Embedding Words as Distributions with a Bayesian Skip-gram Model

Arthur Bražinskas[1,2] Serhii Havrylov[2] Ivan Titov[1,2]

COLING 2018, Santa-Fe, New Mexico

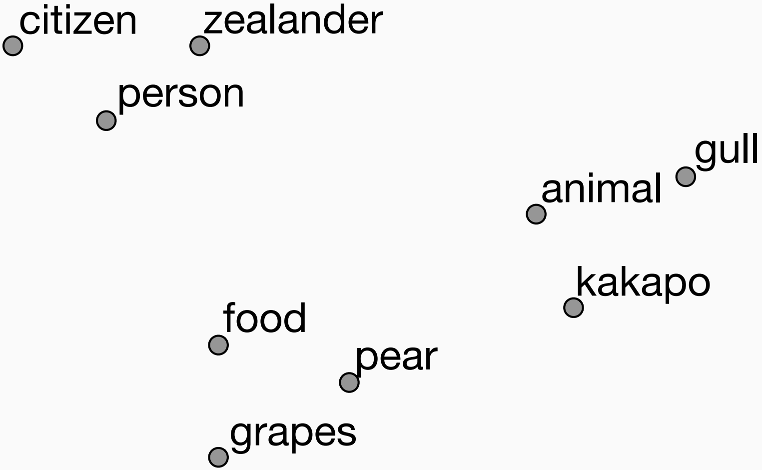
[1]ILLC, University of Amsterdam
[2]ILCC, School of Informatics, University of Edinburgh

Introduction

Word embeddings

- Unsupervised learning
- Distributional hypothesis [Harris, 1954]

Words as vectors



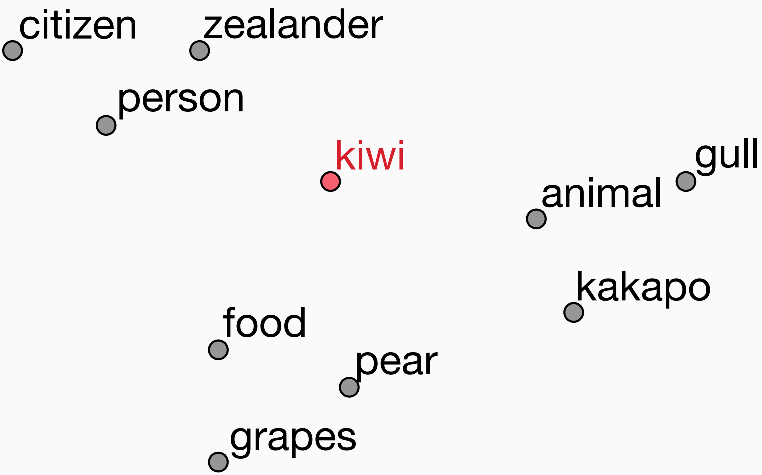
How to embed polysemous words?

(a) Kiwi fruit

(b) Kiwi bird

(c) Kiwi man

Single embeddings per word



Multiple embeddings per word

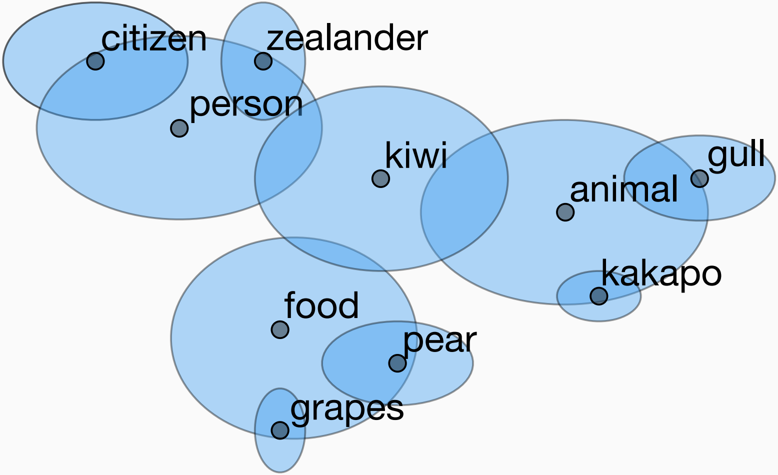
How many?

Multiple embeddings per word

- Pre-processing(e.g. clustering [Huang et al., 2012])
- Expert knowledge or assumptions(e.g. sense per word type [Neelakantan et al., 2015])

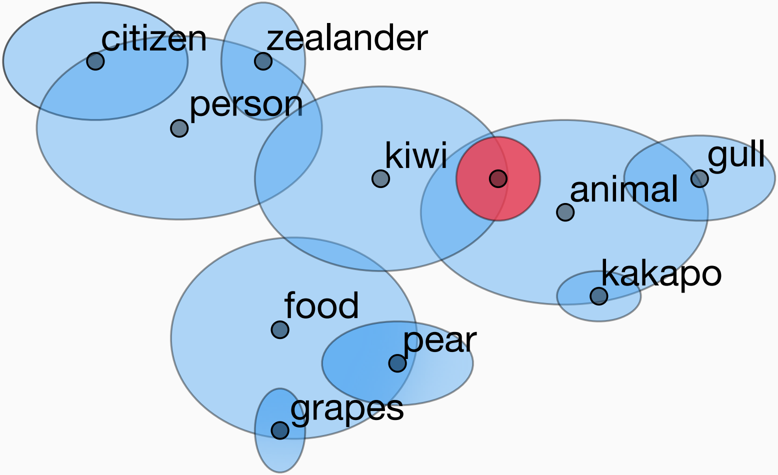
Our approach

Words as Gaussian distributions

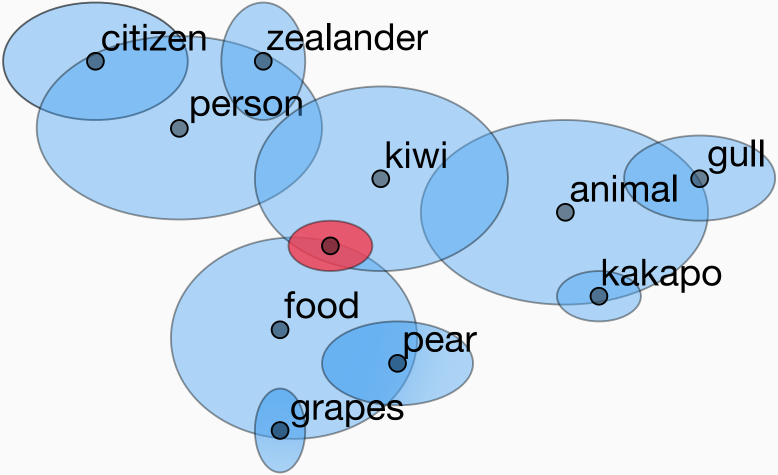


Context sensitive distributions

Ex.: I saw a small flightless **kiwi**

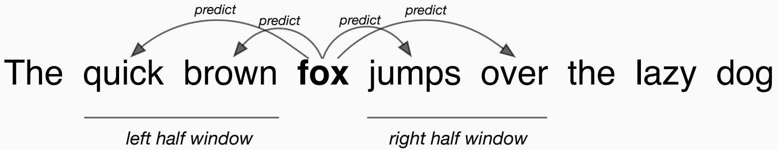


Ex.: I've bought a **kiwi** and an apple



Background

Skip-gram (SG)


predict
predict
predict
predict
The quick brown **fox** jumps over the lazy dog
left half window
right half window

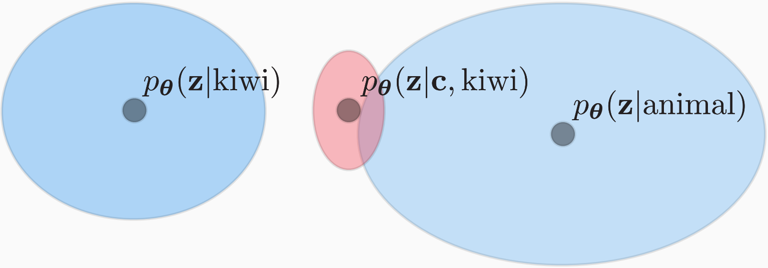
Skip-gram (SG)

- Context words directly depend on center words
- Word embeddings are vectors

Bayesian Skip-gram (BSG)

- Context words depend on '**meanings**' of center words
- 'Meaning' of center words is a **latent vector**
- Word embeddings are Gaussian distributions

BSG: Components of interest



$p_{\boldsymbol{\theta}}(\mathbf{z}|w)$ - prior distribution (static embeddings)
$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{c}, w)$ - posterior distribution (context sensitive embeddings)

BSG: Search for θ

Log-likelihood is **intractable** for gradient optimization.

$$\log \underbrace{p_{\boldsymbol{\theta}}(\mathrm{c}|w)}_{\text{likelihood}} = \log \int p_{\boldsymbol{\theta}}(\mathrm{z}|w) p_{\boldsymbol{\theta}}(\mathrm{c}|\mathrm{z}, w) d\mathbf{z}$$

BSG: Exact inference

Generation of context sensitive embeddings (inference) is also intractable.

$$\underbrace{p_{\boldsymbol{\theta}}(\mathsf{z}|w,\mathsf{c})}_{\text{posterior}} = \frac{p_{\boldsymbol{\theta}}(\mathsf{z},\mathsf{c}|w)}{\underbrace{p_{\boldsymbol{\theta}}(\mathsf{c}|w)}_{\text{likelihood}}}$$

BSG: Variational inference

To mitigate those issues, we've used variational inference: **variational auto-encoders**[Kingma and Welling, 2013]

BSG: Trick 1

Approximate with a **neural network**(ϕ) that generates a Gaussian distribution.

$$\overbrace{p_{\boldsymbol{\theta}}(\mathsf{z}|\mathsf{c}, w)}^{\text{true posterior}} \approx \underbrace{q_{\boldsymbol{\phi}}(\mathsf{z}|\mathsf{c}, w)}_{\text{approximate posterior}} = \mathcal{N}(\mathsf{z}|\underbrace{\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathsf{c}, w)}_{\text{neural network}}, \overbrace{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}(\mathsf{c}, w)}^{\text{neural network}})$$

BSG: Trick 2

Maximize the **lower-bound** instead of the log-likelihood.

$$\log p_{\boldsymbol{\theta}}(\mathbf{c}|w) \geq \overbrace{\sum_{j=1}^{C} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{c},w)} \left[\log p_{\boldsymbol{\theta}}(c_j|\mathbf{z})\right]}^{\text{reconstruction}} - \underbrace{\mathbb{D}_{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{c},w) \| p_{\boldsymbol{\theta}}(\mathbf{z}|w)\right]}_{\text{regularization}}$$

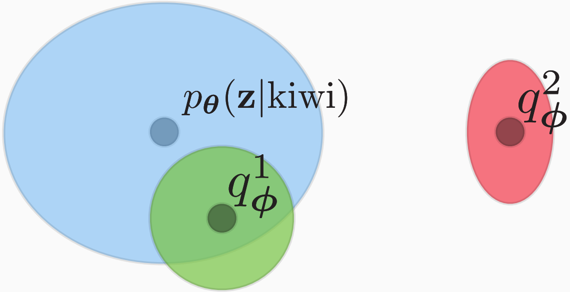
$$= \underbrace{\mathcal{L}(\boldsymbol{\theta},\phi|\mathbf{c},w)}_{\text{lower-bound}}$$

BSG: Intuition behind regularization

green

red

$$\mathbb{D}_{KL}\left[q^1_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\text{kiwi})\right] < \mathbb{D}_{KL}\left[q^2_{\boldsymbol{\phi}} \| p_{\boldsymbol{\theta}}(\mathbf{z}|\text{kiwi})\right]$$



BSG: Trick 3

Perform **gradient optimization** of $\underbrace{\mathcal{L}(\boldsymbol{\theta}, \phi | \mathbf{c}, w)}_{\text{lower-bound}}$ to find locally optimal:

- $\boldsymbol{\theta}$ - context agnostic embeddings
- $\boldsymbol{\phi}$ - context sensitive embeddings(encoder)

Experiments and Results

Setup

- Compared with:
  - Skip-gram [Mikolov et al., 2013]
  - Word2Gauss [Vilnis and McCallum, 2014]
- Trained on a concatenation of ukWaC and WaCkypedia corpora
- Approximately 3 billion tokens

Evaluation plan

Approximate posteriors $q_{\phi}(\mathrm{z}|\mathrm{c}, w)$:

- Lexical substitution

Priors $p_{\boldsymbol{\theta}}(\mathrm{z}|w) = \mathcal{N}(\mathrm{z}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$:

- Semantic word similarity ($\boldsymbol{\mu}_w$)
- Entailment directionality detection ($\boldsymbol{\Sigma}_w$)

Lexical substitution

A way to test context dependent word embeddings.

Lexical substitution

man

fruit

bird

In the $\underbrace{\text{forest I saw a flightless}}_{\textit{left context}}$ $\underbrace{\textbf{kiwi}}_{\textit{center}}$

Lexical substitution

$e(\text{man})$

$e(\text{fruit})$

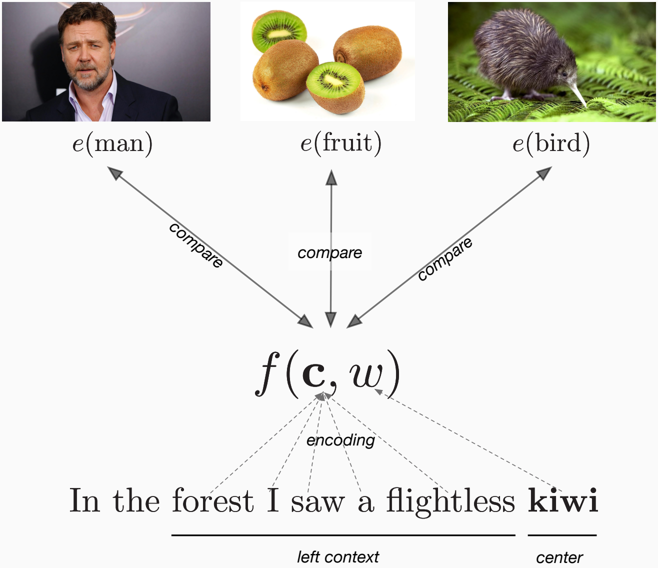
$e(\text{bird})$

$$f(\mathbf{c}, w)$$

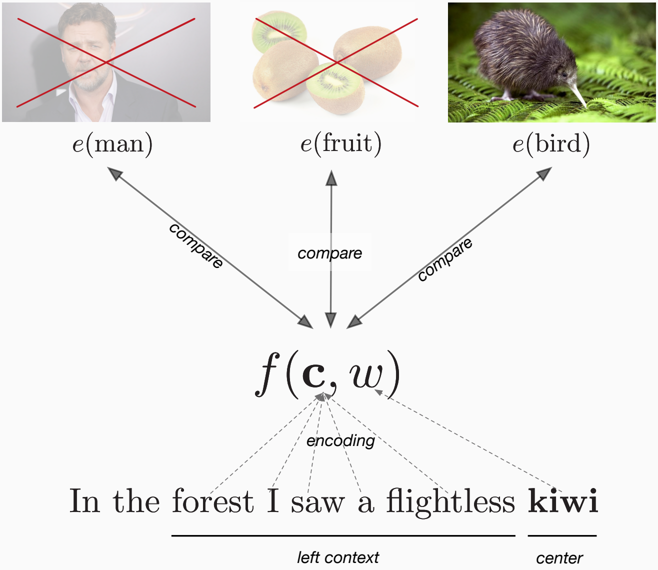
encoding

$$\text{In the } \underbrace{\text{forest I saw a flightless}}_{\textit{left context}} \; \underbrace{\textbf{kiwi}}_{\textit{center}}$$

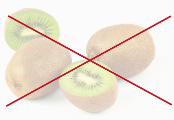
Lexical substitution



Lexical substitution



Lexical substitution: BSG

$p_{\boldsymbol{\theta}}(\mathbf{z}|man)$ $p_{\boldsymbol{\theta}}(\mathbf{z}|fruit)$ $p_{\boldsymbol{\theta}}(\mathbf{z}|bird)$

KLD *KLD* *KLD*

$$q_{\phi}(\mathbf{z}|\mathbf{c}, w)$$

In the $\underbrace{\text{forest I saw a flightless}}_{\textit{left context}}$ $\underbrace{\textbf{kiwi}}_{\textit{center}}$

Lexical substitution: Setup

- For W2G and Skip-gram dynamic embeddings $f(\mathbf{c}, w)$ used the best heuristics from [Melamud et al., 2015]
- SemEval-2007 task 10 dataset [McCarthy and Navigli, 2007]

Lexical substitution: Results

Model	GAP
BSG	**0.461**
W2G	0.432
SG	0.428

Table 1: Results in terms of generalized average precision(GAP). The higher, the better.

Lexical substitution: Conclusion

- Intuition's support for representation of **word senses**
- Effective representations

Lexical substitution: Examples

Excerpts	Top 3 Substitutes
At that size it would have a **mass** of about the same as an average galaxy	conglomeration, magnitude, bulk
Few people parallels the growing poverty of the **masses**	multitude, proletariat, throng

Word semantic similarity

- Prior means were used from W2G and BSG.
- BSG is better on 8/12 datasets than other models.

BSG	WG	SG
7.26	7.10	7.15

Table 2: Results in terms of the sum of Spearman's correlation coefficients. The higher, the better.

Word semantic similarity: Conclusion

Prior means induced by BSG are effective in capturing semantic properties of words.

Entailment directionality detection

A way to test whether $\boldsymbol{\Sigma}$ are capturing relative generality.

Entailment directionality detection

(a) Fish

(b) Shark

fish $\models$ shark or shark $\models$ fish?

Entailment directionality detection: Baseline

(a) Fish

(b) Shark

count(shark) < count(fish)

shark $\models$ fish

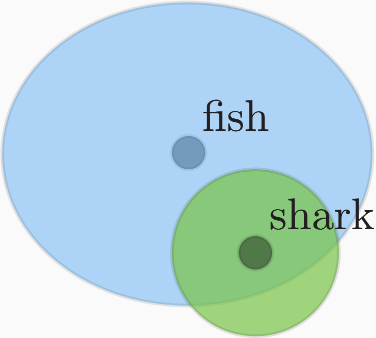
Entailment directionality detection: KLD

(a) Fish

(b) Shark

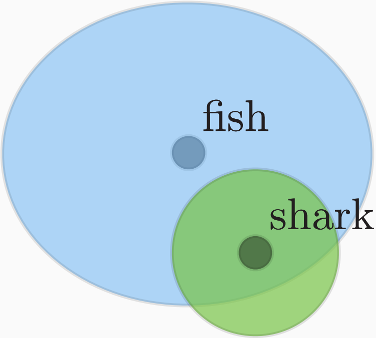
Can use the **asymmetric Kulback-Leibler divergence** function.

Entailment directionality detection: KLD



$$\mathbb{D}_{KL}\left[\text{shark}\|\text{fish}\right] \quad ? \quad \mathbb{D}_{KL}\left[\text{fish}\|\text{shark}\right]$$

Entailment directionality detection: KLD



$$\mathbb{D}_{KL}\,[\text{shark}\|\text{fish}] \quad < \quad \mathbb{D}_{KL}\,[\text{fish}\|\text{shark}]$$

Entailment directionality detection: Results

Model	BBDS	BLESS
BSG	78.23	**67.34**
W2G	78.41	57.50
Baseline	**78.84**	55.26

Table 3: Accuracy of entailment directionality detection. Baseline is based on frequency.

Entailment directionality detection: Conclusion

- BSG priors learn **generality** information beyond frequency
- W2G shows to encode frequency into $\boldsymbol{\Sigma}$

Wrap up

Summary

- Introduced a Bayesian extension of the Skip-gram model
- Static and dynamic embeddings are Gaussian distributions
- Showed an efficient model's training procedure based on the variational auto-encoders framework
- Demonstrated their effectiveness on a number of benchmarks

Special thanks to

Thank you!

Questions?

References

Zellig S Harris. Distributional structure. *Word*, 10(2-3): 146–162, 1954.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diana McCarthy and Roberto Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic*

Evaluations, pages 48–53. Association for Computational Linguistics, 2007.

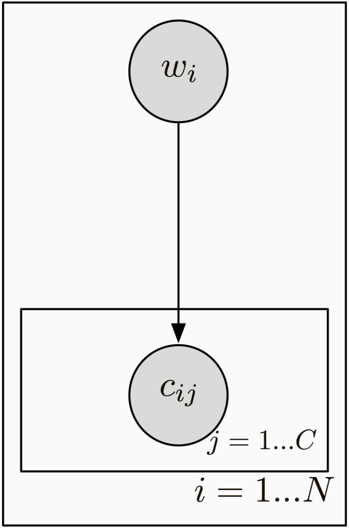
Oren Melamud, Omer Levy, and Ido Dagan. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

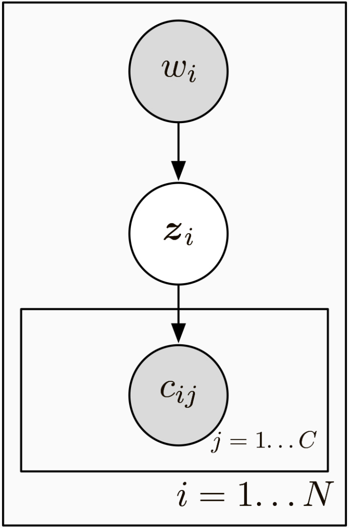
Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2015.

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014.

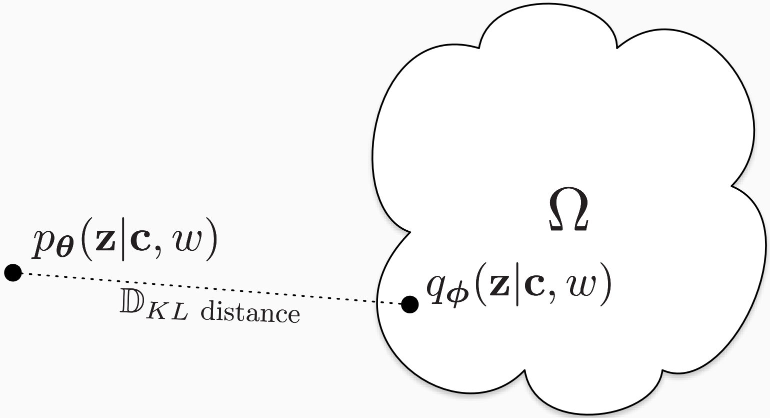
SG: graphical model



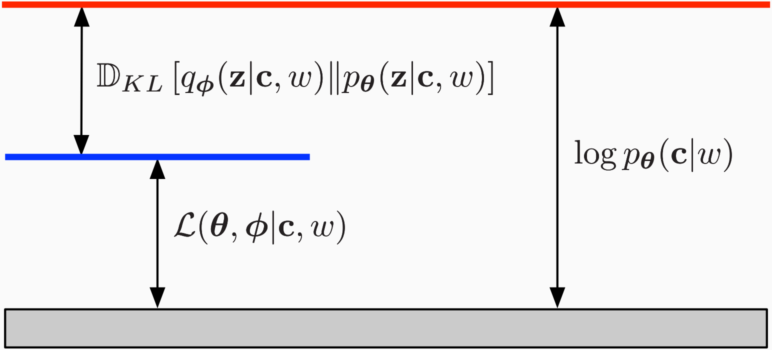
Bayesian Skip-gram (BSG)



BSG: Variational Inference



Variational Inference



Word similarity

Datasets	BSG	WG(S)	WG(D)	SG
MC-30	0.71	0.69	0.70	**0.72**
MEN-TR-3k	**0.73**	0.72	0.71	0.72
MTurk-287	**0.70**	**0.70**	0.69	**0.70**
MTurk-771	**0.67**	0.65	0.64	0.65
RG-65	0.70	0.69	0.71	**0.72**
RW-STNFRD	0.43	0.43	0.42	**0.44**
SIMLEX-999	**0.35**	0.34	0.34	0.34
VERB-143	0.32	**0.38**	0.29	0.36
WS-353-ALL	**0.72**	0.68	0.67	0.69
WS-353-REL	**0.68**	0.66	0.65	0.65
WS-353-SIM	**0.75**	0.70	0.68	0.71
YP-130	**0.50**	0.46	0.46	0.45
Sum	**7.26**	7.10	6.95	7.15

KLD

$$D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^{\mathrm{T}}\Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right)$$