Improving Language Estimation with the Paragraph Vector Model for Ad-hoc Retrieval

Qingyao Ai[1], Liu Yang[1], Jiafeng Guo[2], W. Bruce Croft[1]
[1]College of Information and Computer Sciences,
University of Massachusetts Amherst, Amherst, MA, USA
{aiqy, lyang, croft}@cs.umass.edu
[2]CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology,
Chinese Academy of Sciences, China
guojiafeng@ict.ac.cn

ABSTRACT

Incorporating topic level estimation into language models has been shown to be beneficial for information retrieval (IR) models such as cluster-based retrieval and LDA-based document representation. Neural embedding models, such as paragraph vector (PV) models, on the other hand have shown their effectiveness and efficiency in learning semantic representations of documents and words in multiple Natural Language Processing (NLP) tasks. However, their effectiveness in information retrieval is mostly unknown. In this paper, we study how to effectively use the PV model to improve ad-hoc retrieval. We propose three major improvements over the original PV model to adapt it for the IR scenario: (1) we use a document frequency-based rather than the corpus frequency-based negative sampling strategy so that the importance of frequent words will not be suppressed excessively; (2) we introduce regularization over the document representation to prevent the model overfitting short documents along with the learning iterations; and (3) we employ a joint learning objective which considers both the document-word and word-context associations to produce better word probability estimation. By incorporating this enhanced PV model into the language modeling framework, we show that it can significantly outperform the state-of-the-art topic enhanced language models.

Keywords

Retrieval Model; Language Model; Paragraph Vector

1. INTRODUCTION

Language models have been successfully applied to IR tasks[8, 14]. The core of this approach is to estimate a language model for each document and rank documents according to the likelihood of observing a query given the estimated model. The simple language model approach represents documents and queries under the bag-of-words assumption. This approach fails when query words are not observed in a document. A typical solution to this issue is to apply smoothing techniques by incorporating a corpus language model for "unseen" words, such as the Jelinek-Mercer method, absolute discounting, and Bayesian smoothing using Dirichlet priors [14]. However, smoothing every document with the *same* corpus language model is intuitively not optimal since we essentially assume that all the unseen words in different documents would have similar probabilities [13].

One way to improve the smoothing techniques is to introduce document dependent smoothing that can reflect the content of the document, for example by representing documents and queries in a latent topic space and estimating the generation probability accordingly. By incorporating topic level estimation into language model approaches, previous work such as the cluster-based retrieval model [6] and the LDA-based retrieval model [12] obtained consistent improvements over the basic language models. Nonetheless, the existing topic model based approaches have several drawbacks. Firstly, the model estimation relies on the predefined number of topics. Secondly, the topic models typically assign high probabilities to frequent words. Finally, the learning cost (of the LDA model) is expensive on a large corpus.

Recent advances in Natural Language Processing (NLP) have shown that semantically meaningful representations of words and documents can be efficiently acquired by neural embedding models. In particular, a paragraph vector (PV) model [4] has been proposed to jointly learn word and document embeddings by directly optimizing the generative probabilities of each word given the document. In contrast to existing topic models, PV can automatically cluster topic related words and documents without explicitly defining the number of topics a priori. The negative sampling based optimization strategy makes PV assign high probabilities to discriminative words rather than frequent words. Moreover, the online learning algorithm enables PV to learn over a large-scale corpus efficiently. Existing work has shown that PV can outperform the LDA model on several linguistic tasks [1], but its effectiveness for IR remains mostly unknown.

In this paper, we study how to effectively use the PV model in the language model framework to improve ad-hoc retrieval. Specifically, we use the Distributed Bag of Words version of PV (PV-DBOW) because it naturally constructs a

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914688

document language model that fits the framework of the language modeling approach. However, the original PV-DBOW model is not designed for IR, and we find there are three inherent problems make the original PV-DBOW less effective for ad-hoc retrieval. Firstly, the learning objective of PV-DBOW makes it suppress the importance of frequent words excessively. Secondly, PV-DBOW is prone to over-fit short documents during the training iterations. Finally, PV-DBOW does not model word-context associations, making it difficult to capture word substitution relationships that are important in IR. To address these problems, we proposed three modifications to enhance PV-DBOW model for ad-hoc retrieval, including document-frequency based negative sampling, document regularization and a joint learning objective. Empirical results show that consistent and significant improvements over baselines can be obtained with our enhanced PV model.

2. RELATED WORK

Previous work has shown that generative topic models are beneficial for language model estimation. For example, Liu and Croft [6] showed that document clustering can significantly improve retrieval effectiveness when incorporated in language smoothing. The cluster model, also known as the mixture of unigrams model, groups documents into a finite set of clusters (topics) and associates each cluster with a multinomial distribution over the vocabulary. Later, Wei and Croft [12] proposed an LDA-based retrieval model by combining language estimation based on LDA with query likelihood model. Their results showed that the LDA-based retrieval model can consistently outperform the clustering based model.

Recently, there have been several studies exploring the application of word embeddings in the IR scenario. For example, Vulić and Moens [11] construct dense representations for queries and documents by aggregating word vectors and rank results based on the fusion of cosine similarities and query likelihood scores. Ganguly et al. [2] proposed a generalized language model based on word embeddings by considering three term transformation processes. In contract to these studies that construct retrieval models based on bag of word embeddings, our work mainly focuses on how to effectively use the paragraph vector model to improve estimation in the language model approach.

3. ENHANCED PARAGRAPH VECTOR

In this section, we describe the details of how we enhance the PV model for language estimation in ad-hoc retrieval.

3.1 PV-DBOW

PV-DBOW maps words and documents into low-dimension dense vectors. Each document vector is trained to predict the words it contains. Under the bag-of-words assumption, the generative probability of word w in document d is obtained through a softmax function over the vocabulary:

$$P_{PV}(w|d) = \frac{exp(\vec{w} \cdot \vec{d})}{\sum_{w' \in V_w} exp(\vec{w'} \cdot \vec{d})} \qquad (1)$$

where $\vec{w}$ and $\vec{d}$ are vector representations for w and d; and V_w is the vocabulary of the training collections.

In training, negative sampling is used to approximate the softmax function in Equation (1). Formally, the local objective function for each (w, d) pair in PV-DBOW with negative sampling is

$$\ell = \log(\sigma(\vec{w} \cdot \vec{d})) + k \cdot E_{w_N \sim P_n}[\log \sigma(-\vec{w_N} \cdot \vec{d})] \qquad (2)$$

where $\sigma(x) = 1/(1 + exp(-x))$, k denotes the number of negative samples, w_N denotes the sampled word, and $P_n(w)$ denotes the distribution of negative samples. In [7], $P_n(w)$ is defined as the unigram distribution raised to the power 0.75:

$$P_n(w) = \frac{\#(w)^{0.75}}{|C|} \qquad (3)$$

where $\#(w)$ denotes the corpus frequency of w and $|C| = \sum_{w' \in V_w} \#(w')^{0.75}$.

3.2 PV-DBOW based Retrieval Model

From the learning objective of PV-DBOW, we can see that it can be naturally applied in the probabilistic language model framework for IR. With the learned word and document embeddings, we can directly estimate the generative probability of each word given the document in a latent semantic space. Therefore, we can incorporate the language estimation of PV-DBOW into the query likelihood model as a document dependent smoothing technique:

$$P(w|d) = (1 - \lambda)P_{QL}(w|d) + \lambda P_{PV}(w|d) \qquad (4)$$

where $P_{QL}(w|d)$ and $P_{PV}(w|d)$ represent the word probability estimated with QL and PV-DBOW respectively. λ is the parameter that controls the weights of QL and PV-DBOW.

3.3 Adaptation for IR

Now we describe in detail the major problems of the original PV-DBOW model that makes it less effective for IR, as well as the techniques we employ to solve these issues.

Document Frequency Based Negative Sampling. Following the idea in [5], we can see that PV-DBOW with negative sampling is implicitly factorizing a shifted matrix of point-wise mutual information between words and documents:

$$\vec{w} \cdot \vec{d} = \log(\frac{\#(w,d)}{\#(d)} \cdot \frac{|C|}{\#(w)}) - \log(k) \qquad (5)$$

where $\#(w,d)$ is the term frequency of w in d; $\#(d)$ is the length of d and k is the number of negative instances. From Equation (5), we can see that the original PV-DBOW model implicitly weights words according to inverse corpus frequencies (ICF). However, previous studies have shown that term weighting with ICF may over-penalize frequent words, and often performs worse than term weighting with inverse document frequency (IDF) [9]. Inspired by this, we propose a novel document-frequency based negative sampling strategy for PV-DBOW to better fit the IR scenario. More formally, we replace $P_n(w)$ with a new sample distribution:

$$P_n(w) = \frac{\#D(w)}{\sum_{w' \in V_w} \#D(w')} \qquad (6)$$

where $\#D(w)$ represents the document frequency of w. We can find that the new learning objective of PV-DBOW with document-frequency based negative sampling is equal to the following factorization:

$$\vec{w} \cdot \vec{d} = \log(\frac{\#(w,d)}{\#(d)} \cdot \frac{\sum_{w' \in V_w} \#D(w')}{\#D(w)}) - \log(k) \qquad (7)$$

Since k and $\sum_{w' \in V_w} \#D(w')$ are constants, the training process of PV-DBOW with document-frequency based negative sampling is actually factorizing a shifted tf-idf matrix.

In practice, the exact value of the inverse document frequency is too aggressive for tf-idf weighting and its logarithmic version is more widely used. To achieve similar effects, we adapt a power version of document frequency that uses $\#D(w)^\eta$ $(\eta \leq 1)$ instead of $\#D(w)$.

Document Regularization. The original PV-DBOW does not handle the varied lengths of documents, making it prone to over-fit short documents during the training iterations. Specifically, through the training process of PV-DBOW, vector norms of long documents remain roughly the same while vector norms of short documents keep growing. Increasing vector norms affect the dot product value in Equation (1) and make the language estimation concentrate on the observed words. This in turn significantly decreases the smoothing power of the PV-DBOW model on short documents. To solve this problem, we propose to introduce document regularization into the learning objective to avoid the ever-growing norm of short documents. Specifically, we add an L2 constraint on the document norm to the learning objective of PV-DBOW:

$$\ell = \log(\sigma(\vec{w} \cdot \vec{d})) + k \cdot E_{w_N \sim P_n}[\log \sigma(-\vec{w_N} \cdot \vec{d})] - \frac{\gamma}{\#(d)} ||\vec{d}||^2 \quad (8)$$

where $\#(d)$ is the number of words in d, $||\vec{d}||$ is the norm of vector $\vec{d}$ and γ is a hyper-parameter that control the strength of regularization. Each iteration of the stochastic gradient descent in PV-DBOW goes through each word exactly once, so we use the document length $1/\#(d)$ to ensure equal regularizations over long and short documents.

Joint Objective. The original PV-DBOW model learns over the word-document co-occurrence information as shown in Equation (2), making it focus on capturing syntagmatic relations between words (i.e., words that frequently co-occur in same documents). It lacks the modeling of paradigmatic relations between words (e.g. "car" and "vehicle") since no word-context information is leveraged in its learning process. As suggested by [1, 10], by modeling both word-document and word-context information, one can usually obtain better word and document vectors for NLP tasks. Following the same idea as [10], we introduce a joint learning objective to the PV-DBOW model. Specifically, we apply a two-layer structure that first uses the document to predict the target word and then uses the target word to predict its context. The new objective function is as follows:

$$\begin{aligned} \ell &= \log(\sigma(\vec{w_i} \cdot \vec{d})) + k \cdot E_{w_N \sim P_n}[\log \sigma(-\vec{w_N} \cdot \vec{d})] \\ &+ \sum_{\substack{j=i-L \\ j \neq i}}^{i+L} \log(\sigma(\vec{w_i} \cdot \vec{c_j})) + k \cdot E_{c_N \sim P_n}[\log \sigma(-\vec{w_i} \cdot \vec{c_N})] \quad (9) \end{aligned}$$

where $\vec{c_j}$ is the context vector for word w_j, c_N denotes the sampled context and L represents the context window size.

4. EXPERIMENTS

Experimental Setup. We evaluate three baselines: query likelihood model (QL), LDA-based retrieval model (LDA-LM) and original PV-DBOW model (PV-LM). We add document frequency based negative sampling (D), document regularization (R), and joint objective (J) to PV-DBOW one by one, and refer to the enhanced PV based retrieval model as EPV-D-LM, EPV-DR-LM, and EPV-DRJ-LM respectively. We use two TREC collections, Robust04 and GOV2. We report the results of different versions of enhanced PV based retrieval models on Robust04, but only the full model (EPV-DRJ-LM) on GOV2 due to the space limitation. We use the Galago search engine[1] to index the corpus and report results for both the title and description of each TREC topic (stop words removed). Queries and documents are stemmed with the Krovetz stemmer. For test efficiency, we adopt a re-ranking strategy. An initial retrieval is performed with QL to obtain 2,000 candidate documents, and then a re-ranking is performed with both LDA-LM and EPV based retrieval models. The final evaluation is based on the top 1,000 results. We use a 5-fold cross validation in the same way as [3]: 4 folds are used to tune λ in smoothing process and 1 fold is used to test retrieval performance. We includes three evaluation metrics: mean average precision (MAP), normalized discounted cumulative gain at 20 (nDCG@20) and precision at 20 (P@20).

Parameter Settings. We train both LDA and EPV on the whole Robust04 collection. However, for the GOV2 collection, due to the prohibitive training time, we train both LDA and EPV on a randomly sampled subset with 500k documents for fair comparison. The topic number (K) in LDA and the vector dimension in PV-DBOW/EPV are empirically set as 300. For LDA, we set the hyper-parameters α and β to $50/K$ and 0.01 as described in [12]. For EPV, we tuned γ from 1 to 100 (1, 10 and 100), and η from 0.1 to 0.9 (0.1 per step). The final value for γ is 10 (Robust04/GOV2), for η is 0.1 (Robust04) and 0.2 (GOV2).

Results. The results on Robust04 are shown in the top part of Table 1. As we can see, by incorporating topic level estimation, LDA-LM can outperform the QL model on both topic titles and descriptions. Meanwhile, by estimating the language model using the original PV-DBOW model, PV-LM obtains very similar results as LDA-LM. By adding the proposed techniques one by one to enhance the PV-DBOW model for IR, we obtain better and better retrieval performance. The results indicate the effectiveness of the proposed techniques for the PV based retrieval model. Finally, the full enhanced model EPV-DRJ-LM can outperform both QL and LDA-LM significantly on both topic titles and descriptions. For example, the relative MAP improvement of EPV-DRJ-LM over QL and LDA-LM in Robust04 is 5.5% and 3.5% on titles, 2.5% and 2.4% on descriptions, respectively.

From the results on GOV2, however, we find that the incorporation of the LDA model may even hurt the retrieval performance in most cases. A major reason is that GOV2 is a large Web collection with many diverse and noisy topics. By using only 300 topics, the learned topics in LDA might be too coarse and noisy, which can hurt the language model estimation. Therefore, one may observe better performance with LDA-LM by increasing the number of topics (with correspondingly lower efficiency). On the other hand, although the vector dimension of our enhanced PV model is also 300, the potential number of topics is not limited to that number. Therefore, EPV-DRJ-LM can capture much finer topic relations between words and documents, and produce better language estimation in the latent semantic space. We

[1] http://www.lemurproject.org/galago.php

Table 1: Comparison of different models over Robust04 and GOV2 collection. $*$, $+$ means significant difference over QL, LDA-LM respectively at 0.05 significance level measured by Fisher randomization test.

	Robust04 collection					
	Topic titles			Topic descriptions		
Method	MAP	nDCG@20	P@20	MAP	nDCG@20	P@20
QL	0.253	0.415	0.369	0.246	0.391	0.334
LDA-LM	0.258^{*}	0.421	0.374^{*}	0.247	0.392	0.336
PV-LM	0.259^{*}	0.418	0.371	0.247	0.392	0.335
EPV-D-LM	0.260^{*}	0.417	0.371	0.251^{*}	0.397^{*}	0.340^{*}
EPV-DR-LM	0.262^{*}	0.418	0.368	0.252^{*+}	0.397^{*}	0.338^{*}
EPV-DRJ-LM	**0.267**$^{*+}$	**0.425***	**0.376***	**0.253**$^{*+}$	**0.404**$^{*+}$	**0.347**$^{*+}$
	GOV2 collection					
	Topic titles			Topic descriptions		
Method	MAP	nDCG@20	P@20	MAP	nDCG@20	P@20
QL	0.295	0.409	0.510	0.249	0.371	0.470
LDA-LM	0.292	0.405	0.504	0.244	**0.375**	0.467
EPV-DRJ-LM	**0.297**$^{+}$	**0.415**$^{*+}$	**0.519**$^{*+}$	**0.252**$^{*+}$	0.371	**0.472**

observe much better performance with EPV-DRJ-LM compared with both QL and LDA-LM.

The results in Table 1 also show that the topic level smoothing is more effective on short queries (topic titles) than long queries (topic descriptions). For example, the relative improvement of LDA-LM over QL is 2.0% on titles and 0.4% on descriptions in terms of MAP respectively; while the relative improvement of EPV-DRJ-LM over QL is 5.5% on titles and 2.5% on descriptions in terms of MAP. With fewer words in a query, the language model estimation would be more difficult based on exact matching. Therefore, by involving topic level estimation, the smoothing technique can bring larger benefits by alleviating the vocabulary mismatch problem.

5. CONCLUSION

In this paper, we study how to effectively use the PV model to improve ad-hoc retrieval. We identify several problems that make the original PV-DBOW model less effective for the IR scenario. To solve these issues, we proposed three techniques to enhance the original PV model. The experimental results demonstrate the effectiveness of our enhanced PV based retrieval model compared with the state-of-the-art topic enhanced language models. This is also the first study to show that a PV model can work better than a topic model on language model estimation for IR.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF IIS-1160894, and in part by NSF grant IIS-1419693. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] A. M. Dai, C. Olah, Q. V. Le, and G. S. Corrado. Document embedding with paragraph vectors. In *NIPS Deep Learning Workshop*, 2014.

[2] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 795–798. ACM, 2015.

[3] S. Huston and W. B. Croft. A comparison of retrieval models using term dependencies. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 111–120. ACM, 2014.

[4] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[5] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014.

[6] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM, 2004.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. CJorrado, and M. I. Dean, Jeffdan. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[9] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[10] F. Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 136–145, Beijing, China, 2015.

[11] I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372. ACM, 2015.

[12] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[13] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141, 2008.

[14] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.