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

Anna Potapenko[1], Artem Popov[2], and Konstantin Vorontsov[3]

[1] National Research University Higher School of Economics
anna.a.potapenko@gmail.com
[2] Lomonosov Moscow State University
popov.artem.s@yandex.ru
[3] Moscow Institute of Physics and Technology
vokov@forecsys.ru

Abstract. We consider probabilistic topic models and more recent word embedding techniques from a perspective of learning hidden semantic representations. Inspired by a striking similarity of the two approaches, we merge them and learn probabilistic embeddings with online EM-algorithm on word co-occurrence data. The resulting embeddings perform on par with Skip-Gram Negative Sampling (SGNS) on word similarity tasks and benefit in the interpretability of the components. Next, we learn probabilistic document embeddings that outperform paragraph2vec on a document similarity task and require less memory and time for training. Finally, we employ multimodal Additive Regularization of Topic Models (ARTM) to obtain a high sparsity and learn embeddings for other modalities, such as timestamps and categories. We observe further improvement of word similarity performance and meaningful inter-modality similarities.

1 Introduction

Recent progress in deep natural language understanding prompted a variety of word embedding techniques that work remarkably well for capturing semantics. These techniques are usually considered as general neural networks that predict context words given an input word [3, 27, 17]. Although this perspective is convenient to generalize to more complex neural network architectures, e.g. skip-thought vectors [16], we believe that it is also important to establish connections between neural embeddings and more traditional models of distributional semantics. It gives theoretical insights about certain models and enables to use previous work as a grounding for further advances.

One of the first findings in this line of research is interpreting Skip-Gram Negative Sampling (SGNS, [27]) as an implicit matrix factorization of the shifted Pointwise Mutual Information (PMI) matrix [20]. It brings SGNS to the context of various vector space models (VSMs) developed during the last decades. Pantel and Turney [40] provide a thorough survey of VSMs dividing them into word-word, word-context and word-document categories based on the type of the co-occurrence matrix. According to the distributional hypothesis [12], similar words tend to occur in similar contexts; thus the rows of any of these matrices can be used for estimating word similarities [9].

Gentner [11] defines attributional similarity (e.g. *dog* and *wolf*) and relational similarity (e.g. *dog:bark* and *cat:meow*), which are referred to as similarity and analogy tasks in more recent papers. While Baroni et al. [26] argue that word embeddings inspired by neural networks significantly outperform more traditional count-based approaches for both tasks, Levy et al. [21] tune a shared set of hyperparameters and show that two paradigms give a comparable quality.

We follow this line of research and demonstrate how principle ideas of the modern word embedding techniques and probabilistic topic models can be mutually exchanged to take the best of the two worlds. So far, topic modeling has been widely applied to factorize word-document matrices and reveal hidden topics of document collections [15, 4]. In this paper we apply topic modeling to a word-word matrix to represent words by probabilistic topic distributions. Firstly, we discover a number of practical learning tricks to make the proposed model perform on par with SGNS on word similarity tasks. Secondly, we show that the obtained probabilistic word embeddings (PWE) inherit a number of benefits from topic modeling.

One such benefit is interpretability. Interpretability of each component as a coherent topic is vital for many downstream NLP tasks. To give an example, exploratory search aims not only to serve similar documents by short or long queries, but also to navigate a user through the results. If a model can explain why certain items are relevant to the query in terms of distinct topics, then these topics can be used to arrange the results by categories. Murphy et al. [30] motivated the importance of interpretability and sparsity from the cognitive plausibility perspective and introduced Non-Negative Sparse Embeddings (NNSE), which is a variation of Non-Negative Sparse Coding matrix factorization. State-of-the-art techniques, such as SGNS or GloVe [35] lack both sparsity and interpretability. To address this problem, more recent models [23, 39] extend SGNS and CBOW [27] respectively. However, they do that with explicit modifications of optimization procedure, such as project gradient for SGD. A benefit of topic modeling framework is that interpretability comes naturally with a probabilistic interpretation of parameters.

Furthermore, probabilistic word embeddings can be easily extended with Additive Regularization of Topic Models, ARTM [43]. This is a general framework to combine multiple requirements in one topic model. In this work we use ARTM to obtain sparsity and to learn embeddings for additional *modalities*, such as timestamps, authors, categories, etc. It enables us to investigate inter-modality similarities, because all the embeddings are in the same space. Interestingly, additional modalities also improve performance on word similarity task. Finally, we build probabilistic document embeddings and show that they outperform DBOW architecture of paragraph2vec [17] on a document similarity task. Thus, we get a powerful framework for learning probabilistic embeddings for various items and with various requirements. We train these models with online EM-algorithm similar to [14] in BigARTM open-source library [41].

Related work includes Word Network Topic Model (WNTM, [45]) and Biterm Topic Model (BTM, [44]) that use word co-occurrence data for analyzing short and imbalances texts. However, they do not consider their models as a way to learn word representations. There are also a number of papers on building hybrids of topic models and word embeddings. Gaussian LDA [8] imposes Gaussian priors for topics in a semantic vector space produced by word embeddings. The learning procedure is obtained via Bayesian inference, however a similar idea is implemented more straightforwardly in [38]. They use pre-built word vectors to perform clustering via Gaussian Mixture Model and apply the model to Twitter analysis. Pre-built word embeddings are also used in [33] to improve quality of topic models on small or inconsistent datasets. Another model, called Topical Word Embeddings (TWE, [22]) combines LDA and SGNS. It infers a topic for each word occurrence and learns different embeddings for the same word occurred under different topics. Unlike all these models, we do not combine the models as separate mechanisms, but highlight a striking similarity of optimization objectives and *merge* the models.

The rest of the paper is organized as follows. In section 2 we remind the basics of word embeddings and topic models. In sections 3 and 4 we discuss theoretic insights and introduce our generalized approach. In the experiments section we use 3 text datasets (Wikipedia, ArXiv, and Lenta.ru news corpus) to demonstrate high quality on word similarity and document similarity tasks, drastic improvement of interpretability and sparsity, and meaningful inter-modality similarities.

2 Related work

Definitions and notation. Here we introduce the notation that highlights a common nature of all methods and will be used throughout the paper. Consider a set of documents D with a vocabulary W. Let n_{wd} denote a number of times the word w occurs in the document d. The document can be treated as a *global context*. We will be also interested in a *local context* of each word occurrence, which is a bag of words in a window of a fixed size. Let n_{uv} denote a number of co-occurrences of words u and v in a sliding window, $n_u = \sum_v n_{uv}$, $n_v = \sum_u n_{uv}$, and $n = \sum_u n_u$.

All the models will be parametrized with the matrices $\varPhi$ and $\varTheta$, containing $|T|$-dimensional embeddings.

Skip-Gram model. Skip-gram model learns word embeddings by predicting a local context for each word in a corpus. The probability of word u from a local context of word v is modeled as follows:

$$p(u|v) = \frac{\exp \sum_t \phi_{ut}\theta_{tv}}{\sum_{w \in W} \exp \sum_t \phi_{wt}\theta_{tv}}, \tag{1}$$

where $\varPhi^{|W|\times|T|} = (\phi_{ut})$ and $\varTheta^{|T|\times|W|} = (\theta_{tv})$ are two real-valued matrices of parameters. According to the bag-of-words assumption, each word in the local context is modeled independently, thus one can derive the log-likelihood as follows:

$$\mathcal{L} = \sum_{v \in W} \sum_{u \in W} n_{uv} \ln p(u|v) \to \max_{\varPhi,\varTheta}. \tag{2}$$

where n_{uv} denotes the number of times the two terms co-occurred in a sliding window. However, normalization over the whole vocabulary in formula (1) prevents from learning the model effectively on large corpora. Skip-Gram Negative Sampling (SGNS) is one of possible ways to tackle this problem. Instead of modeling a conditional probability $p(u|v)$, SGNS models the probability of a co-occurrence for a pair of words (u, v). The model is trained on word pairs from the corpus (positive examples) as well as randomly sampled pairs (negative examples):

$$\sum_{v \in W} \sum_{u \in W} n_{uv} \log \sigma \left(\sum_t \phi_{ut} \theta_{tv} \right) \; + \; k \, \mathbb{E}_{\bar{v}} \log \sigma \left(- \sum_t \phi_{ut} \theta_{tv} \right) \;\; \rightarrow \;\; \max_{\Phi, \Theta}, \quad (3)$$

where σ is a sigmoid function, $\bar{v}$ are sampled from unigram distribution and k is a parameter to balance positive and negative examples. SGNS model can be effectively learned via Stochastic Gradient Descent.

SGNS model can be extended to learn document representations if the probabilities in (1) are conditioned on a document instead of a word. This architecture is called DBOW [7] and it is one of the modifications of the popular paragraph2vec approach.

Topic model. Probabilistic Latent Semantic Analysis, PLSA [15] is a topic model that describes words in documents by a mixture of hidden topics:

$$p(w|d) = \sum_{t \in T} p(w|t) p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}, \quad (4)$$

where $\Phi^{|W| \times |T|}$ contains probabilities ϕ_{wt} of words in topics and $\Theta^{|T| \times |D|}$ contains probabilities θ_{td} of topics in documents. The distributions are learned via maximization of the likelihood given normalization and non-negativity constraints:

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{wd} \log p(w|d) \rightarrow \max_{\Phi,\, \Theta} \quad (5)$$

$$\phi_{wt} \geq 0, \quad \sum_w \phi_{wt} = 1 \quad (6)$$

$$\theta_{td} \geq 0, \quad \sum_t \theta_{td} = 1. \quad (7)$$

This task can be effectively solved via EM-algorithm [9] or its online modification [14]. The most popular Latent Dirichlet Allocation [4] topic model extends PLSA by using Dirichlet priors for Φ and Θ distributions.

Additive Regularization of Topic Models, ARTM [43] is a non-Bayesian framework for learning multiobjective topic models. The optimization task (5) is extended with n additive regularizers $R_i(\Phi, \Theta)$ that are balanced with τ_i coefficients:

$$\mathcal{L} + R \rightarrow \max_{\Phi,\, \Theta}; \quad R = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \quad (8)$$

This approach addresses the problem of the non-uniqueness of the likelihood maximization (5) solution and imposes additional criteria to choose Φ and Θ. The optimization is still done with online EM-algorithm, where M-step is modified to use the derivatives of the regularization terms [43].

3 Probabilistic word embeddings

Consider a modification of PLSA to predict the word u in a local context of the word v:

$$p(u|v) = \sum_{t \in T} p(u|t)p(t|v) = \sum_{t \in T} \phi_{ut}\theta_{tv} \tag{9}$$

In this formulation the topic model approximates a word co-occurrence matrix instead of a word-document matrix. Unlike in PLSA, $\Theta^{|T| \times |W|}$ contains probabilities θ_{tv} of topics for *words*. However, from the topic modeling perspective, those words can be treated as *pseudo-documents*. One may think of a pseudo-document *derived by a word* v as a concatenation of all local contexts for all occurrences of the word v in the corpus. A local context is still defined as a fixed-size window, but this definition can be easily extended to use syntactic patterns, sentences, or any other structure.

Interestingly, this approach appears to be extremely similar to Skip-Gram model (1). Both models predict the same probabilities $p(u|v)$ and make use of the observed data by optimizing exactly the same likelihood (2). Both models are parametrized with matrices of hidden representations of words. The only difference is the space of the parameters: while Skip-Gram has no constraints, the topic model learns non-negative and normalized vectors that have a probabilistic interpretation. As a benefit, word probabilities can be predicted with a mixture model of the parameters with no need in explicit *softmax* normalization.

Learning probabilistic word embeddings (PWE) can be treated as a stochastic matrix factorization of probabilities $p(u|v)$ estimated from a corpus. This makes a perfect analogy with matrix factorization formulations of SGNS [19], GloVe, NNSE, and other similar techniques. GloVe uses a squared loss with a weighting function $f(n_{uv})$ that penalizes too frequent co-occurrences. Apart from two real-valued matrices of parameters, it introduces bias terms b_u and $\tilde{b}_v$. NNSE also uses a squared loss, but imposes additional constraints to obtain sparse non-negative embeddings ϕ_u and guarantees the limited $l2$-norm for Θ rows, which are called *dictionary* entries.

We summarize the connections between all mentioned models in Table 1. Each method is decomposed into several components: the type of raw co-occurrence data $F = (f_{uv})^{W \times W}$, the matrix factorization loss, the constraints for a parameter space, and the optimization technique. From this point of view, there is no big difference between so called *count-based* and *predictive* approaches. On the one hand, each method counts f_{uv} values (probably implicitly) and performs dimensionality reduction by a matrix factorization. On the other hand, each matrix factorization objective can be treated as a loss, which is used to train the model from data. More importantly, the unified view provides a powerful tool to analyze a diverse set of existing models and exchange components across them.

4 Additive regularization and embeddings for multiple modalities

The proposed probabilistic embeddings can be easily extended as a topic model. First, there is a natural way to learn document embeddings. Second, additive regularization of topic models [43] can be used to meet further requirements. In this paper we employ

Table 1. Learning word embeddings with a low-rank matrix factorization.

PWE	data type	$F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u \mid v)$
	objective	$\sum_{v \in W} n_v \operatorname{KL}\left(\hat{p}(u \mid v) \Vert \langle \phi_u \theta_v \rangle\right) \to \min\limits_{\Phi,\Theta}$
	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	$\sum_{v \in W} \sum_{u \in W} f(n_{uv}) \big(\langle \phi_u \theta_v \rangle + b_u + \tilde{b}_v - \log n_{uv} \big)^2 \to \min_{\Phi,\Theta,b,\tilde{b}}$
	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	$\sum_{u \in W} \left(\Vert f_u - \phi_u \Theta \Vert^2 + \Vert \phi_u \Vert_1 \right) \to \min_{\Phi,\Theta}$
	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]

it to obtain a high sparsity with no reduction in the accuracy of matrix factorization. The regularization criteria is a sum of cross-entropy terms between the target and fixed distributions:

$$R = -\tau \sum_{t \in T} \sum_{u \in W} \beta_u \ln \phi_{ut} \tag{10}$$

where β_u can be set to the uniform distribution.

Furthermore, we extend the topic model to incorporate meta-data or *modalites*, such as timestamps, categories, authors, etc. Real data often has such type of information associated with each document and it is desirable to build representations for these additional tokens as well as for the usual words.

Recall that each *pseudo-document* v in our training data is formed by collecting words u that co-occur with word v within a sliding window. Now we enrich it by the tokens u of some additional modality m that co-occur with the word v within a document. The only difference here is in using *global* document-based co-occurrences for additional modalities as opposed to *local* window-based co-occurrences for the modality of words. Once the *pseudo-documents* are prepared, we employ Multi-ARTM approach [42] to learn topic vectors for tokens of each modality:

$$\sum_{m \in M} \lambda_m \underbrace{\sum_{v \in W^0} \sum_{u \in W^m} n_{uv} \ln p(u|v)}_{\text{modality log-likelihood } \mathcal{L}_m(\Phi,\Theta)} \to \max_{\Phi,\Theta}, \tag{11}$$

$$\phi_{ut} \geq 0, \quad \sum_{u \in W^m} \phi_{ut} = 1, \, \forall m \in M; \tag{12}$$

$$\theta_{tv} \geq 0, \quad \sum_{t \in T} \theta_{tv} = 1. \tag{13}$$

Table 2. Spearman correlation for word similarities on Wikipedia.

Model	Data	Optimization	Metric	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	SimLex-999
LDA	n_{wd}	online EM	hel	0.530	0.455	0.474	0.583	0.220
PWE	n_{uv}	offline EM	dot	0.709	0.635	0.654	0.658	0.240
PWE	pPMI	offline EM	dot	0.701	0.615	0.647	0.707	0.276
PWE	n_{uv}	online EM	dot	0.718	**0.673**	**0.685**	0.669	0.263
SGNS	sPMI	SGD	cos	**0.752**	0.632	0.666	**0.745**	**0.384**

where $\lambda_m > 0$ are *modality weights*, W^m are modality vocabularies, and $m = 0$ for the basic text modality. Optionally, the tokens of other modalities can also form pseudo-documents and this would restore the symmetric property of the factorized matrix. Regularizers can be still added to the multimodal optimization criteria.

Online EM-algorithm. Regularized multimodal likelihood maximization is performed with online EM-algorithm implemented in BigARTM library [41]. First, we compute all necessary co-occurrences and build the *pseudo-documents* as described before. We store this corpus on disk and process it by batches of $B = 100$ pseudo-documents. The algorithm starts with random initialization of Φ and Θ matrices. The E-step estimates posterior topic distributions $p(t|u, v)$ for words u in a pseudo-document v. These updates are alternating with θ_{tv} updates for the given pseudo-document. After a fixed number of iterations through the pseudo-document, θ_{tv} are thrown away, while $p(t|u, v)$ are used to compute incremental unnormalized updates for ϕ_{ut}. These updates are applied altogether when the whole batch of pseudo-documents is processed. Importantly, these procedure does not overwrite the previous value of Φ, but slowly forgets it with an exponential moving average. The detailed formulas for the case of usual documents can be found in [41]. Note that the only matrix which has to be always stored in RAM is Φ. The number of epochs (runs through the whole corpus) in our experiments ranges from 1 to 6.

5 Experiments

We conduct experiments on three different datasets. Firstly, we compare the proposed Probabilistic Word Embeddings (PWE) to SGNS on Wikipedia dump by word similarities and interpretability of the components. Secondly, we learn probabilistic document embeddings on ArXiv papers and compare them to DBOW on the document similarity task [7]. Finally, we learn embeddings for multiple modalities on a corpus of Russian news Lenta.ru and investigate inter-modality similarities. All topic models are learnt in BigARTM[4] open source library [41] using Python interface[5]. SGNS is taken from Hyperwords[6] package and DBOW is taken from Gensim[7] library.

[4] bigartm.org

[5] github.com/bigartm/bigartm-book/blob/master/applications/word_embeddings.ipynb

[6] bitbucket.org/omerlevy/hyperwords

[7] radimrehurek.com/gensim/

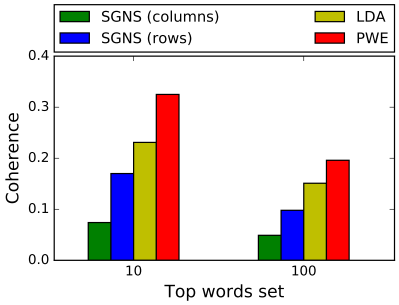

Fig. 1. Coherence scores.

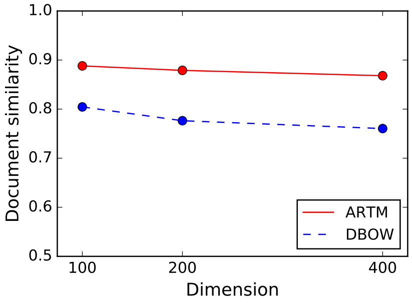

Fig. 2. Document similarities.

Word similarity tasks. We use Wikipedia 2016-01-13 dump and preprocess it with Levy's scripts[2] to guarantee equal conditions for SGNS and topic modeling [21]. We delete top 25 stop-words from the vocabulary, keep the next 100000 words, and delete the word pairs that co-occur less than 5 times. We performed experiments for *windows of size* 2*,* 5*, and* 10, but report here only window-5 results, as the others are analogous. We use *subsampling* with the constant 10^{-5} for all models. While common for SGNS, subsampling has never been used for topic modeling. However, our experiments show that it slightly improves topic interpretability by filtering out too general terms and therefore might be a good preprocessing recommendation. Also, we tried using *dynamic* window, which is a weighting technique based on the distance of the co-occurred words, but we didn't find it beneficial.

Following a traditional benchmark for word similarity tasks, we rank word pairs according to our models and measure Spearman correlation with the human ratings from WordSim353 dataset [10] partitioned into WordSim Similarity and WordSim Relatedness [1], MEN dataset [5], and SimLex-999 [13]. We consider SGNS model as a baseline and investigate if probabilistic word embeddings (PWE) are capable of providing the comparable quality. We start with LDA and Hellinger distance for word vectors as this is the default choice from many papers, e.g. [30]. Table 2 shows that SGNS dramatically outperforms LDA. Our further experiments demonstrate how to make topic models work.

First, we get an improvement by modeling the word-word matrix instead of the word-document matrix. Second, we investigate how to compute word similarity in the obtained space of probabilistic embeddings. We find the topic distributions should be normalized using Bayes' rule $p(t|u) = \frac{\phi_{ut}p(t)}{\sum_t \phi_{ut}p(t)}$ and that dot-product performs better than Hellinger distance or cosine similarity. Third, we find that online EM-algorithm with incremental Φ updates performs better than its offline analogue, where Φ is overwritten once per epoch. We also find that it is beneficial to initialize Θ randomly each time rather than store the values from the previous epoch. This combination of tricks gives the accuracy comparable to SGNS.

To obtain *sparsity*, we add the regularizer at the last iterations of EM-algorithm and observe **93%** of zeros in word embeddings *with the same* performance on word similarity tasks. We also try different co-occurrence scores instead of raw counts such

Table 3. Interpretability of topics.

PWE		SGNS	
art	arbitration	transports	rana
painting	ban	recon	walnut
museum	requests	grumman	rashid
painters	arbitrators	convoys	malek
gallery	noticeboard	piloted	aziz
sculpture	block	stealth	khalid
painter	administrators	flotilla	yemeni
exhibition	arbcom	convoy	andalusian
portraits	sanctions	supersonic	bien
drawings	mediation	bomber	gcc

Table 4. Event timestamps.

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

as $\log n_{uv}$ to penalize frequent co-occurrences or normalized $\frac{n_{uv}}{\sum_u n_{uv}}$ values to obtain a sum of *non-weighted* KL-divergences in the optimization criteria. While most of these weighting schemes give worse results, positive PMI values appear to be beneficial for some testsets.

Interpretability of embedding components. We characterize each component by a set of words with the highest values in the embedding matrix and check if those sets correspond to some aspects that can be named by a human. *Word intrusion* [6] technique is based on the idea that for well formed sets, a human expert can easily detect an intruder, randomly sampled from the vocabulary. This technique has been widely used in topic modeling and also for Non-Negative Sparse Embeddings [30] and Online Interpretable Word Embeddings [24]. Word intrusion requires experts, but it can be automated by the *coherence score*, which is shown to have high correlations with human judgements [32]. It averages pairwise similarities across the set of words. For similarities one can use PMI scores from an external corpus [31], log-conditional probabilities from the same corpus [29], distributional similarities [2], or other variants [36].

In our experiments we use the PMI-based coherence for top-10 and top-100 words for each component. The score is averaged over the components and reported in Figure 1. For SGNS we consider two different schemes of ranking words within each component. First, using the raw values; second, applying softmax *by rows* and using Bayes' rule to convert $p(t|w)$ into $p(w|t)$ probabilities. We show that the coherence for probabilistic word embeddings is consistently higher than that of LDA or SGNS for a range of embedding sizes. Also, this result is confirmed by visual analysis of the obtained components (see Table 3 for the examples).

Document similarity task. In this experiment we learn probabilistic document embeddings on ArXiv corpus and test them on a document similarity task. The testset released by Dai et. al [7] contains automatically generated triplets of a query paper, a similar paper that shares key words, and a dis-similar paper that does not share any key words. The quality is evaluated by the accuracy of identifying the similar one within each triplet.

Table 5. Spearman correlation for word similarities on Lenta.ru.

Model	WordSim Sim	WordSim Rel	MC	RG	HJ	SimLex
SGNS	0.630	0.530	0.377	0.415	0.567	**0.243**
CBOW	0.625	0.513	0.403	0.370	0.551	0.170
PWE	0.649	0.565	0.605	**0.594**	0.604	0.123
Multi-PWE	**0.682**	**0.58**	**0.607**	0.584	**0.611**	0.144

We preprocess[8] plain texts of 963564 ArXiv papers with a total of 1416554733 tokens and reduce the vocabulary size to 122596 words with a frequency-based filtering. The restored mapping between the plain texts and the URLs from the testset[9] covers 15853 triplets out of 20000.

We train embeddings with 1 epoch of online EM-algorithm. Note that the matrix Θ is not stored, so memory consumption does not grow linearly with the number of documents. Afterwards, we infer test embeddings with 10 passes on each document. As a baseline, we train DBOW [7] with 15 epochs and use linear decay of learning rate from 0.025 to 0.001; afterwards we infer test embeddings with 5 epochs. Unlike online EM-algorithm, DBOW needs in-memory storage of document vectors and also takes much longer to train (several hours instead of 30 minutes on the same machine). We do not facilitate training word vectors in DBOW, because it slows down the process dramatically.

Figure 2 shows that our ARTM model consistently outperforms DBOW for a range of embedding sizes. The absolute numbers are also better than for all other methods reported in [7], thus giving a new state-of-the-art on this dataset.

Multimodal embedding similarities. The experiments are held on Russian *lenta.ru* corpus, that contains 100033 news with a total of 10050714 tokens. The corpus has additional modalities of timestamps (825 unique tokens), categories (22 unique tokens) and sub-categories (97 unique tokens). The basic text modality has 54963 unique words.

We produce a collection of pseudo-documents using the window of size 5 and subsampling. For evaluation we use HJ testset [34] with human judgments on 398 word pairs translated to Russian from the widely used English testsets: MC [28], RG [37], and WordSim353 [10]. We also use SimLex-999 testset translation [18].

Table 5 shows that probabilistic word embeddings (PWE) outperform SGNS for most of the testsets even without using additional modalities. One can note that this corpus is relatively small and it might be a reason for poor SGNS performance. We have also tried CBOW [27] following a common recommendation to use it for small data, but it performed even worse. Generally, we observe that topic modeling requires less data for a good performance, thus the proposed PWE approach might be beneficial for applications with limited data.

Next, we use additional modalities and optimize the modality weights in the objective (11). With this approach we observe a further boost in the performance for the word similarity task (see Multi-PWE in Table 5). Finally, we experiment with two different

[8] https://github.com/romovpa/arxiv-dataset

[9] http://cs.stanford.edu/quocle/triplets-data.tar.gz

modes: using modalities only as tokens (a non-symmetric case) and both as tokens and pseudo-documents (a symmetric case). While word similarities are better for the non-symmetric case, we observe better inter-modality similarities for the symmetric case. Table 4 provides several examples of remarkable timestamps and their closest words. The words are manually translated from Russian to English for reporting purposes only. Each column is easily interpretable as a coherent event, namely the release of Star Wars, the Oscars 2016, and Victory Day in Russia.

6 Conclusions

In this work we revisited topic modelling techniques in the context of learning hidden representations for words and documents. Topic models are known to provide interpretable components but perform poorly on word similarity tasks. However, we have shown that topic models and neural word embeddings can be made to predict the same probabilities with the only difference in the probabilistic nature of parameters. This theoretical insight enabled us to merge the models and get practical results. First, we obtained probabilistic word embeddings (PWE) that work on par with SGNS on word similarity tasks, but have high sparsity and interpretability of the components. Second, we learned document embeddings that outperform DBOW on a document similarity task and require less memory and time for training. Furthermore, considering the task as a topic modeling, enabled us to adapt Multi-ARTM approach and learn embeddings for multiple modalities, such as timestamps and categories. We observed meaningful inter-modality similarities and a boost of the quality on the basic word similarity task. In future we plan to apply the proposed probabilistic embeddings to a suite of NLP tasks and take even more advantage of the additive regularization to incorporate task-specific requirements into the models.

Acknowledgements. The work was supported by Government of the Russian Federation (agreement 05.Y09.21.0018) and the Russian Foundation for Basic Research grants 17-07-01536, 16-37-00498.

References

1. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A.: A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 19–27. NAACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
2. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: IWCS (2013)
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. 3, 1137–1155 (Mar 2003)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)

5. Bruni, E., Boleda, G., Baroni, M., Tran, N.K.: Distributional semantics in technicolor. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. pp. 136–145. ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
6. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Neural Information Processing Systems (2009)
7. Dai, A.M., Olah, C., Le, Q.V.: Document embedding with paragraph vectors. CoRR abs/1507.07998 (2015)
8. Das, R., Zaheer, M., Dyer, C.: Gaussian LDA for Topic Models with Word Embeddings. In: ACL (1). pp. 795–804. The Association for Computer Linguistics (2015)
9. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 pp. 391–407 (1990)
10. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. ACM Trans. Inf. Syst. 20(1), 116–131 (Jan 2002)
11. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive Science 7(2), 155–170 (1983)
12. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
13. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with genuine similarity estimation. Comput. Linguist. 41(4), 665–695 (Dec 2015)
14. Hoffman, M.D., Blei, D.M., Bach, F.R.: Online learning for latent dirichlet allocation. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) NIPS. pp. 856–864. Curran Associates, Inc. (2010)
15. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. pp. 289–296. UAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999)
16. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-thought vectors. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. pp. 3294–3302. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
17. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR abs/1405.4053 (2014)
18. Leviant, I., Reichart, R.: Judgment language matters: Towards judgment language informed vector space modeling. In: Preprint pubslished on arXiv (arxiv:1508.00106) (2015)
19. Levy, O., Goldberg, Y.: Linguistic Regularities in Sparse and Explicit Word Representations. In: Morante, R., tau Yih, W. (eds.) CoNLL. pp. 171–180. ACL (2014)
20. Levy, O., Goldberg, Y.: Neural Word Embedding as Implicit Matrix Factorization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2177–2185. Curran Associates, Inc. (2014)
21. Levy, O., Goldberg, Y., Dagan, I.: Improving Distributional Similarity with Lessons Learned from Word Embeddings. TACL 3, 211–225 (2015)
22. Liu, Y., Liu, Z., Chua, T.S., Sun, M.: Topical Word Embeddings. In: AAAI. pp. 2418–2424 (2015)
23. Luo, H., Liu, Z., Luan, H.B., Sun, M.: Online Learning of Interpretable Word Embeddings. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) EMNLP. pp. 1687–1692. The Association for Computational Linguistics (2015)
24. Luo, H., Liu, Z., Luan, H.B., Sun, M.: Online learning of interpretable word embeddings. In: EMNLP (2015)
25. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research 11, 19–60 (2010)

26. Marco Baroni, Georgiana Dinu, G.K.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference 1, 238–247 (2014)
27. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) NIPS. pp. 3111–3119 (2013)
28. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
29. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 262–272. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
30. Murphy, B., Talukdar, P.P., Mitchell, T.M.: Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In: Kay, M., Boitet, C. (eds.) COLING. pp. 1933–1950. Indian Institute of Technology Bombay (2012)
31. Newman, D., Bonilla, E.V., Buntine, W.L.: Improving topic coherence with regularized topic models. In: NIPS (2011)
32. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
33. Nguyen, Q.D., Billingsley, R., Du, L., Johnson, M.: Improving Topic Models with Latent Feature Word Representations. Transactions of the Association of Computational Linguistics 3, 299–313 (2015)
34. Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., Biemann, C.: Human and machine judgements for russian semantic relatedness. In: Analysis of Images, Social Networks and Texts (AIST'2016). Springer (2016)
35. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
36. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 399–408. WSDM '15, ACM, New York, NY, USA (2015)
37. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10), 627–633 (Oct 1965)
38. Sridhar, V.K.R.: Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words. In: Blunsom, P., Cohen, S.B., Dhillon, P.S., Liang, P. (eds.) VS@HLT-NAACL. pp. 192–200. The Association for Computational Linguistics (2015)
39. Sun, F., Guo, J., Lan, Y., Xu, J., Cheng, X.: Sparse Word Embeddings Using 1 Regularized Online Learning. In: Kambhampati, S. (ed.) IJCAI. pp. 2915–2921. IJCAI/AAAI Press (2016)
40. Turney, P.D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, (2010), 37, 141-188 (Mar 2010)
41. Vorontsov, K., Frei, O., Apishev, M., Romov, P., Dudarenko, M.: Bigartm: Open source library for regularized multimodal topic modeling of large collections. In: AIST (2015)
42. Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., Yanina, A.: Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. In: Aletras, N., Lau, J.H., Baldwin, T., Stevenson, M. (eds.) TM@CIKM. pp. 29–37. ACM (2015)
43. Vorontsov, K., Potapenko, A.: Additive regularization of topic models. Machine Learning 101(1), 303–323 (2015)

44. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Schwabe, D., Almeida, V.A.F., Glaser, H., Baeza-Yates, R.A., Moon, S.B. (eds.) WWW. pp. 1445–1456. International World Wide Web Conferences Steering Committee / ACM (2013)
45. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowl. Inf. Syst. 48(2), 379–398 (2016)