UMassAmherst

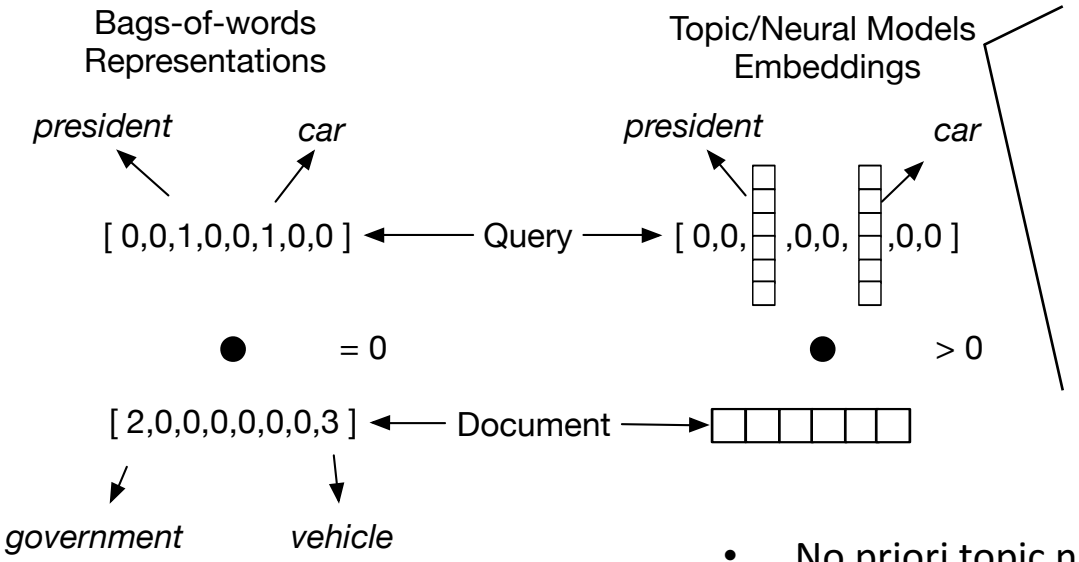
Analysis of the Paragraph Vector Model for Information Retrieval

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

UMASS AMHERST

Motivation

- Most tasks in IR benefit from representations that reflect the semantic relationships between words and documents.
- Word-document matching is essential for language modeling approaches.



- Topic models
  - PLSA
  - LDA
  - …
- Neural models
  - Word2vec
  - **Paragraph vector model**

- No priori topic number
- Highly efficient in training
- Automatically learn document representations
- **Language model**
- **Optimize a weighting scheme widely used in IR**

UMASS AMHERST

Outline

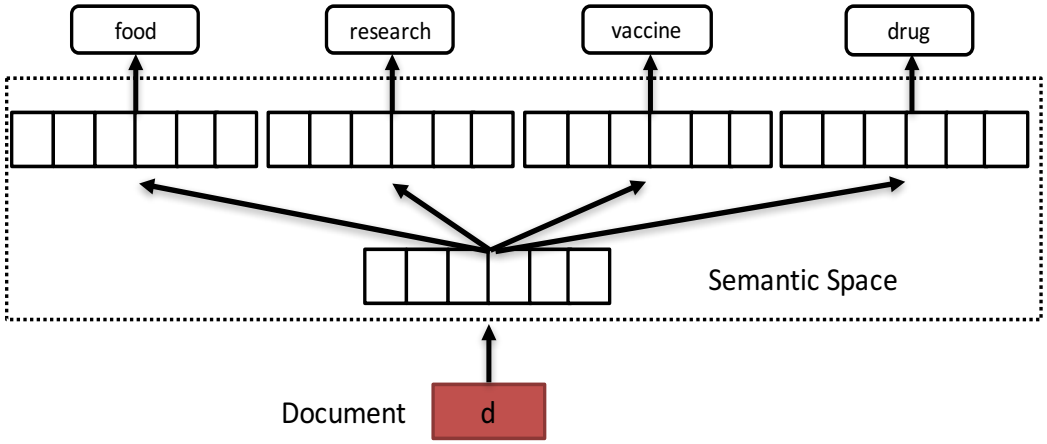
- Paragraph Vector Based Retrieval Model
  - What is paragraph vector model
  - How to use it for retrieval
- Issues of Paragraph Vector Model in Retrieval Scenario
  - Over-fitting on short documents
  - Improper noise distribution
  - Insufficient modeling for word substitution
- Experiments
  - Experiment setup
  - Results
  - Parameter sensitivity

Paragraph Vector Model

- Paragraph vector model [13] jointly learns embedding for words and documents through optimizing the probabilities of observed word-document pairs defined as:

$$P(w|d) = \frac{exp(\vec{w} \cdot \vec{d})}{\sum_{w' \in V_w} exp(\vec{w'} \cdot \vec{d})} \quad (1)$$

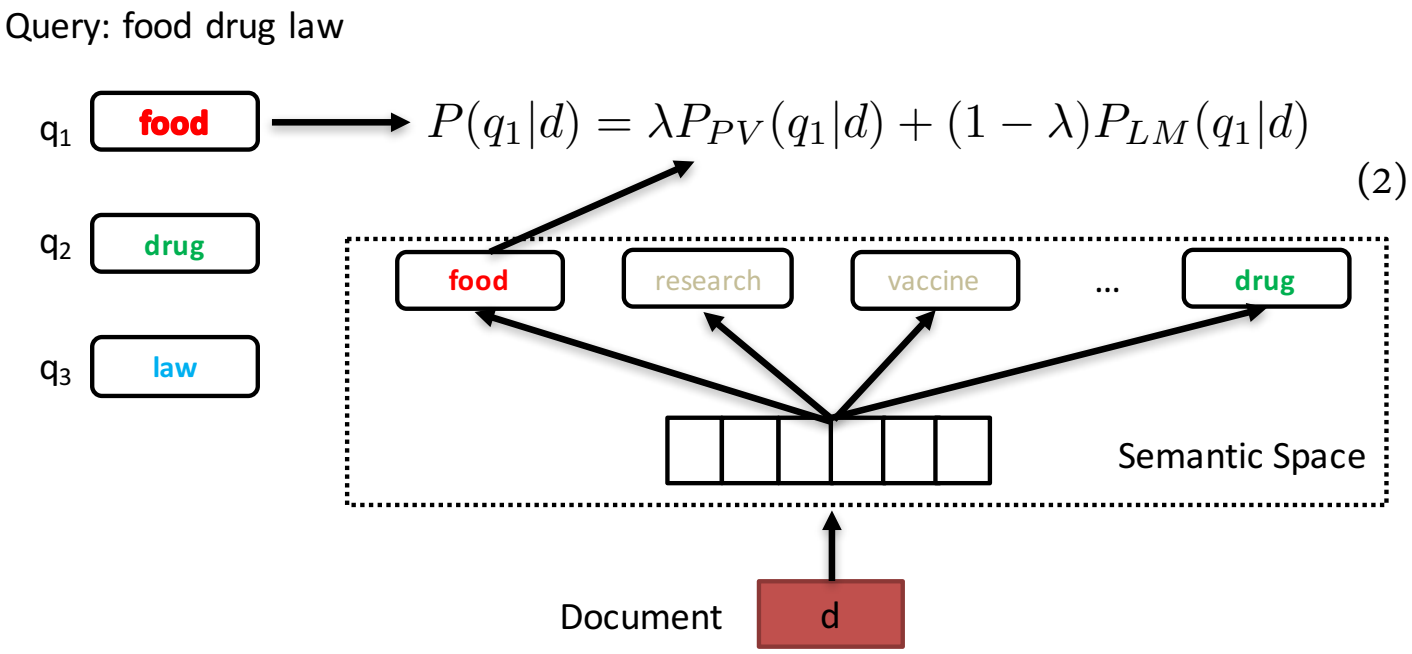
- The following figure describes the structure of Paragraph vector model with distributed bag-of-words assumption (PV-DBOW).



UMASS AMHERST

Language Estimation with Paragraph Vector Model

- Inspired by LDA-based retrieval model [24], we apply paragraph vector model by smoothing the probability estimation in language modeling approaches with PV-DBOW and propose a paragraph vector based retrieval model (PV-LM).



Language Estimation with Paragraph Vector Model

- However, PV-LM did not produce promising results:
  - The performance of PV-LM is highly sensitive to the training iteration of PV-DBOW.
  - The mean average precision (MAP) of PV-LM does not outperform LDA-LM [24] on Robust04 (0.259).

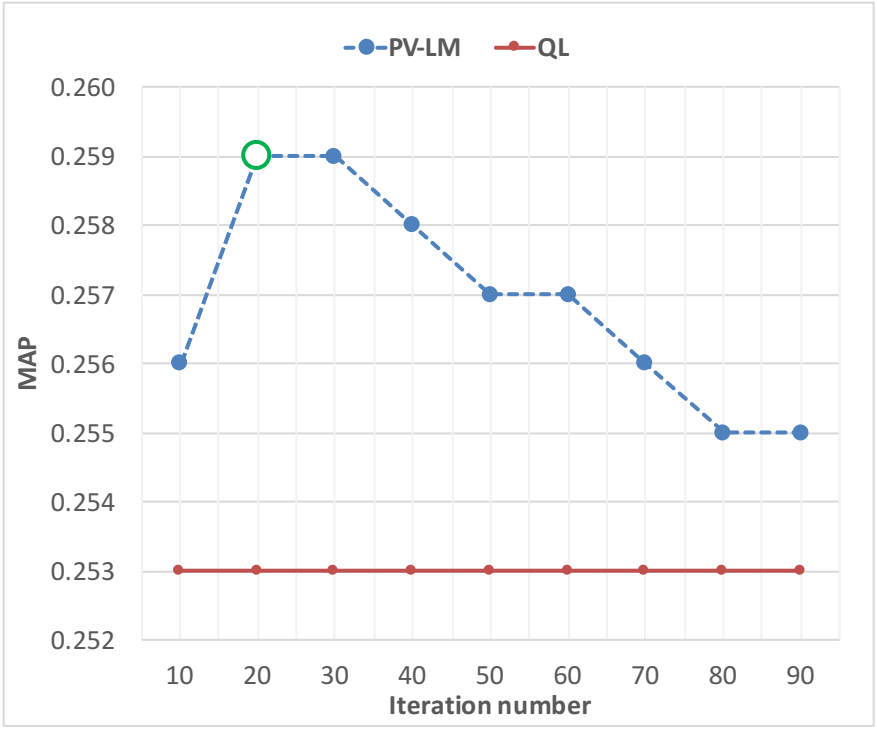


Figure 1: The MAP of QL and the PV-based retrieval model with the original PV-DBOW on Robust04 with title queries in respect of different training iteration.

UMASS AMHERST

Outline

- Paragraph Vector Based Retrieval Model
  - What is paragraph vector model
  - How to use it for retrieval
- Issues of Paragraph Vector Model in Retrieval Scenario
  - Over-fitting on short documents
  - Improper noise distribution
  - Insufficient modeling for word substitution
- Experiments
  - Experiment setup
  - Results
  - Parameter sensitivity

UMASS AMHERST

Overfitting on Short Documents

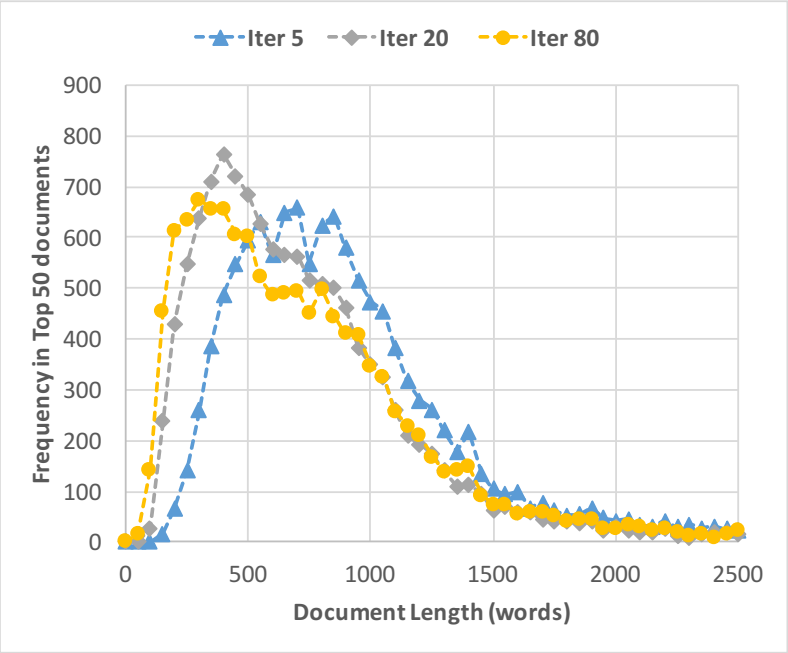


Figure 2: The distribution of documents in respect of document length for top 50 documents retrieved by PV-based retrieval model on Robust04 (title queries).

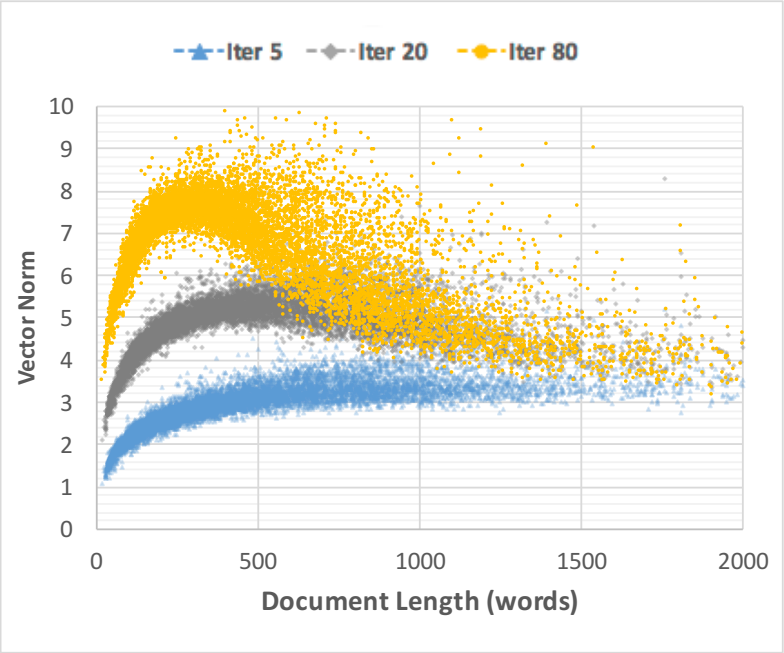


Figure 3: The distribution of vector norms in respect of document length for 10,000 documents randomly sampled from Robust04.

- The PV-based retrieval model tends to retrieve more short documents as training iteration increases.
- In a subset of 10,000 random sampled documents, we observed significant norm increase for short documents' vectors.

UMASS AMHERST

Overfitting on Short Documents

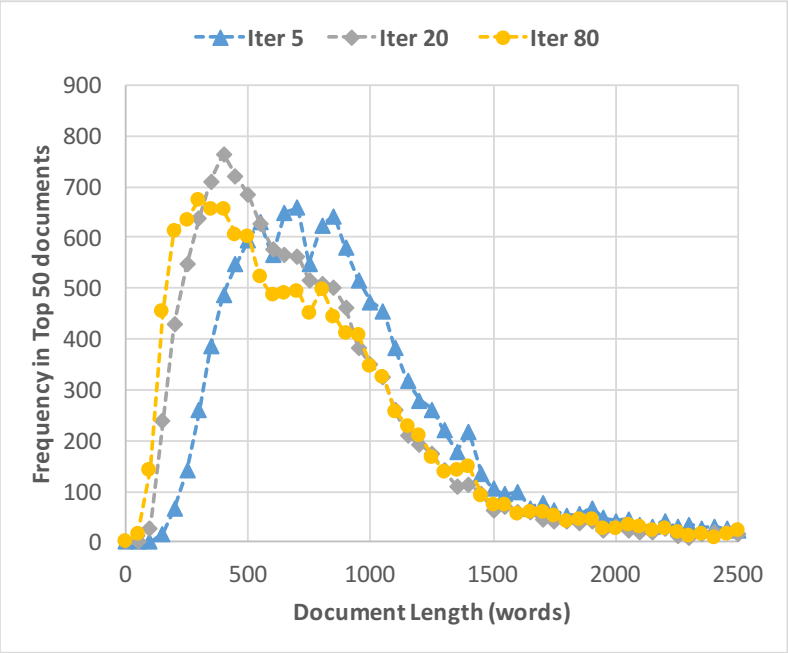


Figure 2: The distribution of documents in respect of document length for top 50 documents retrieved by PV-based retrieval model on Robust04 (title queries).

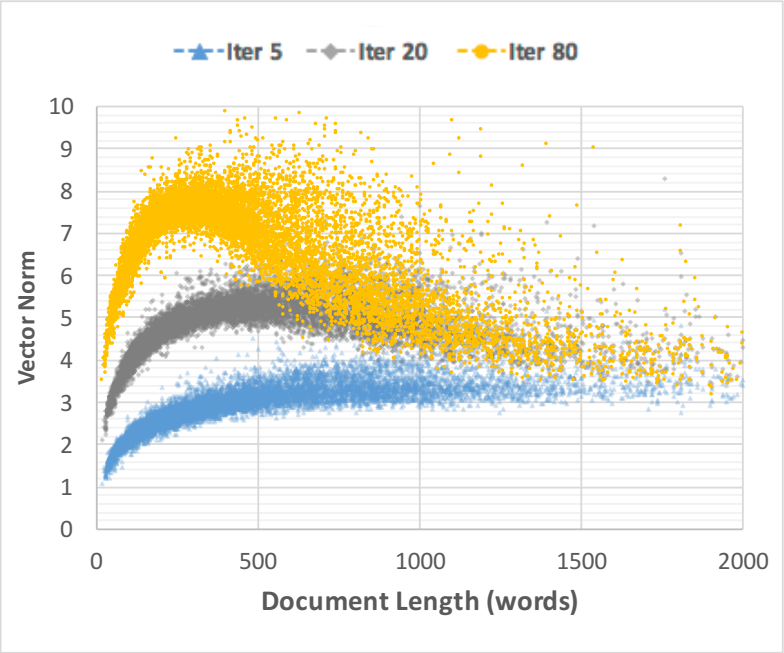
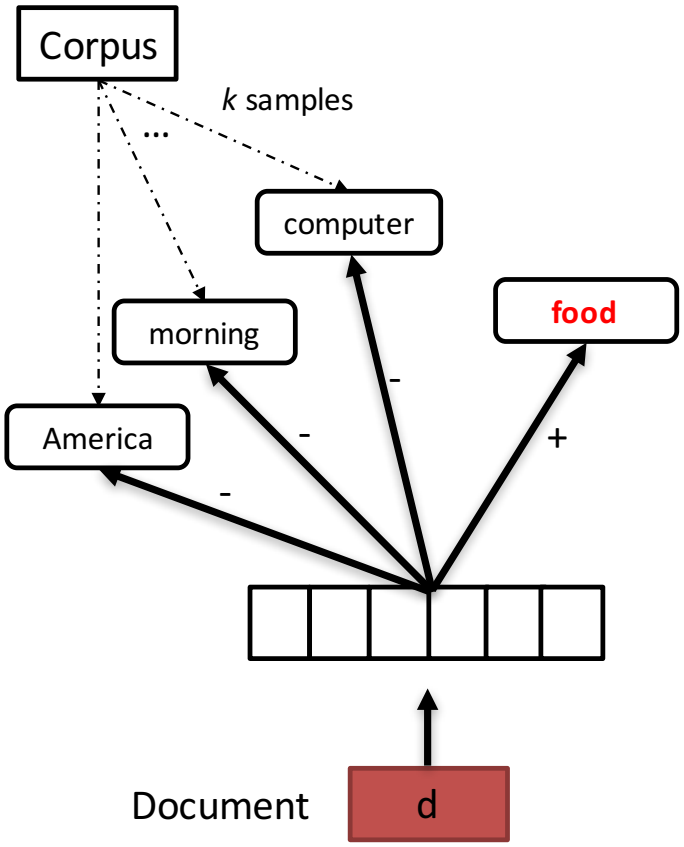


Figure 3: The distribution of vector norms in respect of document length for 10,000 documents randomly sampled from Robust04.

- Long document vector norms change the probability distribution of document language models and makes them focus on observed words.
- One direct solution to this problem is L2 regularization:

$$\ell'(w,d) = \ell(w,d) - \frac{\gamma}{\#d}||\vec{d}||^2 \qquad (3)$$

Negative Sampling



- Proposed by Mikolov et al. [17], negative sampling is a technique that approximates the global objective of PV-DBOW by sampling "negative" terms from corpus:

$$\ell = \sum_{w \in V_w} \sum_{d \in V_d} \#(w,d) \log(\sigma(\vec{w} \cdot \vec{d})) + \sum_{w \in V_w} \sum_{d \in V_d} \#(w,d)(k \cdot E_{w_N \sim P_V}[\log \sigma(-\vec{w_N} \cdot \vec{d})]) \quad (4)$$

- If we derived the local objective of a specific word-doc pair and let its partial derivative equal to zero. Then we have:

$$\vec{w} \cdot \vec{d} = \log(\frac{\#(w,d)}{\#(d)} \cdot \frac{1}{P_V(w)}) - \log k \quad (5)$$

UMASS AMHERST

Improper Noise Distribution

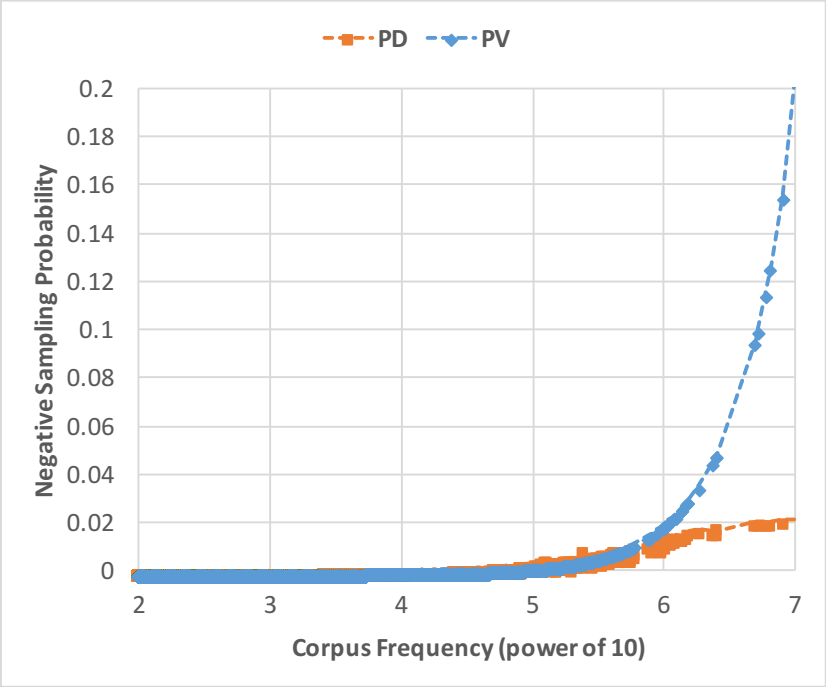


Figure 4: The distribution of the original negative sampling (PV) and the document-frequency based negative sampling (PD). The horizontal axis represents log value of word frequency (base 10).

$$\vec{w} \cdot \vec{d} = \log(\frac{\#(w,d)}{\#(d)} \cdot \frac{1}{P_V(w)}) - \log k \quad (5)$$

- The original negative sampling technique adopts a empirical word distribution as:

$$P_V(w_N) = \frac{\#w_N}{|C|} \quad (6)$$

  which makes the original PV-DBOW optimizing a variation of TF-ICF weighting scheme.

- Empirically:
  - CF-based negative sampling suppresses frequent words too much.
  - TF-ICF weighting loses the document structure information

- We proposed a document-frequency based noise distribution:

$$P_D(w_N) = \frac{\#D(w_N)}{\sum_{w' \in V_w} \#D(w')} \quad (7)$$

  which makes the PV-DBOW optimizing a variation of TF-IDF weighting scheme.

Insufficient Modeling for Word Substitution

Table 1: The cosine similarities between "*clothing*", "*garment*" and four relevant documents in Robust04 query 361 ("*clothing sweatshops*").

	PV-DBOW	
	clothing	*garment*
clothing	1.000	0.632
LA112689-0194 ($TF_{clothing} = 2, TF_{garment} = 26$)	0.044	0.134
LA112889-0108 ($TF_{clothing} = 0, TF_{garment} = 10$)	-0.003	0.100
LA021090-0137 ($TF_{clothing} = 7, TF_{garment} = 9$)	0.052	0.092
LA022890-0105 ($TF_{clothing} = 6, TF_{garment} = 6$)	0.066	0.079

- Existing topic models and embedding models mainly focus on two types of word relations: co-occurrence (e.g. topic related words) and substitution (e.g. synonyms)
- PV-DBOW focuses on capturing word co-occurrence but ignores word-context information, which makes it difficult to understand word substitution relation (e.g. "*clothing*" and "*garment*").

Insufficient Modeling for Word Substitution

Table 1: The cosine similarities between "*clothing*", "*garment*" and four relevant documents in Robust04 query 361 ("*clothing sweatshops*").

	PV-DBOW		PV joint objective	
	clothing	*garment*	*clothing*	*garment*
clothing	1.000	0.632	1.000	0.638
LA112689-0194 ($TF_{clothing} = 2, TF_{garment} = 26$)	0.044	0.134	0.107	0.169
LA112889-0108 ($TF_{clothing} = 0, TF_{garment} = 10$)	-0.003	0.100	0.126	0.155
LA021090-0137 ($TF_{clothing} = 7, TF_{garment} = 9$)	0.052	0.092	0.147	0.119
LA022890-0105 ($TF_{clothing} = 6, TF_{garment} = 6$)	0.066	0.079	0.107	0.107

- As suggested by Dai et al. [5] and Sun et al. [22], one approach to alleviate the problem is regularizing PV-DBOW by requiring word vectors to predict their context. Specifically, we apply a joint objective as:

$$\ell = \log(\sigma(\vec{w_i} \cdot \vec{d})) + k \cdot E_{w_N \sim P_V}[\log \sigma(-\vec{w_N} \cdot \vec{d})] + \sum_{\substack{j=i-L \\ j \neq i}}^{i+L} \log(\sigma(\vec{w_i} \cdot \vec{c_j})) + k \cdot E_{c_N \sim P_V}[\log \sigma(-\vec{w_i} \cdot \vec{c_N})] \quad (8)$$

UMASS AMHERST

Outline

- Paragraph Vector Based Retrieval Model
  - What is paragraph vector model
  - How to use it for retrieval
- Issues of Paragraph Vector Model in Retrieval Scenario
  - Over-fitting on short documents
  - Improper noise distribution
  - Insufficient modeling for word substitution
- Experiments
  - Experiment setup
  - Results
  - Parameter sensitivity

UMASS AMHERST

Experiment Setup

- Datasets:
  - TREC collections: Robust04, GOV2* with title and description queries
  - Five-fold cross validation
  - Evaluation: mean average precision (MAP), normalized discounted cumulative gain (NDCG@20) and precision (P@20)

- Reported Models:
  - QL: Query likelihood model [19] with Dirichlet smoothing.
  - LDA-LM: LDA-based retrieval model proposed by Wei and Croft [15].
  - PV-LM: the PV-based retrieval model with the PV-DBOW proposed by Le et al. [13]
  - EPV-R-LM: the PV-LM model with L2 regularization.
  - EPV-DR-LM: the EPV-R-LM model with document frequency based negative sampling.
  - EPV-DRJ-LM: the EPV-DR-LM model with joint objective.

* Due to the efficiency issues, we used a random subset with 500k documents to train LDA and PV on GOV2

UMASS AMHERST

Experiment Results

Table 2: 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. The best performance is highlighted in boldface.

	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-R-LM	0.259*	0.418	0.370	0.247	0.393	0.336
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	0.290	0.406	0.505	0.245	**0.376**	0.468
PV-LM	0.294	0.409	0.510+	0.246	0.364	0.463
EPV-R-LM	0.295+	0.410	0.511+	0.250+	0.368	0.467
EPV-DR-LM	0.296+	0.412	0.512	0.250+	0.371	0.470
EPV-DRJ-LM	**0.297**+	**0.415***+	**0.519***+	**0.252***+	0.371	**0.472**

UMASS AMHERST

Parameter Sensitivity

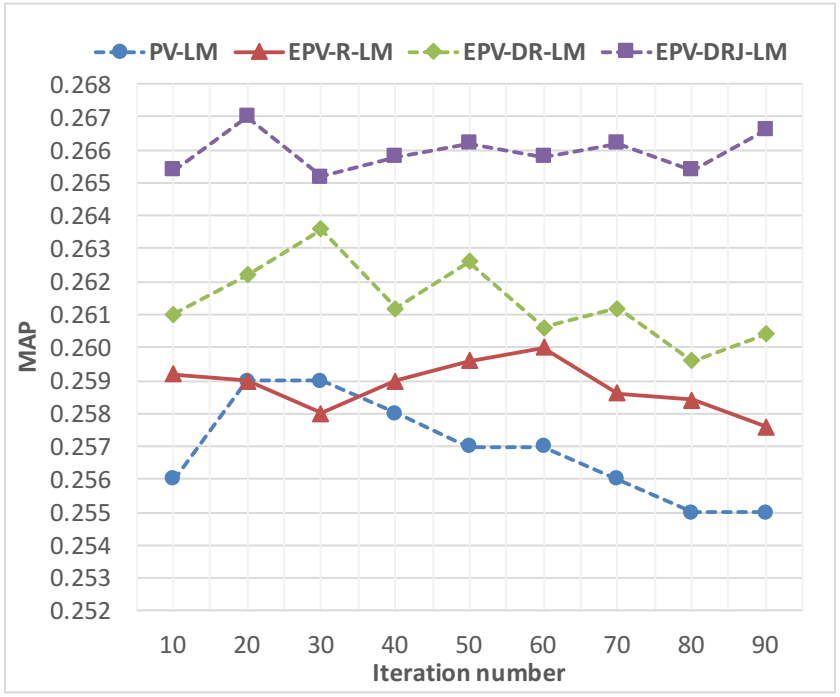


Figure 5: MAP variation of PV-based retrieval models with respect to iteration number on Robust04 title queries.

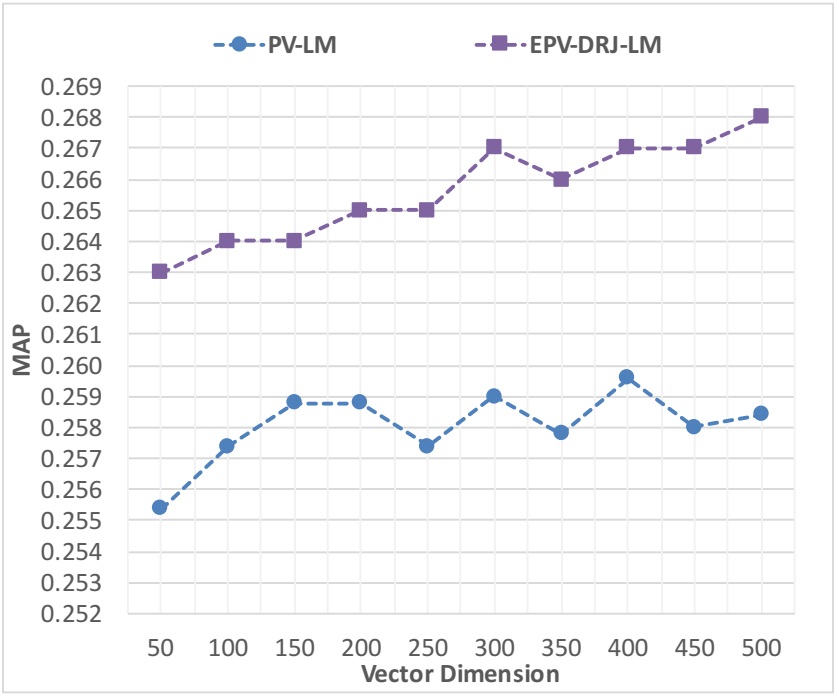


Figure 6: MAP variation of PV-based retrieval models with respect to vector dimensions on Robust04 title queries.

- With L2 regularization, EPV models showed better stability in long term training.
- Increasing vector dimension does not consistently improve the performance of PV-LM, but it seems beneficial for EPV-DRJ-LM.

UMASS AMHERST

Conclusion

- In this work, we focus on the theoretic and empirical analysis of the paragraph vector model for language estimation in IR.
- We identify three issues of PV-DBOW:
  - It is vulnerable to over-fitting short documents.
  - Its original negative sampling strategy suppresses frequent words too much.
  - It lacks sufficient modeling for word substitution relations.
- Things to note:
  - Vector norms affect the language estimation of paragraph vector models.
  - The noise distribution of negative sampling determines the optimization objective of paragraph vector models

UMASS AMHERST

Thanks for listening!

Thanks SIGIR travel grants for supporting the presentation of this work

aiqy@cs.umass.edu
http://www.cs.umass.edu/~aiqy/

UMASS AMHERST

Reference

- [1] Q. Ai, L. Yang, J. Guo, and W. B. Croft. Improving language estimation with the paragraph vector model for ad-hoc retrieval. In Proceedings of the 39th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2016.
- [2] A. Atreya and C. Elkan. Latent semantic indexing (lsi) fails for trec collections. ACM SIGKDD Explorations Newsletter, 12(2):5–10, 2011.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, Mar. 2003.
- [4] K. W. Church and W. A. Gale. Poisson mixtures. Natural Language Engineering, 1(02):163–190, 1995.
- [5] A. M. Dai, C. Olah, Q. V. Le, and G. S. Corrado. Document embedding with paragraph vectors. In NIPS Deep Learning Workshop, 2014.
- [6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. JAsIs, 41(6):391–407, 1990.
- [7] 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.
- [8] T. L. Griffiths and M. Steyvers. Finding scientific topics. PNAS, 101(suppl. 1):5228–5235, 2004.
- [9] D. Hiemstra and W. Kraaij. Twenty-one at trec-7: Ad-hoc and cross-language track. 1999.
- [10] T. Hofmann. Probabilistic latent semantic indexing. InProceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM, 1999.
- [11] 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.
- [12] R. Krovetz. Viewing morphology as an inference process. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 191–202. ACM, 1993.
- [13] 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.
- [14] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems, pages 2177–2185, 2014.
- [15] 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.

UMASS AMHERST

Reference

- [16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- [17] 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.
- [18] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving document ranking with dual word embeddings. In Proceedings of the 25th International Conference Companion on World Wide Web, pages 83–84. International World Wide Web Conferences Steering Committee, 2016.
- [19] 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.
- [20] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation, 60(5):503–520, 2004.
- [21] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 623–632. ACM, 2007.
- [22] 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 Annual Meeting of the Association for Computational Linguistics, 2015.
- [23] I. Vulí c 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.
- [24] 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.
- [25] 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.
- [26] L. Zhao and J. Callan. Term necessity prediction. InProceedings of the 19th ACM international conference on Information and knowledge management, pages 259–268. ACM, 2010.
- [27] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In Proceedings of the 20th Australasian Document Computing Symposium, pages Article–No. ACM, 2015.