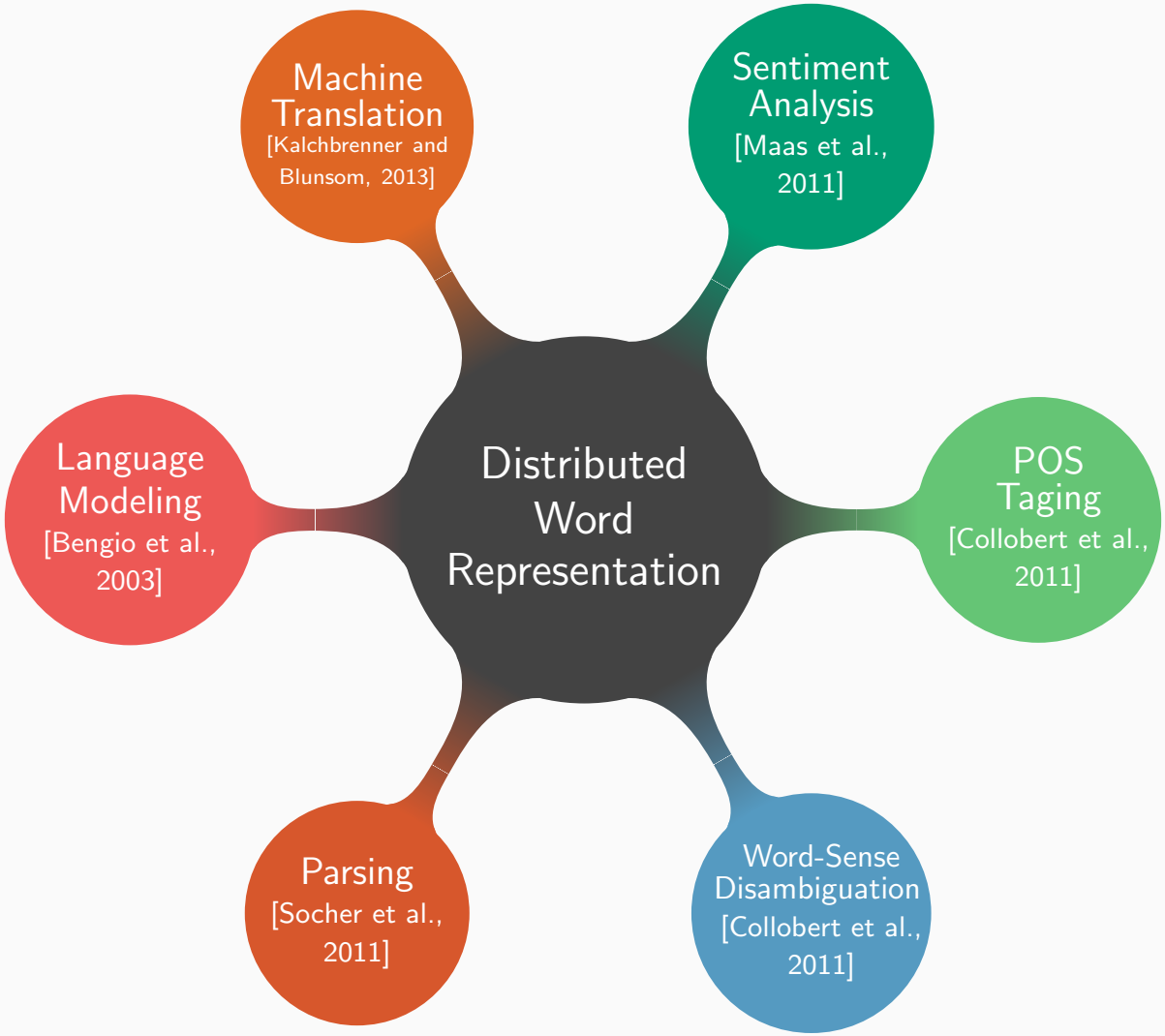
Sparse Word Embeddings Using ℓ_1 Regularized Online Learning

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng

July 14, 2016

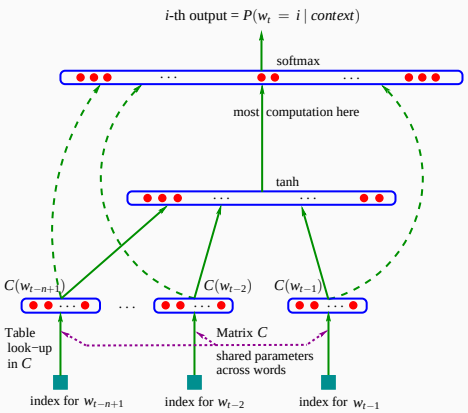
ofey.sunfei@gmail.com, {guojiafeng, lanyanyan,junxu, cxq}@ict.ac.cn
CAS Key Lab of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences

Distributed Word Representation

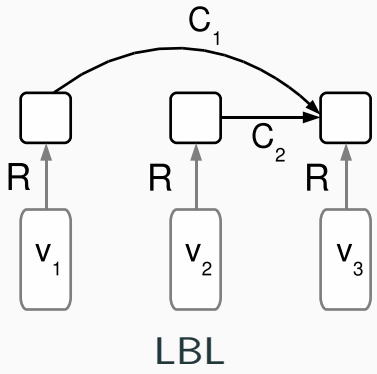


Distributed word representation is so hot in NLP community.

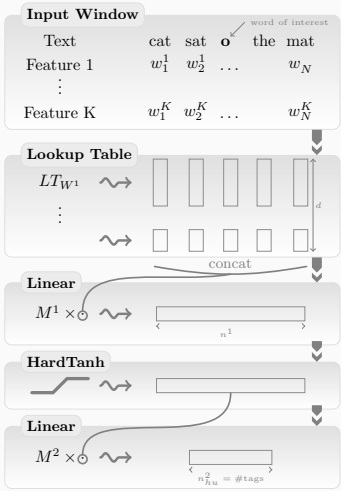
Models



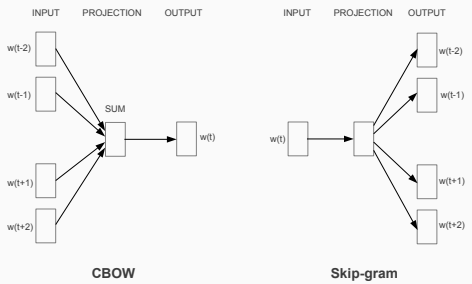
NPLM



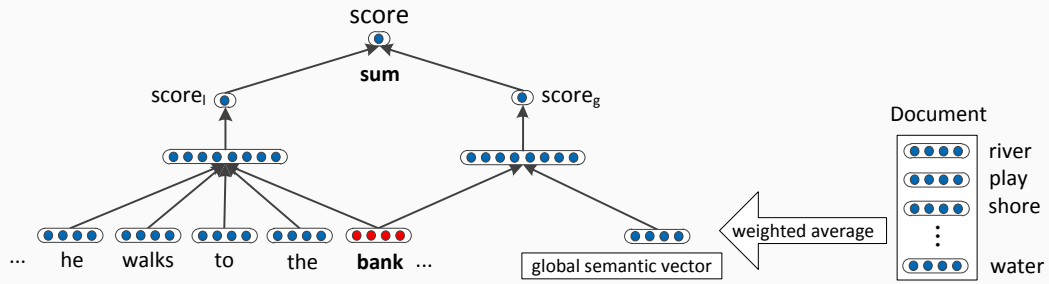
LBL



C&W



Word2Vec



Huang

State-Of-The-Art: CBOW and SG.

Dense Representation and Interpretability

Example[1]

man	$[0.326172, \ldots, 0.00524902, \ldots, 0.0209961]$
woman	$[0.243164, \ldots, -0.205078, \ldots, -0.0294189]$
dog	$[0.0512695, \ldots, -0.306641, \ldots, 0.222656]$
computer	$[0.107422, \ldots, -0.0375977, \ldots, -0.0620117]$

[1] Vectors from GoogleNews-vectors-negative300.bin.

Dense Representation and Interpretability

Example[1]

man	$[0.326172, \ldots, 0.00524902, \ldots, 0.0209961]$
woman	$[0.243164, \ldots, -0.205078, \ldots, -0.0294189]$
dog	$[0.0512695, \ldots, -0.306641, \ldots, 0.222656]$
computer	$[0.107422, \ldots, -0.0375977, \ldots, -0.0620117]$

- Which dimension represents the gender of *man* and *woman*?

[1] Vectors from GoogleNews-vectors-negative300.bin.

Dense Representation and Interpretability

Example[1]

man	$[0.326172, \ldots, 0.00524902, \ldots, 0.0209961]$
woman	$[0.243164, \ldots, -0.205078, \ldots, -0.0294189]$
dog	$[0.0512695, \ldots, -0.306641, \ldots, 0.222656]$
computer	$[0.107422, \ldots, -0.0375977, \ldots, -0.0620117]$

- Which dimension represents the gender of *man* and *woman*?
- What sort of value indicates male or female?

[1] Vectors from GoogleNews-vectors-negative300.bin.

Dense Representation and Interpretability

Example[1]

man	$[0.326172, \ldots, 0.00524902, \ldots, 0.0209961]$
woman	$[0.243164, \ldots, -0.205078, \ldots, -0.0294189]$
dog	$[0.0512695, \ldots, -0.306641, \ldots, 0.222656]$
computer	$[0.107422, \ldots, -0.0375977, \ldots, -0.0620117]$

- Which dimension represents the gender of *man* and *woman*?
- What sort of value indicates male or female?
- Gender dimension(s) would be active in all the word vectors including irrelevant words like *computer*.

[1] Vectors from GoogleNews-vectors-negative300.bin.

Dense Representation and Interpretability

Example[1]

man	$[0.326172, \ldots, 0.00524902, \ldots, 0.0209961]$
woman	$[0.243164, \ldots, -0.205078, \ldots, -0.0294189]$
dog	$[0.0512695, \ldots, -0.306641, \ldots, 0.222656]$
computer	$[0.107422, \ldots, -0.0375977, \ldots, -0.0620117]$

- Which dimension represents the gender of *man* and *woman*?
- What sort of value indicates male or female?
- Gender dimension(s) would be active in all the word vectors including irrelevant words like *computer*.
- Difficult in interpretation and uneconomic in storage.

[1] Vectors from GoogleNews-vectors-negative300.bin.

Sparse Word Representation

Non-Negative Sparse Embedding (NNSE) [Murphy et al., 2012]

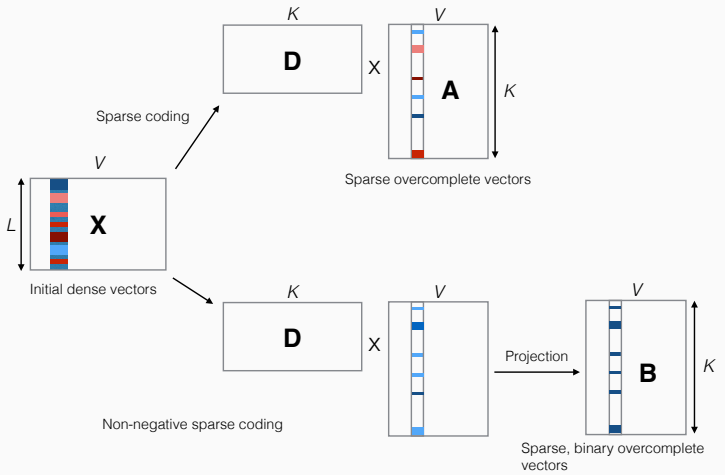
$$\arg \min_{\substack{\boldsymbol{A}\in\mathbb{R}^{m\times k}\\ \boldsymbol{D}\in\mathbb{R}^{k\times n}}} \sum_{i=1}^{m}\left(\|\boldsymbol{X}_i-\boldsymbol{A}_i\times\boldsymbol{D}\|^2+\lambda\|\boldsymbol{A}_i\|_1\right)$$

$$\text{where, } \boldsymbol{A}_{i,j}\geq 0, \forall 1\leq i\leq m, \forall 1\leq j\leq k$$

$$\boldsymbol{D}_i\boldsymbol{D}_i^{\top}\leq 1, \forall 1\leq i\leq k$$

Introduce sparse and non-negative constraints into MF.

Sparse Coding (Word2Vec) [Faruqui et al., 2015]



Convert dense vector using sparse coding in a post-processing way.

Sparse Word Representation

Non-Negative Sparse Embedding (NNSE) [Murphy et al., 2012]

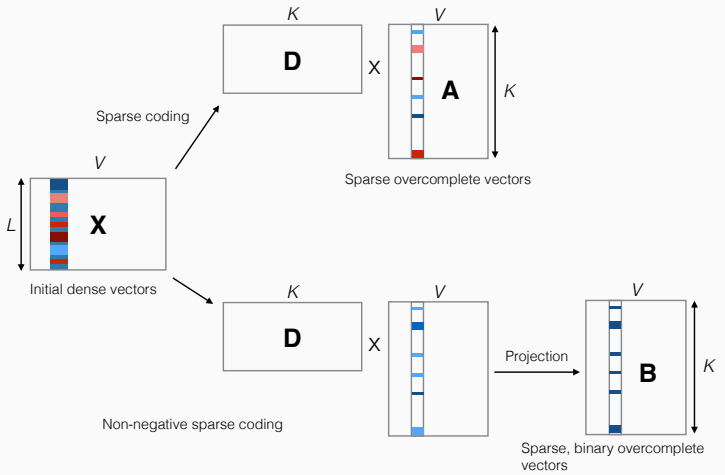
$$\arg \min_{\substack{\boldsymbol{A}\in\mathbb{R}^{m\times k}\\ \boldsymbol{D}\in\mathbb{R}^{k\times n}}} \sum_{i=1}^{m}\left(\|\boldsymbol{X}_i-\boldsymbol{A}_i\times\boldsymbol{D}\|^2+\lambda\|\boldsymbol{A}_i\|_1\right)$$

$$\text{where, } \boldsymbol{A}_{i,j}\geq 0, \forall 1\leq i\leq m, \forall 1\leq j\leq k$$

$$\boldsymbol{D}_i\boldsymbol{D}_i^{\top}\leq 1, \forall 1\leq i\leq k$$

Introduce sparse and non-negative constraints into MF.

Sparse Coding (Word2Vec) [Faruqui et al., 2015]



Convert dense vector using sparse coding in a post-processing way.

Thay are difficult to train on large-scale data for heavy memory usage!

Our Motivation

Good Performance Fast Large Scale Interpretable

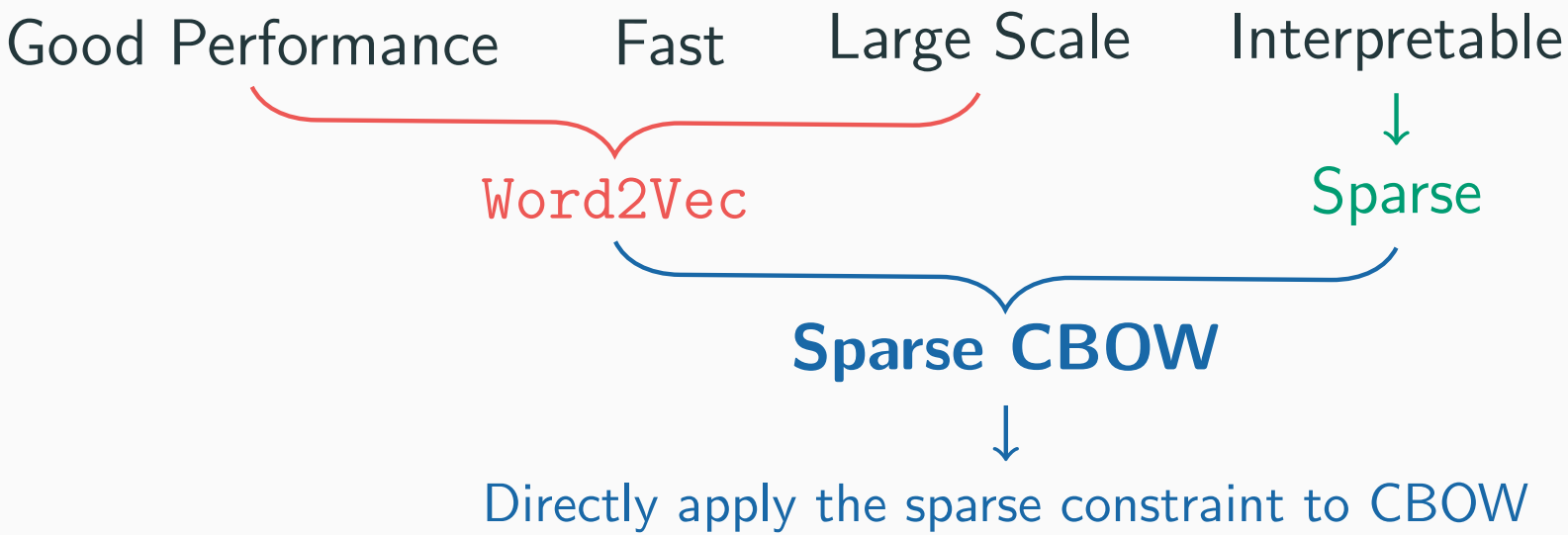
Our Motivation

Good Performance Fast Large Scale Interpretable

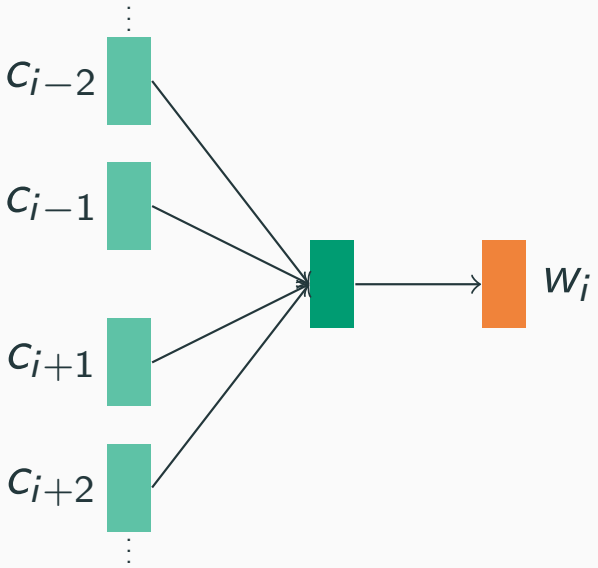
`Word2Vec`

Our Motivation

Our Motivation



CBOW



CBOW

$$\mathcal{L}_{cbow} = \sum_{i=1}^{N} \log p(w_i|h_i)$$

$$p(w_i|h_i) = \frac{\exp(\vec{w}_i \cdot \vec{h}_i)}{\sum_{w \in W} \exp(\vec{w} \cdot \vec{h}_i)}$$

$$\vec{h}_i = \frac{1}{2l} \sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \vec{c}_j$$

$$\mathcal{L}^{\text{ns}}_{\text{cbow}} = \sum_{i=1}^{N} \left(\log \sigma(\vec{w}_i \cdot \vec{h}_i) + k \cdot \mathbf{E}_{\tilde{w} \sim P_{\tilde{W}}} \log \sigma(-\vec{\tilde{w}} \cdot \vec{h}_i) \right)$$

Sparse CBOW

$$\mathcal{L}^{\text{ns}}_{\text{s}-\text{cbow}} = \mathcal{L}^{\text{ns}}_{\text{cbow}} - \lambda \sum_{w \in W} \|\vec{w}\|_1$$

Sparse CBOW

$$\mathcal{L}^{\text{ns}}_{\text{s}-\text{cbow}} = \mathcal{L}^{\text{ns}}_{\text{cbow}} - \lambda \sum_{w \in W} \|\vec{w}\|_1$$

$\downarrow$

Online, SGD

Sparse CBOW

$$\mathcal{L}^{\text{ns}}_{\text{s-cbow}} = \mathcal{L}^{\text{ns}}_{\text{cbow}} - \lambda \sum_{w \in W} \|\vec{w}\|_1$$

Online, SGD — failed

Sparse CBOW

$$\mathcal{L}^{\text{ns}}_{\text{s-cbow}} = \mathcal{L}^{\text{ns}}_{\text{cbow}} - \lambda \sum_{w \in W} \|\vec{w}\|_1$$

Online, SGD — failed

Regularized Dual Averaging (RAD) [Xiao, 2009]

Truncating using online average subgradients

RDA algorithm for Sparse CBOW

procedure SPARSECBOW($\mathcal{C}$)

  Initialize: $\vec{w}, \forall w \in W,\ \vec{c}, \forall c \in C,\ \bar{g}^0_{\vec{w}} = \vec{0}, \forall w \in W$

  for $i = 1, 2, 3, \ldots$ **do**

    $t \leftarrow$ update time of word w_i

    $$\vec{h}_i = \frac{1}{2l} \sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \vec{c}_j$$

    $$g^t_{\vec{w}_i} = \left[\mathbb{1}_h(w_i) - \sigma(\vec{w}^t_i \cdot \vec{h}_i)\right] \vec{h}_i$$

    $$\bar{g}^t_{\vec{w}_i} = \frac{t-1}{t} \bar{g}^{t-1}_{\vec{w}_i} + \frac{1}{t} g^t_{\vec{w}_i}$$ ▷ Keeping track of the online average subgradients

    Update $\vec{w}_i$ element-wise according to

    $$\vec{w}^{t+1}_{ij} = \begin{cases} 0 & \text{if } |\bar{g}^t_{\vec{w}_{ij}}| \leq \frac{\lambda}{\#(w_i)}, \\ \eta t \left(\bar{g}^t_{\vec{w}_{ij}} - \frac{\lambda}{\#(w_i)} \operatorname{sgn}(\bar{g}^t_{\vec{w}_{ij}})\right) & \text{otherwise,} \end{cases}$$ ▷ Truncating

    where, $j = 1, 2, \ldots, d$

    for $k = -l, \ldots, -1, 1, \ldots, l$ **do**

      update $\vec{c}_{i+k}$ according to

      $$\vec{c}_{i+k} := \vec{c}_{i+k} + \frac{\alpha}{2l} \left[\mathbb{1}_{h_i}(w_i) - \sigma(\vec{w}^t_i \cdot \vec{h}_i)\right] \vec{w}^t_i$$

    end for

  end for

end procedure

Evaluation

Baselines & Tasks

Baseline

- Dense representation models
  - GloVe [Pennington et al., 2014]
  - CBOW and SG [Mikolov et al., 2013]
- Sparse representation models
  - Sparse Coding (SC) [Faruqui et al., 2015]
  - Positive Pointwise Mutual Information (PPMI) [Bullinaria and Levy, 2007]
  - NNSE [Murphy et al., 2012]

Tasks

- Interpretability
  - Word Intrusion
- Expressive Power
  - Word Analogy
  - Word Similarity

Experimental Settings

Corpus: Wikipedia 2010 (1B words)

Parameters Setting:

window	negative	iteration	λ	learning rate	noise distribution
10	10	20	grid search	0.05	$\propto \#(w)^{0.75}$

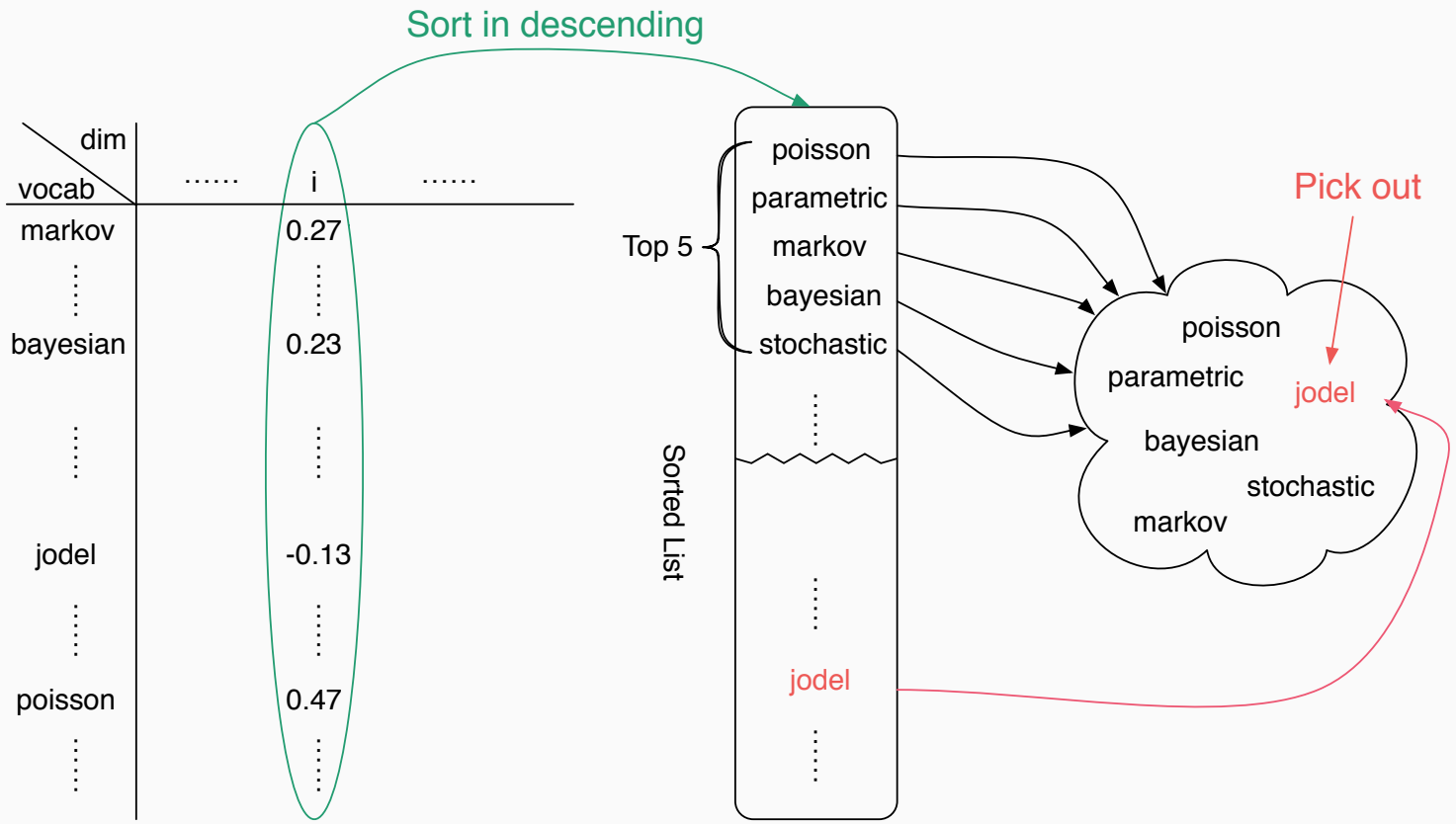
Baseline:

Model	Setting
GloVe, CBOW, SG	same setting with released tools
SC	Embeddings of CBOW as input
NNSE	PPMI matrix, 4,0000 words, SPAMS[1]

[1] http://spams-devel.gforge.inria.fr

Interpretability

Word Intrusion [Chang et al., 2009]



1. Sort dimension i in descending order.
2. A set: {top 5 words, 1 word (bottom 50% in i & top 10% in $j, j \neq i$)}
3. Pick out the intruder word

More interpretable, more easy to pick out.

Evaluation Metric

traditional metric: human assignment, subjective, costly.

Definition

$$\text{IntraDist}_i = \sum_{w_j \in \text{top}_k(i)} \sum_{\substack{w_k \in \text{top}_k(i) \\ w_k \neq w_j}} \frac{\text{dist}(w_j, w_k)}{k(k-1)}$$

$$\text{InterDist}_i = \sum_{w_j \in \text{top}_k(i)} \frac{\text{dist}(w_j, w_{b_i})}{k}$$

$$\text{DistRatio} = \frac{1}{d} \sum_{i=1}^{d} \frac{\text{InterDist}_i}{\text{IntraDist}_i}$$

IntraDist$_i$: Average distance between top k words

InterDist$_i$: Average distance between top k words and bottom word

Intuition: The intruder word should be dissimilar to the top words while those top words should be similar to each other.

In this work, we use euclidean distance as $\text{dist}(\cdot, \cdot)$

Word Intrusion Results

Table 1: 300 dimension, running 10 times.

Model	Sparsity	DistRatio
GloVe	0%	1.07
CBOW	0%	1.09
SG	0%	1.12
NNSE (PPMI)	89.15%	1.55
SC (CBOW)	88.34%	1.24
Sparse CBOW	90.06%	1.39

- Sparse is better than dense.
- Sparse CBOW **VS.** SC (CBOW): information loss in a separate sparse coding step.
- Non-negative constraint might also be a good choice.

Case Study

Model	Top 5 Words	
CBOW	beat, finish, wedding, prize, read	
	rainfall, footballer, breakfast, weekdays, angeles	
	landfall, interview, asked, apology, dinner	
	becomes, died, feels, resigned, strained	
	best, safest, iucn, capita, tallest	
Sparse CBOW	poisson, parametric, markov, bayesian, stochastic	statistical learning
	ntfs, gzip, myfile, filenames, subdirectories	file system
	hugely, enormously, immensely, wildly, tremendously	adverb for degree
	earthquake, quake, uprooted, levees, spectacularly	disasters
	bosons, accretion, higgs, neutrinos, quarks	particles

The dimensions of Sparse CBOW reveal some clear and consistent semantic meanings.

Expressive Power

Expressive Power

Task	Word Analogy	Word Similarity
Testset	Google [Mikolov et al., 2013]	Rare Word (RW) [Luong et al., 2013] WordSim-353 (WS-353) [Finkelstein et al., 2002] SimLex-999 (SL-999) [Hill et al., 2015]
Example	Beijing: China ∼ Paris: ? big: bigger ∼ deep:?	(tiger cat 7.35)
Solution	3CosMul [Levy and Goldberg, 2014]	cos
Metric	Percentage	Spearman Rank Correlation

Results

Table 2: Precision (%) for analogy and spearman correlation for similarity.

Model	Dim	Sparsity	Sem	Syn	Total	WS-353	SL-999	RW
GloVe	300	0%	79.31	61.48	69.57	59.18	32.35	34.13
CBOW	300	0%	79.38	68.80	73.60	67.21	38.82	45.19
SG	300	0%	77.79	67.32	72.09	70.74	36.07	45.55
PPMI(W-C)	40000	86.55%	74.02	38.99	53.02	62.35	24.10	30.45
PPMI(W-C)	388723	99.61%	58.55	31.19	43.60	58.99	23.01	27.98
NNSE (PPMI)[2]	300	89.15%	29.89	27.68	28.56	68.61	27.60	41.82
SC (CBOW)	300	88.34%	28.99	28.43	28.68	59.85	30.44	38.75
SC (CBOW)	3000	95.85%	74.71	61.24	67.35	68.22	39.12	44.75
Sparse CBOW	300	90.06%	73.24	67.48	70.10	68.29	44.47	42.30

- Sparse CBOW is a competitive model using much less memory.
- Similar performance on analogy if we reduce its sparsity level $< 85\%$.

[2] The input matirx of NNSE is the 40000-dimensional representations of PPMI in fourth row.

Summary

- A sparse word representation model.
- A new evaluation metric for word intrusion task.
- Improvement in word vector interpretability.
- Similar performance with less memory usage.

Thanks!

Q & A

References I


Bullinaria, J. A. and Levy, J. P. (2007).
Extracting semantic representations from word co-occurrence statistics: A computational study.
Behavior Research Methods, 39(3):510–526.


Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009).
Reading tea leaves: How humans interpret topic models.
In *NIPS*, pages 288–296. Curran Associates, Inc.


Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., and Smith, N. A. (2015).
Sparse overcomplete word vector representations.
In *Proceedings of ACL*, pages 1491–1500.


Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., andGadi Wolfman, Z. S., and Ruppin, E. (2002).
Placing search in context: The concept revisited.
ACM Trans. Inf. Syst., 20(1):116–131.

References II


Hill, F., Reichart, R., and Korhonen, A. (2015).
Simlex-999: Evaluating semantic models with (genuine) similarity estimation.
Computational Linguistics, pages 665–695.


Levy, O. and Goldberg, Y. (2014).
Linguistic regularities in sparse and explicit word representations.
In *Proceedings of CoNLL*, pages 171–180.


Luong, M.-T., Socher, R., and Manning, C. D. (2013).
Better word representations with recursive neural networks for morphology.
In *Proceedings of CoNLL*, pages 104–113.


Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
In *Proceedings of ICLR*.


Murphy, B., Talukdar, P., and Mitchell, T. (2012).
Learning effective and interpretable semantic models using non-negative sparse embedding.
In *Proceedings of COLING*, pages 1933–1950.

References III

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Proceedings of EMNLP*, pages 1532–1543.

Xiao, L. (2009).
Dual averaging method for regularized stochastic learning and online optimization.
In *Proceedings of NIPS*, pages 2116–2124.