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
arti.lehtonen@gmail.com
[3] Moscow Institute of Physics and Technology
vokov@forecsys.ru

Abstract. We consider probabilistic topic models and more recent word embedding techniques from one 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 SGNS on the word similarity task and benefit in interpretability of the components. Next, we learn probabilistic document embeddings that outperform DBOW on the document similarity task and require less memory and time for training. Finally, we employ Multi-ARTM approach to obtain 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, 24, 16]. Although this perspective is convenient to generalize to more complex neural network architectures, e.g. skip-thought vectors [15], we believe that it is also important to establish connections between neural embeddings and more traditional models of distributional semantics. It gives theoretical insights on why certain models work and enables to use previous work as a grounding for further advances.

One of the first findings on this line of research is interpreting Skip-Gram Negative Sampling (SGNS, [24]) as an implicit matrix factorization of shifted Pointwise Mutual Information (PMI) matrix [18]. It brings SGNS to the context of various vector space models (VSMs) developed during the last decades. Pantel and Turney [38] provide a thorough survey of VSMs dividing them to 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 serve 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 as similarity and analogy tasks in

more recent papers. While Baroni et al. [23] argue that word embeddings inspired by neural networks significantly outperform more traditional count-based approaches for both tasks, Levy et al. [19] reveal and 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 probabilistic topic models and neural word embeddings 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 [14, 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 for a word similarity task. Secondly, we show that the obtained embeddings inherit the benefits of topic models, namely interpretability and sparseness.

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. Murphy et al. [27] motivate the importance of interpretability and sparseness from the cognitive plausibility perspective and introduce Non-Negative Sparse Embedding (NNSE), which is a variation of Non-Negative Sparse Coding matrix factorization. Popular word embedding techniques, such as PMI-SVD [19], GloVe [32], or SGNS, lack both interpretability and sparsity. NNSE is the first attempt to combine the desired properties, but the research is still ongoing, e.g. two recent approaches [21, 37] extend SGNS and CBOW [24] models respectively.

While interpretability comes with the probabilistic nature of our embeddings, the sparsity requirement is explicitly imposed by Additive Regularization of Topic Models, ARTM [41]. This is a general framework to build extensible multi-objective topic models. We also use Multi-ARTM [40] approach to learn embeddings for multiple *modalities*, such as timestamps, authors, categories, etc. It improves word similarity performance and enables to investigate inter-modality similarities, since all the embeddings are in the same space. Finally, we build probabilistic document embeddings in the same framework and show that they outperform DBOW architecrute of paragraph2vec [16] on document similarity task. Thus, we get a powerful tool for learning probabilistic embeddings for various items and with various requirements.

It is important to note that we do not need to store the word-word matrix in memory, even though topic modeling is a matrix factorization technique. We concatenate the contexts of each word across the collection and treat the resulting pseudo-documents as a new corpus. Then we employ online EM-algorithm [13] to process it by batches and update the parameters of the model.

Related work includes Word Network Topic Model (WNTM, [43]) and Biterm Topic Model (BTM, [42]) that use word co-occurrence data for analyzing short and imbalances texts, but not for learning 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 em-

beddings. The learning procedure is obtained via Bayesian inference, however a similar idea is implemented more straightforwardly in [36]. 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 [30] to improve quality of topic models on small or inconsistent datasets. Another model, called Topical Word Embeddings (TWE, [20]) combines LDA and SGNS to learn different embeddings for a word occurred under different topics. Unlike all these models, we do not combine the models as 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

Skip-Gram model. Skip-gram model learns word embeddings by predicting local contexts for each word in a corpus. The probability of word u given its neighbouring word v is modeled as:

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

According to the bag-of-words assumption each word in a context is modeled independently and the log-likelihood of the corpus is as follows:

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

where n_{uv} denotes the number of times the two terms co-occurred in a local context, matrix Φ contains word embeddings, and matrix Θ contains context embeddings.

Calculating *softmax* function for each word prevents from learning this model effectively on large corpora. Skip-Gram Negative Sampling (SGNS) is one of possible ways to tackle this problem. Instead of modeling conditional probabilities $p(u|v)$ normalized over the whole vocabulary, 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):

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

where σ is a sigmoid function, $\bar{v}$ are randomly 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 documents representations if the probabilities in (1) are conditioned on a document instead of a word. This architecture is called DBOW [7] and is one of the modifications of the popular paragraph2vec approach.

Topic model. Probabilistic Latent Semantic Analysis, PLSA [14] is a topic model that describes words in documents by a mixture of hidden topics:

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

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

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

$$\sum_{w} \phi_{wt} = 1; \quad \sum_{t} \theta_{td} = 1, \tag{6}$$

where n_{wd} denotes the number of times word w occurs in document d. This task can be effectively solved via EM-algorithm [9] or its online modification [13]. 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 [41] 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:

$$\mathcal{L} + R \to \max_{\Phi, \Theta}; \quad R = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \tag{7}$$

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

3 Probabilistic word embeddings

Consider a modification of PLSA to predict word u given a co-occurred word v:

$$p(u|v) = \sum_{t \in T} p(u|t)p(t|v) = \sum_{t \in T} \phi_{ut}\theta_{tv} \tag{8}$$

Throughout this paper the co-occurrences are obtained by a sliding fixed-size window, but this notion can be easily extended to use syntactic patterns or any other structure information. In this formulation the topic model approximates a word co-occurrence matrix instead of a word-document matrix; and $\Theta^{T \times W}$ contains probabilities of topics for *words*. However, in this case the words are treated as *contexts* or *pseudo-documents*. One may think of a pseudo-document *derived by word* v that concatenates all fixed-size windows for all occurrences of word v in the corpus. Then the model is obtained by training PLSA on these W pseudo-documents.

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

PWE	data type	$F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u\|v)$
	criteria	$\sum_{v \in W} n_v \operatorname{KL}\left(\hat{p}(u\|v) \middle\| \langle \phi_u \theta_v \rangle\right) \to \min\limits_{\Phi,\Theta}$
	constrains	$\phi_{ut} > 0, \quad \sum_u \phi_{ut} = 1; \quad \theta_{tv} > 0, \quad \sum_t \theta_{tv} = 1$
	technique	Online EM-algorithm
SGNS	data type	$F_{uv} = \log \frac{n_{uv} n}{n_u n_v} - \log k$ [17]
	criteria	$\sum_{u \in W} \sum_{v \in W} n_{uv} \log \sigma\left(\langle \phi_u \theta_v \rangle\right) + k\, \mathbb{E}_{\bar{v}} \log \sigma\left(-\langle \phi_u \theta_v \rangle\right) \to \max_{\Phi,\Theta}$
	constrains	No constraints
	technique	SGD (online by corpus)
GloVe	data type	$F_{uv} = \log n_{uv}$
	criteria	$\sum_{v \in W} \sum_{u \in W} f(n_{uv}) \left(\langle \phi_u \theta_v \rangle + b_u + \tilde{b}_v - \log n_{uv}\right)^2 \to \min_{\Phi,\Theta,b,\tilde{b}}$
	constrains	No constraints
	technique	AdaGrad (online by the elements)
NNSE	data type	$F_{uv} = max(0, \log \frac{n_{uv} n}{n_u n_v})$ or SVD low-rank approximation
	criteria	$\sum_{u \in W} \left(\left\|f_u - \phi_u \Theta\right\|^2 + \left\|\phi_u\right\|_1\right) \to \min_{\Phi,\Theta}$
	constrains	$\theta_u \theta_u^T \leq 1, \forall u \in W \quad \phi_{uv} \geq 0, \forall u \in W, v \in W$
	technique	Online algorithm from (Mairal et al., 2010)

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 for words and contexts. The only difference is in the space of the parameters: while Skip-Gram has no constraints, the topic model learns non-negative and normalized vectors that have 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 corpus. This makes a perfect analogy with matrix factorization formulations of SGNS, GloVe, NNSE, and other similar techniques. We summarize the connections between them in Table 1. Each method is decomposed into several steps: word co-occurrence type $F = (F_{uv})^{W \times W}$, a loss for matrix factorization, constraints for a parameter space, and an optimization technique. From this point of view, there is no difference between count-based and predictive approaches. More importantly, the unified view provides a powerful tool to analyze a diverse set of existing model variants, and enables to exchange the steps 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 [41] can be used to meet further requirements. In this paper we employ it to obtain high sparsity with no loss of matrix factorization accuracy. The regularization

Table 2. Word similarities on Wikipedia.

Model	Data	Optimization	Metric	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	Radinsky M. Turk
LDA	n_{wd}	online EM	hel	0.530	0.455	0.474	0.583	0.483
PWE	n_{uv}	offline EM	dot	0.71	0.62	0.65	0.67	0.59
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
SGNS	sPMI	SGD	cos	**0.752**	0.632	0.666	**0.745**	**0.661**

criteria is a sum of KL-divergences between the target and fixed distributions:

$$R = -\tau \sum_{t \in T} \sum_{u \in W} \beta_u \ln \phi_{ut} \tag{9}$$

where β_u can be set to uniform distribution.

Furthermore, we extend the model for multiple *modalites*, such as timestamps, categories, authors, etc. 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 tokens u of some additional modality m that co-occur with word v *within a document*. Once the *pseudo-documents* are prepared, we employ Multi-ARTM approach [40] to learn topic vectors for tokens of each modality:

$$\sum_{m \in M} \lambda_m \underbrace{\sum_{v \in W^0} \sum_{u \in W^m} n_{uv} \ln p(u|v)}_{\text{modality log-likelihood } \mathcal{L}_m(\Phi,\Theta)} \to \max_{\Phi,\Theta}, \tag{10}$$

$$\phi_{ut} \geq 0, \quad \sum_{u \in W^m} \phi_{ut} = 1, \forall m \in M; \tag{11}$$

$$\theta_{tv} \geq 0, \quad \sum_{t \in T} \theta_{tv} = 1. \tag{12}$$

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 easily added to the multimodal optimization criteria.

Regularized multimodal likelihood maximization is performed with online EM-algorithm, that does not require to store the word-word matrix. *Pseudo-documents* are processed by batches; Φ matrix is stored and updated with exponential moving average; Θ matrix can is randomly initialized for each epoch and re-iterated on each document until convergence.

5 Experiments

We conduct experiments on three different datasets. Firstly, we compare our models 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 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 [39]. SGNS is taken from Hyperwords[5] package and DBOW is taken from Gensim[6] library.

Word similarity task. We use Wikipedia 2016-01-13 dump and preprocess it with Levy's scripts[2] to guarantee equal conditions for SGNS and topic modeling [19]. 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. 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 much beneficial.

Following a traditional benchmark for word similarity task, 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 Mechanical Turk dataset [33]. 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 Helinger distance for word vectors as this is the default choice from many papers, e.g. [27]. Table 2 shows that SGNS significantly outperforms LDA. Our further experiments demonstrate how to make topic models work.

First, using word-word instead of word-document matrix leads to a significant improvement. Second, inner-product of $p(t|w)$ gives better results than Helinger distance or cosine similarity. Next, we try substituting Θ with Φ, transposed and normalized by Bayes' rule along the topics dimension, which is equivalent to learning BTM [42]. Interestingly, this model gives exactly the same quality in our experiments, while it has a factor of two fewer parameters. Initializing Θ randomly per each epoch in online fashion gives the best results that are on par with SGNS. We also try different co-occurrence scores instead of raw counts such as $\log n_{uv}$ or normalized $\frac{n_{uv}}{\sum_u n_{uv}}$ values to obtain a sum of *non-weighted* KL-divergences in optimization criteria. While most of these schemes give worse results, positive PMI values appear to be beneficial for some testsets. To obtain sparsity, we add the regularizer at the last iterations of EM-algorithm and observe **93%** of zeros in word embeddings *with no loss* on word similarity task.

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

[4] bigartm.org

[5] bitbucket.org/omerlevy/hyperwords

[6] radimrehurek.com/gensim/

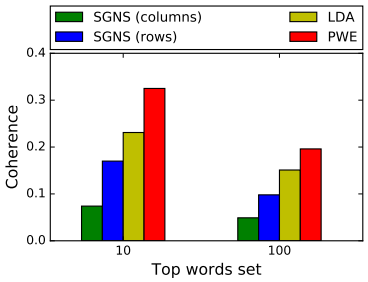

Fig. 1. Coherence scores.

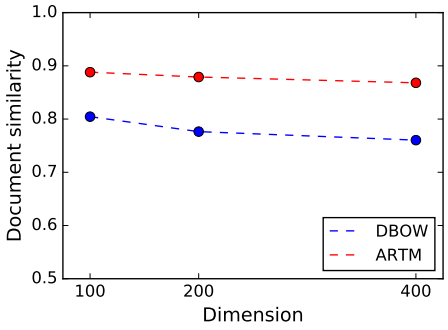

Fig. 2. Document similarities.

Table 3. Word similarities on Lenta.ru.

Model	WordSim Sim	WordSim Rel	MC	RG	ALL
SGNS	0.630	0.530	0.377	0.415	0.567
PWE (cos / dot)	0.612 / 0.649	0.540 / 0.565	0.648 / 0.605	**0.63** / 0.594	0.583 / 0.604
Multi-PWE (cos / dot)	0.646 / **0.682**	0.550 / **0.58**	**0.675** / 0.607	0.617 / 0.584	0.583 / **0.611**

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 [27] and Online Interpretable Word Embeddings [22]. Word intrusion requires experts, but it can be automatized by *coherence score*, which is shown to have high correlations with human judgements [29]. It averages pairwise similarities across the set of words. For similarities one can use PMI scores from an external corpus [28], log-conditional probabilities from the same corpus [26], distributional similarities [2], or other variants [34].

In our experiments we use PMI-based coherence score 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 of probabilistic word embeddings is significantly higher than that of LDA or any SGNS option. Also, this result is confirmed by visual analysis of the obtained components.

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

Plain texts of 963564 ArXiv papers are preprocessed[7] and the vocabulary is further reduced to 122596 words by frequency-based filtering. Restored mapping between the

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

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

Table 5. Coherence.

PWE		SGNS	
topic examples		topic examples	
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

plain texts and the urls from the testset[8] covers 15853 triplets out of 20000. We train embeddings with 1 epoch of online EM-algorithm in BigARTM. Note that Θ matrix 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 ARTM consistently outperforms DBOW for several dimensions. The absolute numbers are also better than for all other models reported in [7], thus giving a new state-of-the-art on this dataset.

Multimodal embedding similarities. Experiments are held on Russian *lenta.ru* corpus, that contains 100033 news and 54693 words in the vocabulary. The text of each document is available along with additional modalities of a timestamp, a category and a subcategoty. We produce a collection of pseudo-documents using the window of size 5 and subsampling. For word similarity benchmark we use HJ dataset [31] with human judgments on 398 word pairs translated to Russian from the widely used English datasets: MC [25], RG [35], and WordSim353 [10]. Table 3 shows that probabilistic word embeddings outperform SGNS for all testsets. Interestingly, there is no consistency in weather cosine similarity or inner product gives better results across the testsets.

The usage of additional modalities significantly boosts the performance. We optimize modality weights and experiment with two different 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 tokens-only mode, we observe better inter-modality similarities for the other mode. 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

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

as a coherent event, namely the release of Star Wars, the Oscars 2016, and Victory Day in Russia.

One can note that this corpus is relatively small and it might be a reason for poor SGNS performance. We have also tried CBOW [24] following a common recommendation to use it for small data, but it performed even worse. Generally, we observe that topic modeling generally requires less data for good performance, thus the proposed PWE approach might be beneficial for applications with limited data.

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 task. However, we showed that WNTM and PLSA predict the same probabilities as SGNS and DBOW respectively with the only difference of a probabilistic nature of the parameter. This theoretical insight enabled us to merge the models and get practical results. Firstly, we obtained probabilistic word embeddings (PWE) that work on par with SGNS on word similarity task, but have high sparsity and interpretability of components. Second, we learned document embeddings that outperform DBOW on document similarity task and are require less memory and time for learning. 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 significant boost of the quality on the basic word similarity task.

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. 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)
14. 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)
15. 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)
16. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR abs/1405.4053 (2014)
17. 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)
18. 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)
19. Levy, O., Goldberg, Y., Dagan, I.: Improving Distributional Similarity with Lessons Learned from Word Embeddings. TACL 3, 211–225 (2015)
20. Liu, Y., Liu, Z., Chua, T.S., Sun, M.: Topical Word Embeddings. In: AAAI. pp. 2418–2424 (2015)
21. 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)
22. Luo, H., Liu, Z., Luan, H.B., Sun, M.: Online learning of interpretable word embeddings. In: EMNLP (2015)
23. 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)
24. 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)
25. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
26. 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)
27. 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)

28. Newman, D., Bonilla, E.V., Buntine, W.L.: Improving topic coherence with regularized topic models. In: NIPS (2011)
29. 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)
30. 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)
31. 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)
32. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
33. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time: Computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th International World Wide Web Conference. pp. 337–346. Hyderabad, India (March 2011)
34. 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)
35. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10), 627–633 (Oct 1965)
36. 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)
37. 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)
38. 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)
39. 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)
40. 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)
41. Vorontsov, K., Potapenko, A.: Additive regularization of topic models. Machine Learning 101(1), 303–323 (2015)
42. 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)
43. 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)