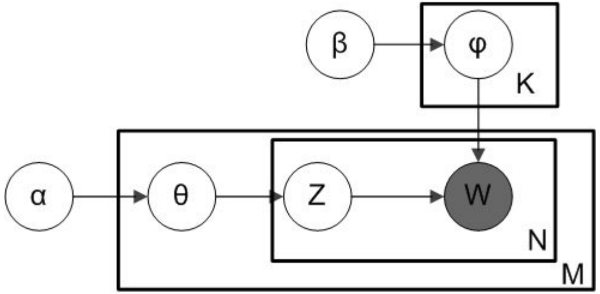
Interpretable probabilistic embeddings:
bridging the gap between topic models and neural networks

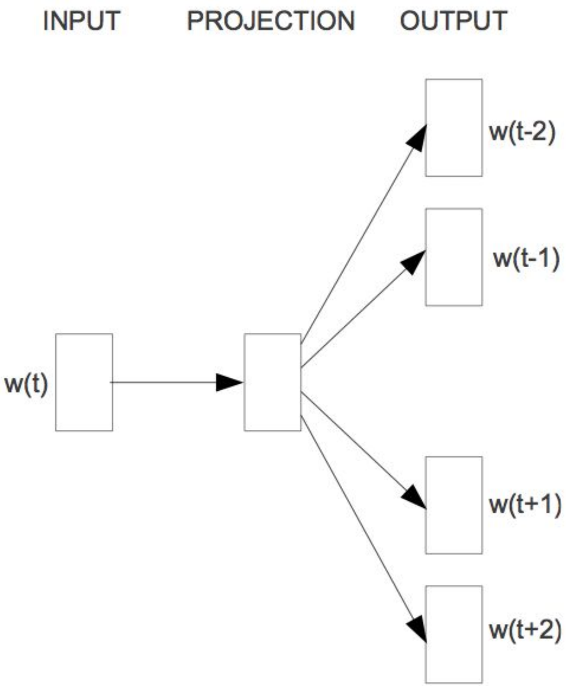
Anna Potapenko, Artem Popov, and Konstantin Vorontsov

HSE, September 13

Topic models and word embeddings (at a first glance)



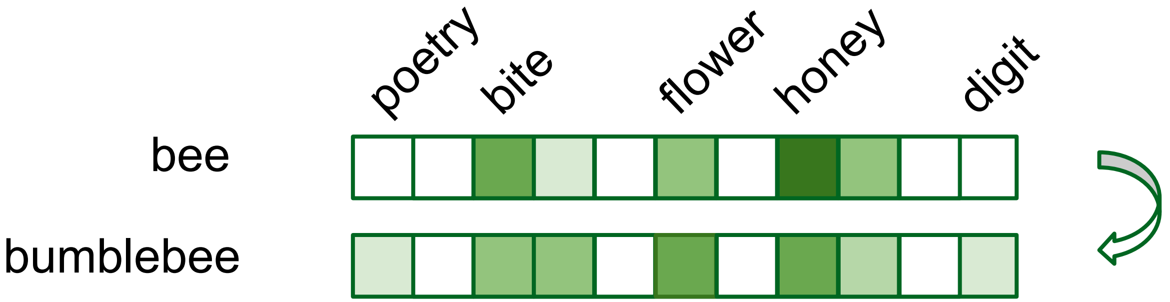
LDA



word2vec

Brief introduction to distributional semantics

- First order co-occurrences
  syntagmatic associates / relatedness (bee and honey)
- Second order co-occurrences
  paradigmatic parallels / similarity (bee and bumblebee)

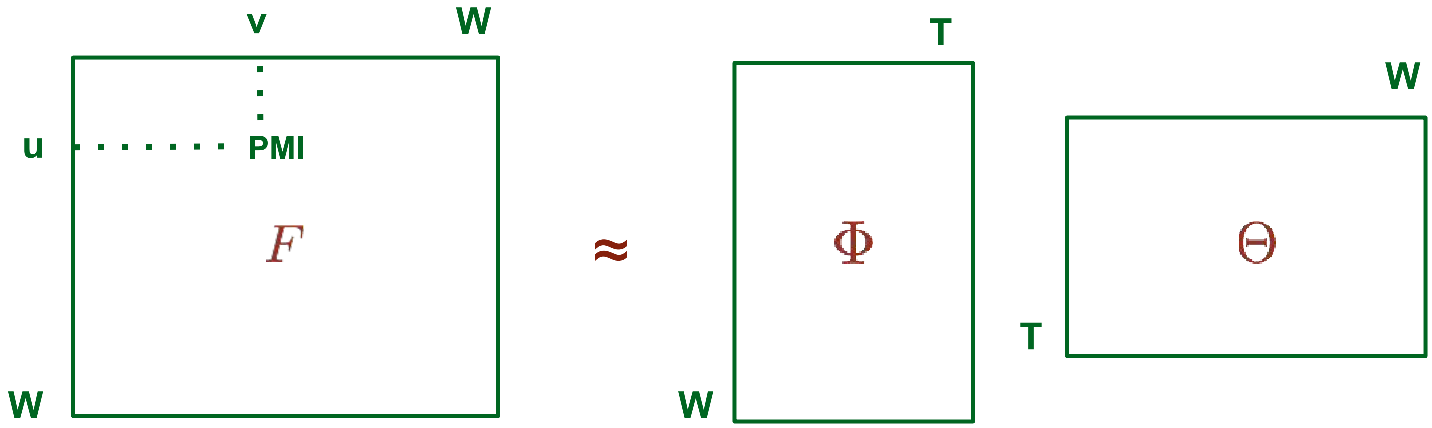


cosine similarity

Schutze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In Making Sense of Words: Proceedings of the Conference, pp. 104–113, Oxford, England.

Brief introduction to distributional semantics

- **Input:** word-word co-occurrences (counts, PMI, …)
- **Method:** dimensionality reduction (SVD, …)
- **Output:** vector representations of words



Turney, P.D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics, 2010.

Topic models and word embeddings (at a second glance)

- PLSA model:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Notation note:

- ➢ w - word
- ➢ t - topic
- ➢ d - document

- Two matrices of parameters
- Parameters are probabilities
- Utilize word-document statistics
- Learnt by EM-algorithm

Topic models and word embeddings (at a second glance)

- PLSA model:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

- Two matrices of parameters
- Parameters are probabilities
- Utilize word-document statistics
- Learnt by EM-algorithm

- Skip-Gram model:

$$p(u|v) = softmax\left(\sum_{t \in T} \phi_{ut}\theta_{tv}\right)$$

- Two matrices of parameters
- Parameters are real values
- Utilize word-word statistics
- Learn by SGD modifications

Probabilistic word embeddings (PWE)

- Probabilistic model:

$$p(u|v) = \sum_{t \in T} p(u|t)p(t|v) = \sum_{t \in T} \phi_{ut}\theta_{tv}$$

- Likelihood maximization:

$$\mathcal{L} = \sum_{v \in W} \sum_{u \in W} n_{uv} \ln p(u|v) \to \max_{\Phi,\Theta},$$

$$\phi_{ut} \geq 0, \quad \sum_{u} \phi_{ut} = 1 \qquad \theta_{td} \geq 0, \quad \sum_{t} \theta_{td} = 1$$

Zuo, Zhao, Xu: Word network topic model: a simple but general solution for short and imbalanced texts, 2016.

Additive Regularization for Topic Models

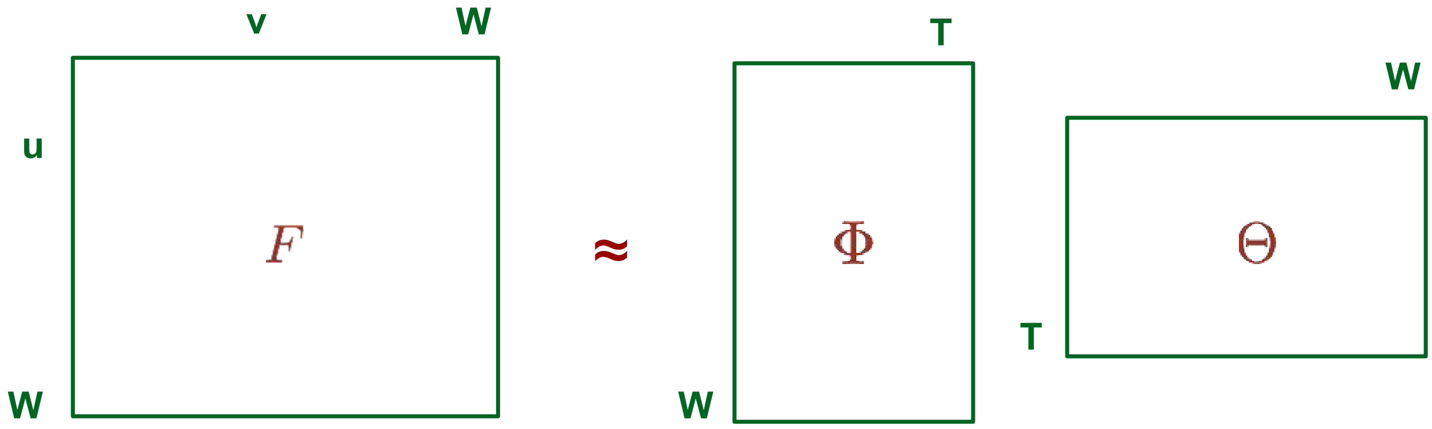
- Easy way to impose additional requirements for the embeddings
- Deals with non-uniqueness of matrix factorization

$$\mathcal{L} + R \to \max_{\Phi,\Theta}; \qquad R = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

- Examples of regularizers:
  - Sparsity: KL-divergence between topic and uniform distributions
  - Diversity: pairwise correlations of the topics

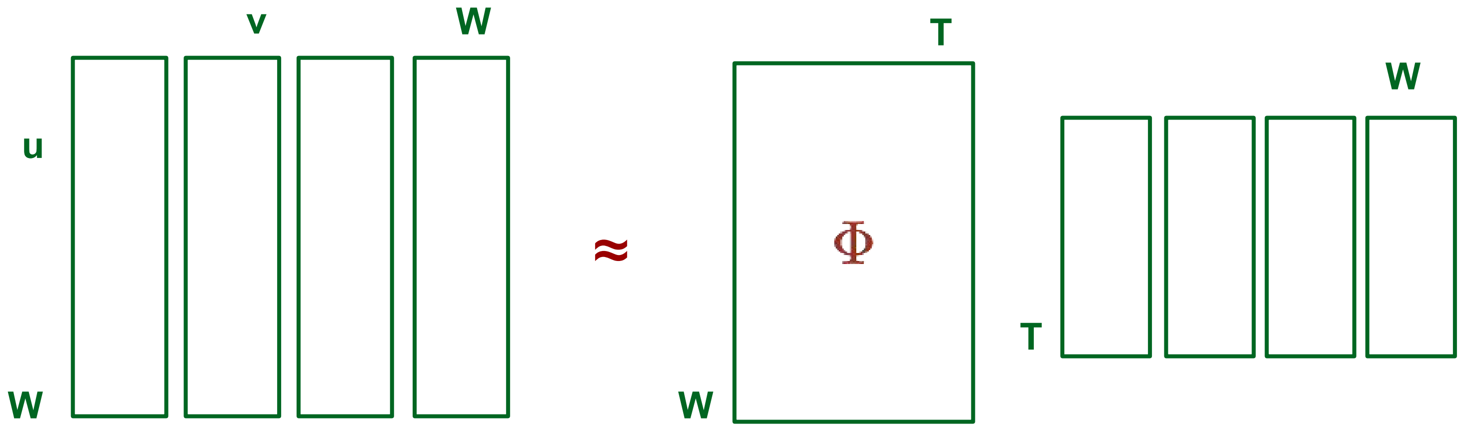
Online EM-algorithm in BigARTM

- E-step: $p(t|u,v)$ - estimate posterior probabilities for hidden variables
- M-step: Φ, Θ - update parameter to maximize expected log-likelihood



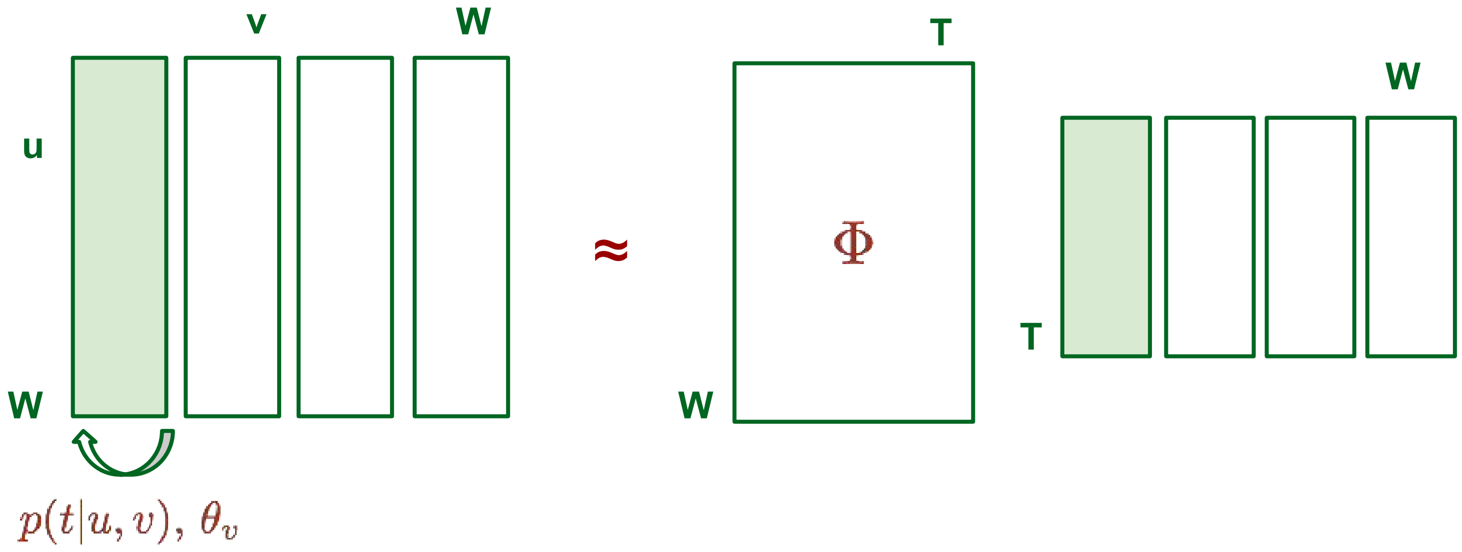
Online EM-algorithm in BigARTM

- E-step: $p(t|u,v)$ - estimate posterior probabilities for hidden variables
- M-step: Φ, Θ - update parameter to maximize expected log-likelihood



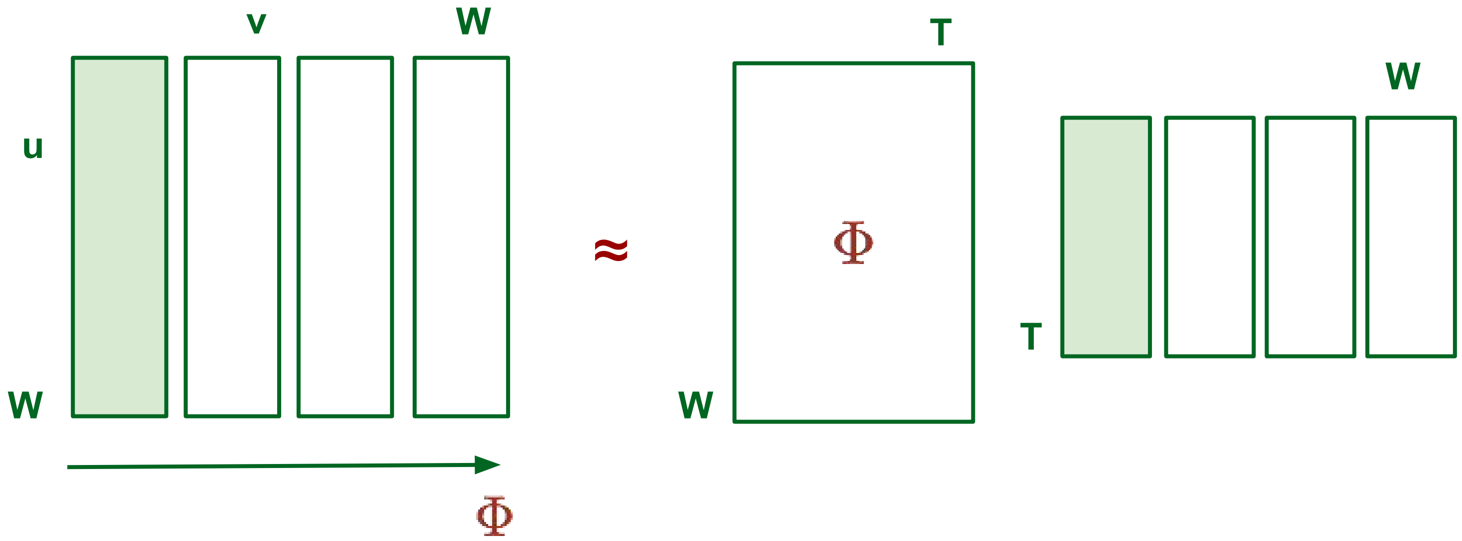
Online EM-algorithm in BigARTM

- E-step: $p(t|u,v)$ - estimate posterior probabilities for hidden variables
- M-step: Φ, Θ - update parameter to maximize expected log-likelihood



Online EM-algorithm in BigARTM

- E-step: $p(t|u,v)$ - estimate posterior probabilities for hidden variables
- M-step: Φ, Θ - update parameter to maximize expected log-likelihood



Another perspective: (implicit) matrix factorization

PWE	data type	$F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u \mid v)$
	objective	$\sum_{v \in W} n_v \, \mathrm{KL}\left(\hat{p}(u \mid v) \middle\| \langle \phi_u \theta_v \rangle\right) \to \min\limits_{\Phi, \Theta}$
	constrains	$\phi_{ut} > 0, \quad \sum_u \phi_{ut} = 1; \quad \theta_{tv} > 0, \quad \sum_t \theta_{tv} = 1$
	technique	EM-algorithm (online by F columns)
SGNS	data type	$F_{uv} = \log \frac{n_{uv} n}{n_u n_v} - \log k$
	objective	$\sum_{u \in W} \sum_{v \in W} n_{uv} \log \sigma\left(\langle \phi_u \theta_v \rangle\right) + k \, \mathbb{E}_{\bar{v}} \log \sigma\left(-\langle \phi_u \theta_v \rangle\right) \to \max_{\Phi, \Theta}$
	constrains	No constraints
	technique	SGD (online by corpus)
GloVe	data type	$F_{uv} = \log n_{uv}$
	objective	$\sum_{v \in W} \sum_{u \in W} f(n_{uv}) \left(\langle \phi_u \theta_v \rangle + b_u + \tilde{b}_v - \log n_{uv} \right)^2 \to \min_{\Phi, \Theta, b, \tilde{b}}$
	constrains	No constraints
	technique	AdaGrad (online by F elements)
NNSE	data type	$F_{uv} = max(0, \log \frac{n_{uv} n}{n_u n_v})$ or SVD low-rank approximation
	objective	$\sum_{u \in W} \left(\lVert f_u - \phi_u \Theta \rVert^2 + \lVert \phi_u \rVert_1 \right) \to \min_{\Phi, \Theta}$
	constrains	$\phi_{ut} \geq 0, \forall u \in W, t \in T \quad \theta_t \theta_t^T \leq 1, \forall t \in T$
	technique	Online algorithm from [25]

Take away from this part:

- Topic models can be considered as a way of learning word embeddings!
  - PLSA and Skip-Gram optimize very similar objectives
  - Both are parametrized with two matrices (probabilistic vs. real-valued)
  - PLSA, (SGNS), GloVe, NNSE, and many others perform low-rank matrix factorization
- Can probabilistic word embeddings perform on par with SGNS?

Word Similarity task: setup

- How do we test that similar words have similar vectors?
  - What do we mean by “similar words”? Computational linguists know a lot.
  - We can use human judgments of word pairs similarity.
  - Depends on downstream tasks! So extrinsic evaluation would be better.

Word Similarity task: setup

- How do we test that similar words have similar vectors?
  - What do we mean by “similar words”? Computational linguists know a lot.
  - We can use human judgments of word pairs similarity.
  - Depends on downstream tasks! So extrinsic evaluation would be better.

tiger	cat	7.35
tiger	tiger	10.00
plane	car	5.77
train	car	6.31
television	radio	6.77
media	radio	7.42
bread	butter	6.19
cucumber	potato	5.92
doctor	nurse	7.00
professor	doctor	6.62
student	professor	6.81
smart	stupid	5.81

Word Similarity task: results

Model	Data	Algorithm	Metric	WordSim Similarity	WordSim Related.	WordSim Joint	Bruni et. al MEN	Radinsky M. Turk
SGNS	sPMI	SGD	cos	**0.752**	0.632	0.666	**0.745**	**0.661**
LDA	n_{wd}	online EM	*hel*	0.530	0.455	0.474	0.583	0.483
PWE	n_{uv}	offline EM	*dot*	0.709	0.635	0.654	0.658	0.590
PWE	pPMI	offline EM	*dot*	0.701	0.615	0.647	0.707	0.613
PWE	n_{uv}	online EM	*dot*	0.718	**0.673**	**0.685**	0.669	0.639
sPWE	n_{uv}	online EM	*dot*	0.728	0.672	0.680	0.675	0.635

All models trained on Wikipedia 2016-01-13 dump with W = 100000. Sparsity of embeddings: 93%.

For the rest of the talk:

- What are the benefits of the new approach?
    - Sparsity (and any further requirements)
    - Interpretability of the components
    - Easy way to get document embeddings
    - Multimodal embeddings (e.g. for authors, timestamps, categories...)

For the rest of the talk:

- What are the benefits of the new approach?
  - Sparsity (and any further requirements)
  - Interpretability of the components
  - Easy way to get document embeddings
  - Multimodal embeddings (e.g. for authors, timestamps, categories...)

Can the components have meaning?

- Unfortunately, this is definitely not happening in reality
- But we could try to make a step in this direction...

No interpretability for SGNS:

- **Component 1:** avg hearth soc protector decomposition whip stochastic sewer splinter accessory howie thief thermodynamic boltzmann equilibrium kingship unconscious sophomore
- **Component 2:** rainy miocene snowy horner cfb triassic eleventh amadeus dams tenth mesozoic fourteenth thirteenth ninth diaries bight demographics seventh almanac eocene
- **Component 3:** gnis usda bloomberg usgs regulator nhk gerd magnetism capacitor fed classifies capacitance stadt bipolar multilateral trpod kunst reciprocal smiths potassium ipc

Some interpretability for PWE:

- **Component 1:** scottish scotland edinburgh glasgow mps oxford educated cambridge college aberdeen dundee royal uk scots fellows fife corpus kingdom thistle eton angus mac trinity stirling
- **Component 2:** game games video gameplay multiplayer puzzle mario nintendo player gaming pok playable mortal super kombat adventure rpg ds puzzles online smash zelda ign poker
- **Component 3:** election party elected elections parliament assembly seats members minister legislative electoral liberal council representatives parliamentary democratic senate seat prime

Interpretability: setup

- **With experts:** word intrusion task
- **Without experts:** coherence (many variants)

scottish

scotland

edinburgh

glasgow

…

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

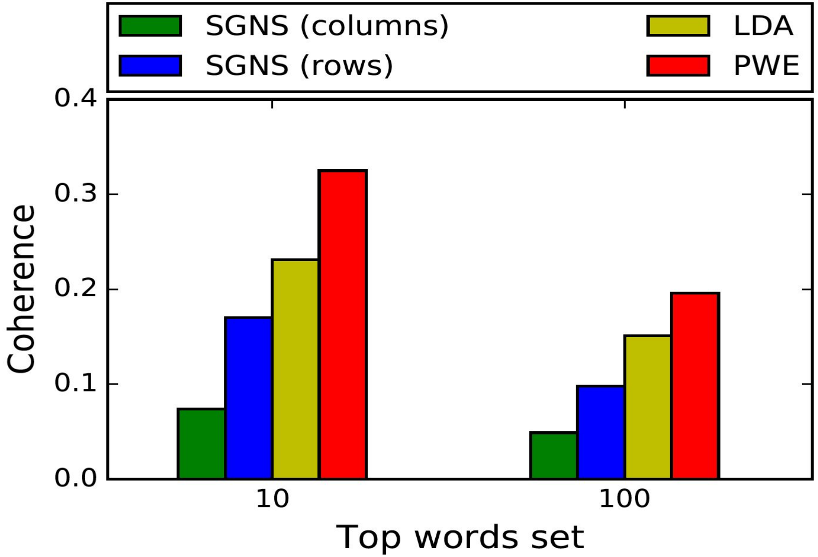
$$\mathcal{C} = \frac{2}{k(k-1)} \sum_{j=2}^{k} \sum_{i=1}^{j} PMI(w_i, w_j)$$

Murphy, Talukdar, Mitchell: Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding, 2012.

Newman, D., Bonilla, E.V., Buntine,W.L.: Improving topic coherence with regularized topic models. NIPS-2011.

Roder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. WSDM-2015.

Interpretability: results



- All models trained on Wikipedia
- Two options for SGNS
- PWE outperforms both

For the rest of the talk:

- What are the benefits of the new approach?
  - Sparsity (and any further requirements)
  - Interpretability of the components
  - Easy way to get document embeddings
  - Multimodal embeddings (e.g. for authors, timestamps, categories...)

Document Similarity Task: setup

- ArXiv triplets: paper A, similar paper B, dissimilar paper C
- Similarity defined automatically by shared subjects

http://arxiv.org/pdf/1206.5743	http://arxiv.org/pdf/cond-mat/0403258	http://arxiv.org/pdf/1408.0189
http://arxiv.org/pdf/1209.0268	http://arxiv.org/pdf/1307.7598	http://arxiv.org/pdf/math/0504051
http://arxiv.org/pdf/hep-ph/9908436	http://arxiv.org/pdf/nucl-th/9707019	http://arxiv.org/pdf/1112.3014
http://arxiv.org/pdf/1111.2905	http://arxiv.org/pdf/1303.2538	http://arxiv.org/pdf/1109.1922
http://arxiv.org/pdf/nucl-ex/0112013	http://arxiv.org/pdf/physics/9704013	http://arxiv.org/pdf/1408.4595
http://arxiv.org/pdf/0709.3419	http://arxiv.org/pdf/quant-ph/0611134	http://arxiv.org/pdf/0902.0616
http://arxiv.org/pdf/hep-th/9609148	http://arxiv.org/pdf/solv-int/9710009	http://arxiv.org/pdf/astro-ph/0508060

20 000
triplets

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015.

Document Similarity Task: setup

Integral formula of Minkowski type and new characterization of the Wulff shape

Yijun He *   Haizhong Li †

Abstract

Given a positive function F on S^n which satisfies a convexity condition, we introduce the r-th anisotropic mean curvature M_r for hypersurfaces in $\mathbb{R}^{n+1}$ which is a generalization of the usual r-th mean curvature H_r. We get integral formulas of Minkowski type for compact hypersurfaces in R^{n+1}. We give some new characterizations of the Wulff shape by use of our integral formulas of Minkowski type, in case $F = 1$ which reduces to some well-known results.

2000 Mathematics Subject Classification: Primary 53C42, 53A30; Secondary 53B25.

Key words and phrases: Wulff shape, F-Weingarten operator, anisotropic principal curvature, r-th anisotropic mean curvature, integral formula of Minkowski type.

aper C

df/1408.0189

df/math/0504051

pdf/1112.3014

pdf/1109.1922

pdf/1408.4595

pdf/0902.0616

pdf/astro-ph/0508060

ragraph Vectors, *CoRR, 2015.*

Document Similarity Task: setup

aper C

Integral formula of Minkowski type and new characterization of the Wulff shape

Yijun He * Haizhong Li †

Abst

Given a positive function F on S^n w
introduce the r-th anisotropic mean curva
is a generalization of the usual r-th mean
of Minkowski type for compact hypersurf
terizations of the Wulff shape by use of c
in case $F = 1$ which reduces to some well-

2000 Mathematics Subject Classificat
53B25.

Key words and phrases: Wulff shape, F-W
curvature, r-th anisotropic mean curvature,

COMPLEX CURVES IN ALMOST-COMPLEX MANIFOLDS AND MEROMORPHIC HULLS

Sergei IVASHKOVICH – Vsevolod SHEVCHISHIN

Preface

Chapter I. Local Properties of Complex Curves.

Lecture 1. Complex Curves in Almost-Complex Manifolds. . . . pp. 1–12
1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Document Similarity Task: setup

Integral formula o

characterizati

Yijun He

Given a positive function F o
introduce the r-th anisotropic mea
is a generalization of the usual r-t
of Minkowski type for compact hy
terizations of the Wulff shape by
in case $F = 1$ which reduces to so

2000 Mathematics Subject Cla
53B25.

Key words and phrases: Wulff shape, F-
curvature, r-th anisotropic mean curvature,

Accepted for publication in *Solar Physics*, waiting for the authoritative version and a DOI which will be available at http://www.springerlink.com/content/0038-0938

Time-dependent Stochastic Modeling of Solar Active Region Energy

M. Kanazir and M. S. Wheatland[1]

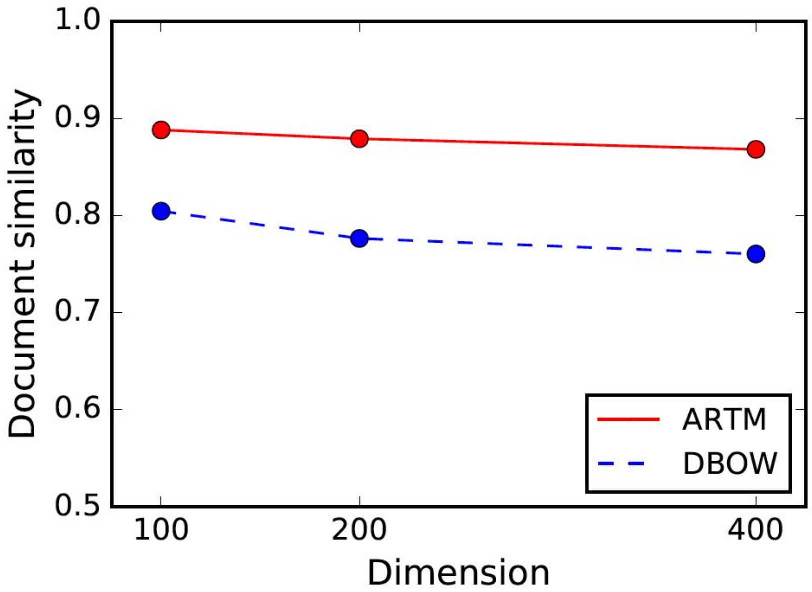
Received: 7 July 2010 / Accepted: 31 July 2010 / Published online: ●●●●●●●●●●●

Abstract A time-dependent model for the energy of a flaring solar active region

1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Document Similarity Task: results

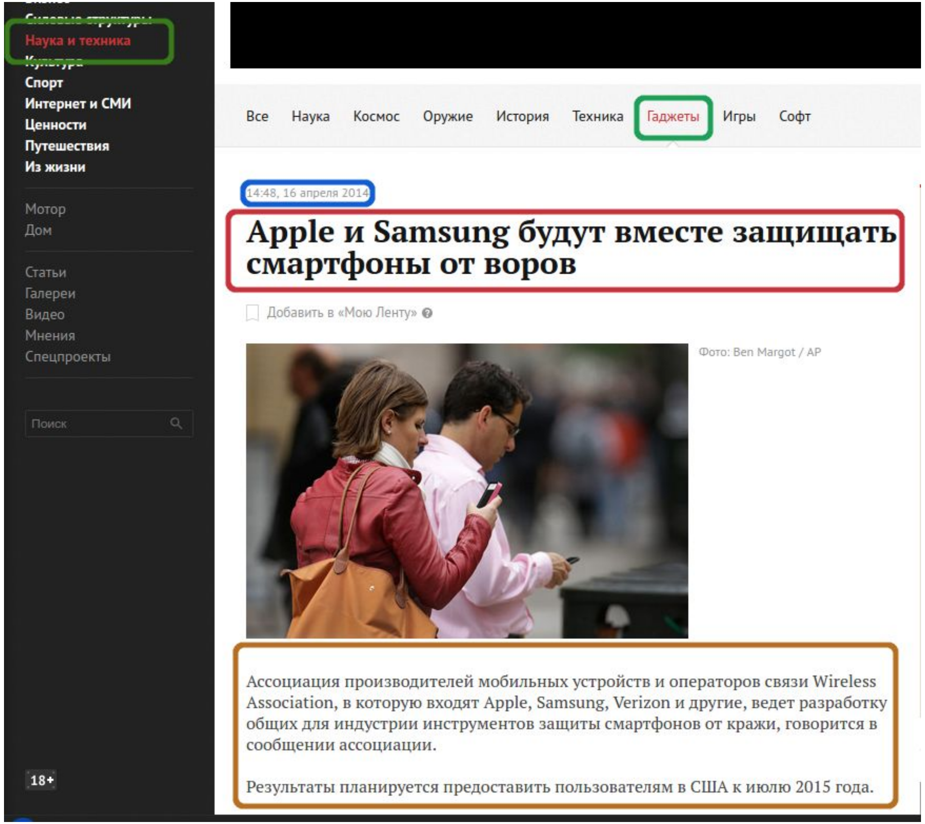
- Trained on ~1 mln ArXiv plain texts, tested on the ArXiv triplets
- DBOW is a well-known paragraph2vec architecture [Dai et. al, 2015]



For the rest of the talk:

- What are the benefits of the new approach?
    - Sparsity (and any further requirements)
    - Interpretability of the components
    - Easy way to get document embeddings
    - Multimodal embeddings (e.g. for authors, timestamps, categories...)

Multimodal news corpus Lenta.ru



ARTM is on rescue!

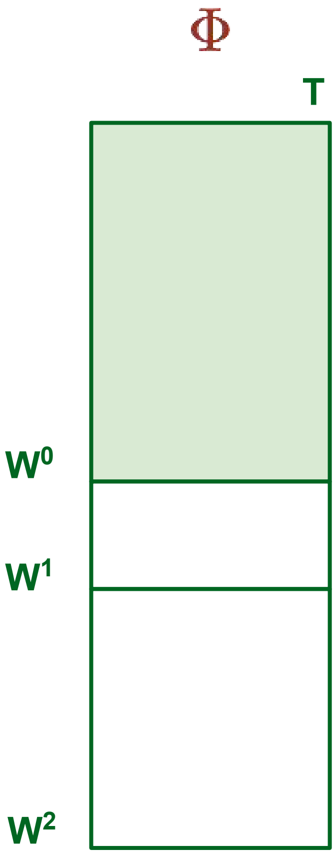
- Log-likelihood for multiple modalities:

$$\sum_{m\in M} \lambda_m \underbrace{\sum_{v\in W^0} \sum_{u\in W^m} n_{uv} \ln p(u|v)}_{\text{modality log-likelihood } \mathcal{L}_m(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

$$\phi_{ut} \geq 0, \quad \sum_{u\in W^m} \phi_{ut} = 1, \ \forall m \in M;$$

$$\theta_{tv} \geq 0, \quad \sum_{t\in T} \theta_{tv} = 1.$$

- Still trained by (modified) EM-algorithm
- Gives embeddings for all modalities in the same space



Word similarities

- Trained on ~100000 Lenta.ru new, tested on word similarity testsets
- Our approach outperforms SGNS on most of the datasets
- Interestingly, additional modalities improve similarities between words

Model	WordSim Similarity	WordSim Relatedness	WordSim+RG+ MC	SimLex
SGNS	0.630	0.530	0.567	**0.24**
PWE	0.649	0.565	0.604	0.12
Multi-PWE	**0.682**	**0.580**	**0.611**	0.14

We used the testsets translated to Russian:
http://russe.nlpub.ru/downloads/
http://www.leviants.com/ira.leviant/MultilingualVSMdata.html

Intermodality similarities

2015-12-18 Star Wars Release	2016-02-29 The Oscars	2015-05-09 Victory Day
jedi sith fett anakin chewbacca film series hamill prequel awaken boyega	statuette award nomination linklater oscar birdman win criticism director lubezki	great anniversary normandy parade demonstration vladimir celebration concentration auschwitz photograph

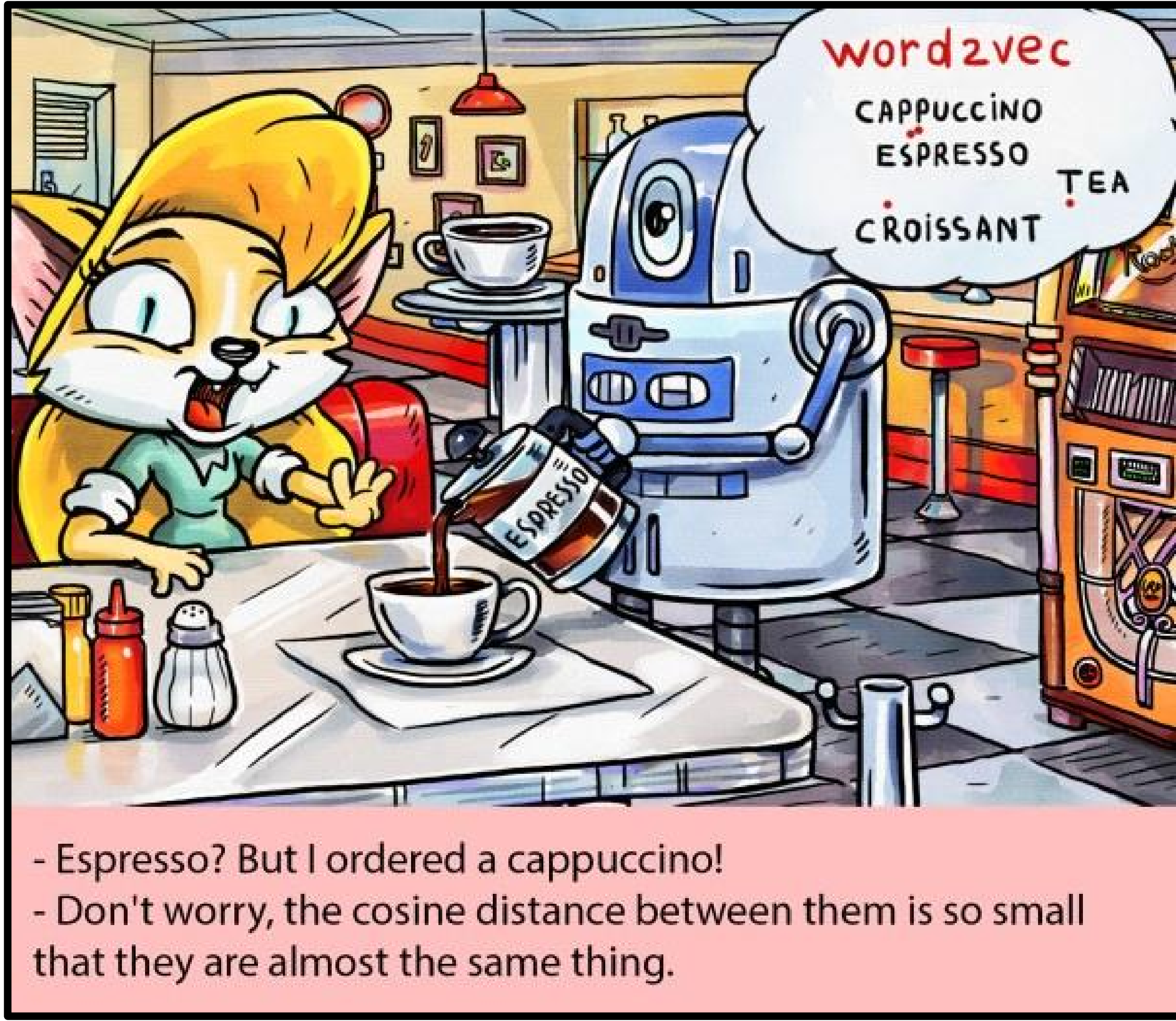
Since words and timestamps are embedded to the same space - we can look for the closest neighbours.

Conclusion:

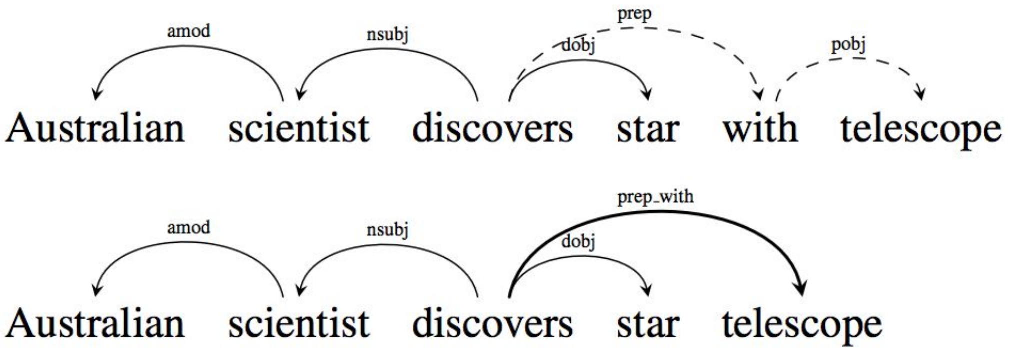
- Experiments just covered:
  - Word similarities vs. interpretability (Wikipedia)
  - Document similarities (ArXiv)
  - Multiple modalities (Letna.ru)

- Contribution: we proposed to learn probabilistic word embeddings with topic modeling and could take the best of the two worlds:
  - Good performance on word similarity tasks
  - Interpretability of the components
  - Easy extensions with additive regularization for topic models

Thanks!
Questions?



What is a context?

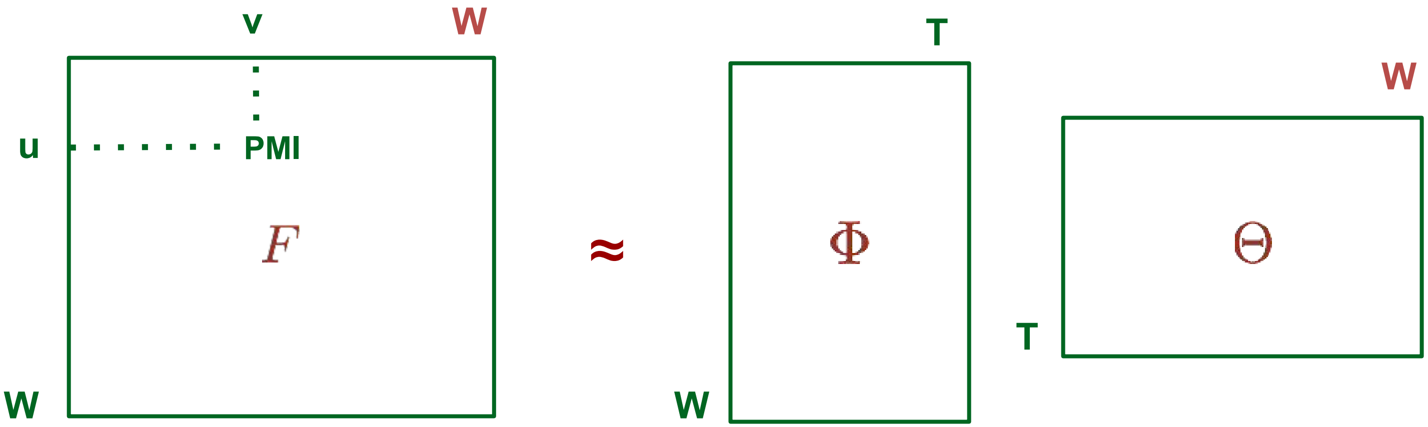


WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Omer Levy, Yoav Goldber, Dependency-Based Word Embeddings, ACL-2014.

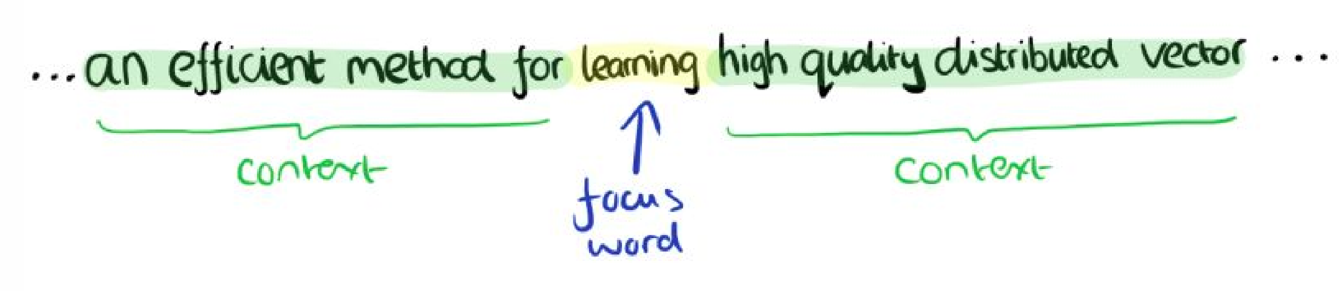
What is a context?

- Usually we use words form a sliding window (no dependency parses)
- Thus W = C, and F is a symmetric word co-occurrence matrix



What is a context?

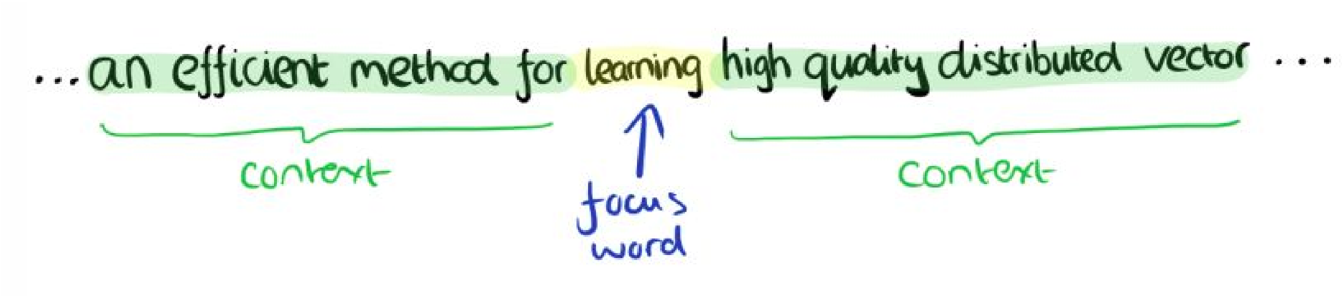
- By context, people can mean lots of different things:



CBOW	$p(w_j \mid w_{j-h}, \dots w_{j+h})$
Skip-Gram	$p(w_{j-h}, \dots w_{j+h} \mid w_j)$
PV-DBOW	$p(w_j \mid d)$

What is a context?

- By context, people can mean lots of different things:



CBOW $p(w_j | w_{j-h}, \dots w_{j+h})$

Skip-Gram $p(w_{j-h}, \dots w_{j+h} | w_j) = \prod_{i=j-h}^{j+h} p(w_i | w_j)$

PV-DBOW $p(w_j | d)$