

Towards Interpretable Word Embeddings through Probabilistic Topic Modeling

Konstantin Vorontsov

Head of Machine Intelligence Laboratory,
Moscow Institute of Physics and Technology, Russia

Korean–Russian Artificial Intelligence Workshop
Moscow • December 18, 2017

1 Theory

- Probabilistic topic modeling
- ARTM: Additive Regularization for Topic Modeling
- Multimodal ARTM

2 Implementation

- BigARTM project
- The regularizers zoo
- Benchmarking

3 Applications

- Exploratory search
- Probabilistic topical word embeddings
- Dialog segmentation

Topic modeling applications

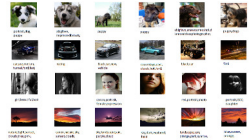
exploratory search
in digital libraries



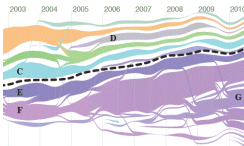
personalized search
in social media



multimodal search
for texts and images



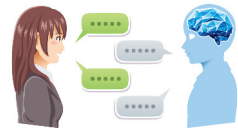
topic detection and
tracking in news flows



navigation in big
text collections



dialog manager in
chatbot intelligence



What is a “topic” in a text collection

Intuitively,

- *Topic* is a specific terminology of a particular domain area
- *Topic* is a set of terms that often co-occur in documents

More formally,

- *topic* is a probability distribution over terms (words, tokens):
 $p(w|t)$ is the frequency of term w in topic t
- *document profile* is a probability distribution over *topics*:
 $p(t|d)$ is the frequency of topic t in document d

When writing term w in document d author thought of topic t .

Topic model uncovers the set T of latent topics in a text collection.

Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned articles.
Top 10 words and their probabilities $p(w|t)$ in %:

topic #68				topic #79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Assessors evaluated 396 topics from 400 as paired and interpretable.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Example. Multilingual topic model of Wikipedia

216 175 of Russian–English parallel not-aligned articles.
Top 10 words and their probabilities $p(w|t)$ in %:

topic #88				topic #251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Assessors evaluated 396 topics from 400 as paired and interpretable.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Problem setup

Given: a set of terms W , a set of documents D ,

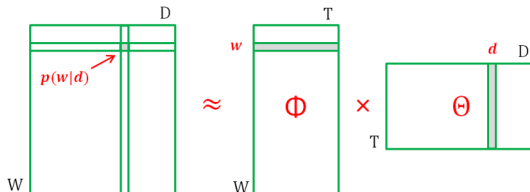
n_{dw} = how many times term w appears in document d

Find: parameters $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ of the topic model

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

subject to $\phi_{wt} \geq 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geq 0$, $\sum_t \theta_{td} = 1$.

This is a problem of *nonnegative matrix factorization*:



PLSA — Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]

Constrained maximization of the log-likelihood:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

EM-algorithm is a simple iteration method for the nonlinear system

$$\begin{cases} \text{E-step:} & p_{tdw} \equiv p(t|d, w) = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} \right) \end{cases} \end{cases}$$

where $\text{norm}_{t \in T} x_t = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ is vector normalization.

Well-posed and ill-posed problems in the sense of Hadamard (1923)

The problem is *well-posed* if

- a solution exists,
- the solution is unique,
- the solution is stable w.r.t. initial conditions.



Jacques Hadamard
(1865–1963)

Matrix factorization is an *ill-posed* inverse problem.

If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

- $\Phi' \Theta' = (\Phi S)(S^{-1} \Theta)$, where $\text{rank } S = |T|$
- $\mathcal{L}(\Phi', \Theta') = \mathcal{L}(\Phi, \Theta)$
- $\mathcal{L}(\Phi', \Theta') \leq \mathcal{L}(\Phi, \Theta) + \varepsilon$ for approximate solutions

Additional *regularizing criteria* should narrow the set of solutions.

LDA — Latent Dirichlet Allocation [D.Blei, A.Ng, M.Jordan, 2003]

Maximum a posteriori probability (MAP) **with Dirichlet prior**.

The prior can be reinterpreted as cross-entropy minimization:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{cross-entropy regularization}} \rightarrow \max_{\Phi, \Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \beta_w \right) \\ \theta_{td} = \text{norm}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \alpha_t \right) \end{cases} \end{cases}$$

ARTM — Additive Regularization for Topic Modeling

Maximum log-likelihood **with regularization criterion** $R(\Phi, \Theta)$:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{array}{l} \text{E-step:} \\ \text{M-step:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

K. Vorontsov. Additive regularization for topic models of text collections. 2014.

Combining topic models via additive regularization

Maximum log-likelihood **with additive combination** of regularizers:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^n \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

where τ_i are regularization coefficients.

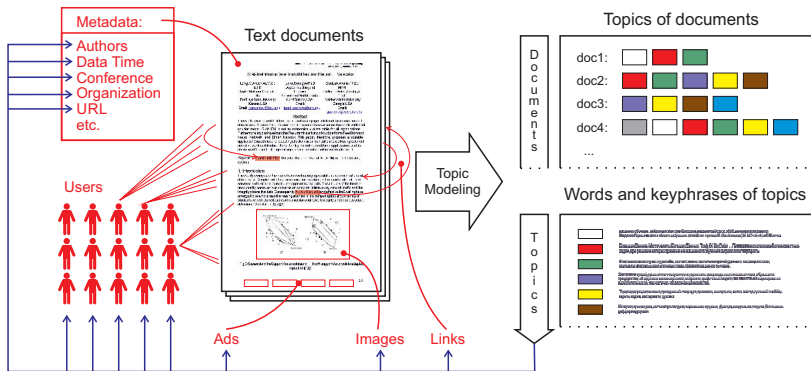
EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \sum_{i=1}^n \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in d} n_{dw} p_{tdw} + \sum_{i=1}^n \tau_i \theta_{td} \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

K. Vorontsov, A. Potapenko. Additive regularization of topic models. Machine Learning, 2015.

Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topic distributions of terms $p(w|t)$ and other modalities: $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{category}|t)$, $p(\text{tag}|t)$, $p(\text{link}|t)$, $p(\text{object-on-image}|t)$, $p(\text{user}|t)$, etc.



Multimodal extension of ARTM

W^m is a vocabulary of tokens of m -th modality, $m \in M$.

Maximum **multimodal** log-likelihood with regularization:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right. \end{cases}$$

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. 2015.

Why BigARTM simplifies topic modeling for applications

Stages	Bayesian Inference for PTMs	ARTM	
<i>Requirements analysis:</i>	Requirements analysis	Requirements analysis	
<i>Model formalization:</i>	Generative model design	predefined criteria	user-defined criteria
<i>Model inference:</i>	Bayesian inference for the generative model (VI, GS, EP)	One regularized EM-algorithm for any combination of criteria	
<i>Model implementation:</i>	Researchers coding (Matlab, Python, R)	Production code (C++)	
<i>Model evaluation:</i>	Researchers coding (Matlab, Python, R)	predefined measures	user-defined measures
<i>Deployment:</i>	Deployment	Deployment	

conventions: ::: not unified stages ::: ::: unified stages :::

Bayesian modeling requires maths and coding at each stage.

ARTM introduces the modular LEGO-style technology, packing each our requirement into a ready-to-use unified building block.

BigARTM: open source for fast modular topic modeling

BigARTM features:

- Parallel + online + multimodal + regularized Topic Modeling
- Out-of-core one-pass processing of large text collections
- Built-in library of regularizers and quality measures

BigARTM community:

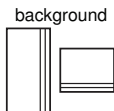
- Open-source <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Documentation <http://bigartm.org>



BigARTM license and programming environment:

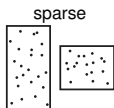
- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Regularizers for the interpretability of topics



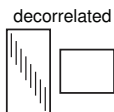
LDA: Smoothing background topics $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$



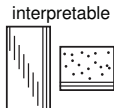
“Anti-LDA”: Sparsing subject domain topics $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \phi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$



Making topics as different as possible:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \phi_{wt} \phi_{ws}$$



Making topics more interpretable
by combining the above regularizers

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

hierarchy



Hierarchical links between topics t and subtopics s :

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}.$$

temporal



Topics dynamics over the modality of time intervals i :

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} |\phi_{it} - \phi_{i-1,t}|.$$

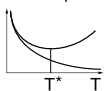
regression



Linear predictive model $\hat{y}_d = \langle v, \theta_d \rangle$ for documents:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2.$$

n of topics



Sparsing $p(t)$ for topic selection:

$$R(\Theta) = -\tau \sum_{t \in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_d p(d) \theta_{td}.$$

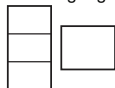
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories for text classification and categorization.

multilanguage



The modalities of languages with translation dictionary $\pi_{uwt} = p(u|w, t)$ for the $k \rightarrow \ell$ language pair:

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \phi_{wt}$$

graph



The modality of graph vertices v with doc sets D_v :

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} S_{uv} \sum_{t \in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|} \right)^2.$$

geospatial



The modality of geolocations g with proximity $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g, g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}} \right)^2$$

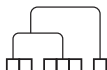
Beyond the “bag-of-words” (too restrictive) hypothesis

n-gram



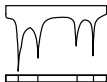
The modalities of n -grams, collocations, named entities

syntax



The modality of n -grams after SyntaxNet preprocessing

segmentation



E-step regularization affecting $p(t|d, w)$ distributions for segmentation and sentence topic models

coherence



Modeling co-occurrence data n_{uv} for biterms (u, v) :

$$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_t n_t \phi_{ut} \phi_{vt}$$

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

- 3.7M articles from Wikipedia, 100K unique words

	procs	$T = 50$		$T = 200$	
		time, m	perplexity	time, m	perplexity
BigARTM	1	42	5117	83	3347
BigARTM <i>async</i>	1	25	5131	53	3362
VowpalWabbit	1	50	5413	154	3960
Gensim	1	142	4945	637	3241
BigARTM	4	12	5216	26	3520
BigARTM <i>async</i>	4	7	5353	16	3634
Gensim	4	88	5311	315	3583
BigARTM	8	8	5648	15	3929
BigARTM <i>async</i>	8	5	6220	10	4309
Gensim	8	88	6344	288	4263

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Exploratory search in tech news

Goal: exploratory search by long text queries in digital libraries and tech news.



The bag-of-regularizers:

$$\mathcal{L}\left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array}\right) + R\left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart icon]} \quad \text{[Scatter plot icon]} \end{array}\right) + R\left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked boxes icon]} \quad \text{[Square icon]} \end{array}\right) + R\left(\begin{array}{c} \text{n-gram} \\ \text{[Grid of boxes icon]} \end{array}\right) \rightarrow \max$$

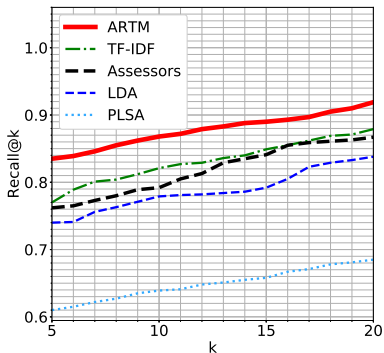
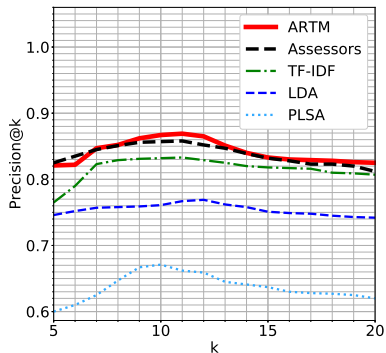
Results:

- Precision and Recall augmented +8% on Habrahabr.ru and TechCrunch.com tech news collections.
- Precision and Recall are comparable with assessors' quality.
- The topic-based search engine instantly performs the work that people typically complete in about 30 minutes.

A.Ianina, L.Golitsyn, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Precision and Recall: comparison against baselines

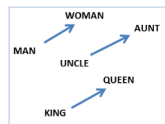
TechCrunch.com text collection, 760K documents
Precision and Recall at top k search result positions



A. Ianina, K. Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Sparse topically interpretable probabilistic word embeddings

Goal: build additively regularizable probabilistic word embeddings $p(t|w)$ with sparse interpretable topical coordinates and semantic properties similar to word2vec.



The bag-of-regularizers:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \left[\begin{array}{|c|} \hline \Phi \\ \hline \end{array} \right] \left[\begin{array}{|c|} \hline \Theta \\ \hline \end{array} \right] \end{array} \right) + R \left(\begin{array}{c} \text{co-occurrence} \\ \left[\begin{array}{|c|} \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \end{array} \right] \left[\begin{array}{|c|} \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \end{array} \right] \end{array} \right) + R \left(\begin{array}{c} \text{sparse} \\ \left[\begin{array}{|c|} \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \end{array} \right] \left[\begin{array}{|c|} \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \cdot \\ \hline \end{array} \right] \end{array} \right) + R \left(\begin{array}{c} \text{multimodal} \\ \left[\begin{array}{|c|} \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \end{array} \right] \left[\begin{array}{|c|} \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \text{---} \\ \hline \end{array} \right] \end{array} \right) \rightarrow \max$$

Results:

- Word vector coordinates are sparse and interpretable
- Performance on word similarity tasks is comparable
- Performance on document similarity tasks is better
- Modalities improve word similarity performance

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

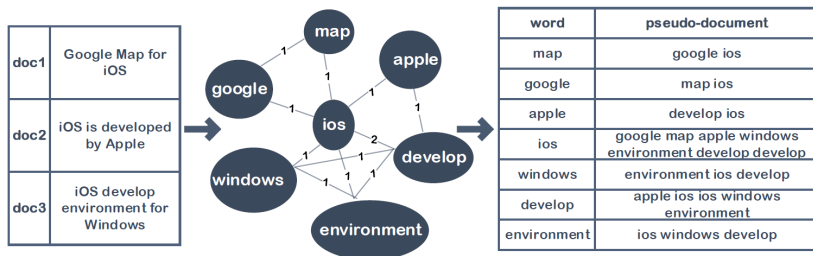
WNTM: Word Network Topic Model

Idea: let us model local contexts of words instead of documents.

d_u — pseudo-document that collects all contexts of the word u .

n_{uw} — the number of co-occurrences of words w and u .

The context is a short message / sentence / window of $\pm h$ words.



Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Examples of vector operations in word similarity tasks

Take the best of the two approaches:

- **ARTM**: sparse interpretable vector components
- **word2vec**: interpretable vector addition and subtraction

vector operation	ARTM result	word2vec result
king - boy + girl	<i>queen</i> , princess, lord, prince	<i>queen</i> , princess, regnant, kings
moscow - russia + spain	<i>madrid</i> , barcelona, aires, buenos	<i>madrid</i> , barcelona, valladolid, malaga
india - russia + ruble	<i>rupee</i> , birbhum, pradesh, madhaya	<i>rupee</i> , rupiah, devalued, debased
cars - car + computer	<i>computers</i> , software, servers, implementations	<i>computers</i> , software, hardware, microcomputers

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Quantitative estimation on word similarity tasks

Wikipedia 2016-01-13 dump, $|W| = 100K$, sparsity 93%.

Baselines: LDA, SVD-PPMI, SGNS (word2vec).

ARTM-based PWE (probabilistic word embeddings):

offline, online, online with sparsing.

	WordSim similarity	WordSim relatedness	WordSim joint	Bruni et al. MEN	Radinsky m.turk
LDA	0.530	0.455	0.474	0.583	0.483
SVD-PPMI	0.711	0.648	0.672	0.236	0.616
SGNS	0.752	0.632	0.666	0.745	0.661
ARTM off	0.701	0.615	0.647	0.707	0.613
ARTM on	0.718	0.673	0.685	0.669	0.639
ARTM on sp	0.728	0.672	0.680	0.675	0.635

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Qualitative estimation of the interpretability (some examples)

No interpretability for SGNS (word2vec):

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

Good interpretability for PWE (probabilistic word embeddings):

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

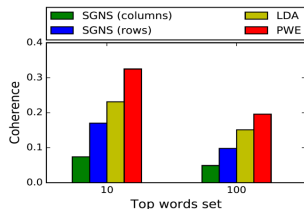
A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Quantitative estimation of the interpretability via coherence

Coherence shows how often the most probable words of the topic t (in the t -th embedding coordinate) co-occur in texts:

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

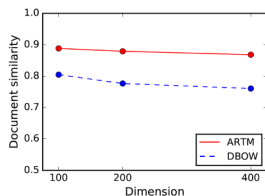
$\text{PMI}(u, v)$ is pointwise mutual information of word pair u, v .



- all models trained on Wikipedia
- two options for SGNS
- LDA is a baseline topic model
- ARTM-PWE outperforms both

Quantitative estimation on document similarity tasks

ArXiv triplets dataset of 20K triplets of papers:
(paper A, similar paper B, dissimilar paper C)



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

ARTM-PWE outperforms DBOW (distributed bag-of-words) model.

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

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Multimodal news corpus Lenta.ru

The screenshot shows a news article on the Lenta.ru website. The left sidebar contains a navigation menu with categories like 'Наука и техника', 'Спорт', and 'Интернет и СМИ'. The main content area has a navigation bar with 'Гаджеты' highlighted. The article title is 'Apple и Samsung будут вместе защищать смартфоны от воров', dated '16 апреля 2015'. Below the title is a photo of a woman and a man looking at a smartphone. The text below the photo discusses the Wireless Association's efforts to protect smartphones from theft. Annotations include a green box around the 'Наука и техника' menu item, a blue box around the date, a red box around the title, an orange box around the text, and a green box around the 'Гаджеты' menu item.

← category 1

← category 2

← time stamp

← title

← text

← author/source

Quantitative estimation of multimodal word embedding

- Trained on 100K news, tested on word similarity testsets
- We used the testsets translated to Russian:

<http://russe.nlpub.ru/downloads>

<http://www.leviants.com/ira.leviant/MultilingualVSMdata.html>

	WordSim similarity	WordSim relatedness	WordSim +RG+MC	SimLex
SGNS	0.630	0.530	0.567	0.24
ARTM-PWE	0.649	0.565	0.604	0.12
Multi-ARTM-PWE	0.682	0.580	0.611	0.14

- ARTM-PWE outperforms SGNS on most of the datasets
- Additional modalities improve similarities between words

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL-6, 2017.

Scenario analysis of call center records (on-going research)

Goals:

- determine typical scenarios of dialogues between operators and customers
- elaborate the quantitative measure of how well operator works
- provide online tips for help operator handle customer's objections



The bag-of-regularizers:

$$\begin{aligned}
 \mathcal{L} & \left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{seed words} \\ \hline \text{[Bar chart]} \quad \square \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{[Bar chart]} \quad \text{[Scatter plot]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{segmentation} \\ \hline \text{[Waveform]} \\ \hline \end{array} \right) \\
 & + R \left(\begin{array}{|c|} \hline \text{n-gram} \\ \hline \text{[Grid]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{syntax} \\ \hline \text{[Tree diagram]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{dialog} \\ \hline \text{[Stacked bars]} \\ \hline \end{array} \right) \rightarrow \max
 \end{aligned}$$

Result: the quality of the topical segmentation augmented from 40% for baselines to 75% for ARTM

ARTM — Additive Regularization for Topic Modeling:

- ARTM is a non-Bayesian regularization framework for PTM
- ARTM gives the easy way to formalize and combine PTMs
- ARTM makes it easier to understand and explain PTMs
- ARTM originates the modular “LEGO-style” PTM technology
- ARTM is implemented in the open source project BigARTM

PWE — Probabilistic (Topical) Word Embeddings:

- PWE are sparse and interpretable like PTMs
- PWE solve word similarity tasks like SGNS
- PWE solve document similarity tasks like SGNS
- PWE can use regularization and modalities like ARTM



ARTM and BigARTM references

- [1] *K.Vorontsov*. Additive regularization for topic models of text collections. Doklady Mathematics, 2014.
- [2] *K.Vorontsov, A.Potapenko*. Additive regularization of topic models. Machine Learning, 2015.
- [3] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina*. Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM, 2015.
- [4] *K.Vorontsov, A.Potapenko, A.Plavin*. Additive regularization of topic models for topic selection and sparse factorization. SLDS, 2015.
- [5] *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova*. BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST, 2015.
- [6] *O.Frei, M.Apishev*. Parallel non-blocking deterministic algorithm for online topic modeling. AIST, 2016.
- [7] *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov*. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.
- [8] *N.Chirkova, K.Vorontsov*. Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- [9] *A.Ianina, L.Golitsyn, K.Vorontsov*. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- [10] *A.Potapenko, A.Popov, K.Vorontsov*. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- [11] *D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov*. Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.