A Study of Document Expansion using Translation Models and Dimensionality Reduction Methods

Textual Data Analytics (TEANA) lab

Saeid Balaneshin-kordan
saeid@wayne.edu

Alexander Kotov
kotov@wayne.edu

Document expansion: can be done using smoothing methods, translation models, and dimensionality reduction techniques, such as matrix decompositions and topic models.

Problem: these research avenues have been individually explored in many previous studies, but there is still a lack of understanding of how state-of-the-art methods for each of them compare with each other in terms of retrieval accuracy.

Goal: fill in this void by reporting the results of an empirical comparison of document expansion methods using translation models estimated based on word co-occurrence and cosine similarity between low-dimensional word embeddings, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), on standard TREC collections (TREC 7-8, ROBUST04, GOV)

Translation Model:

- Quantifies the strength of semantic relationship between pairs of words
- **Method 1 (TM-CX)**: Probability of translating word u to word w [Karimzadehgan and Zhai, ECIR'12]:

$$p_{\mathrm{tr}}(w|u) = \frac{c(w,u)}{\sum_{v \in \mathbb{V}} c(v,u) + |V|}$$

- **Method 2 (TM-WE)**: Semantic similarity between the words in the word embeddings space is calculated based on the cosine similarity of their corresponding word vectors [Zuccon, Koopman, et al. ADCS'15].
- **Document expansion LM**:

$$p_{\mathrm{t}}(w|d) = \sum_{u \in \mathbb{V}} p_{\mathrm{tr}}(w|u) p_{\mathrm{ml}}(u|d)$$

Latent Dirichlet Allocation:

- Approximates documents as mixtures of latent topics
- Models document collection with a probabilistic generative process:
  - draw latent topics $p_\varphi(w|z)$ from Dir(β)
  for each document d:
    - draw a distribution over topics (i.e., $p_\theta(z|d)$) from Dir(α)
    for each word position in d:
      - draw a topic z from the distribution $p_\theta(z|d)$.
      - draw a word w from the distribution $p_\varphi(w|z)$.
- Number of topics determine the dimensionality of latent space
- **Document expansion LM** [Wei, Croft, SIGIR'06]:

$$p_{\mathrm{lda}}(w|d) = \sum_{z \in Z} p_\phi(w|z) p_\theta(z|d)$$

Non-negative Matrix Factorization:

- Approximates sparse high-dimensional TF-IDF term-document matrix **P** with a product of dense lower dimensional matrices:

$$\mathbf{P} = \mathbf{P}_{\mathrm{b}} \mathbf{P}_{\mathrm{e}}$$

by solving the following optimization problem:

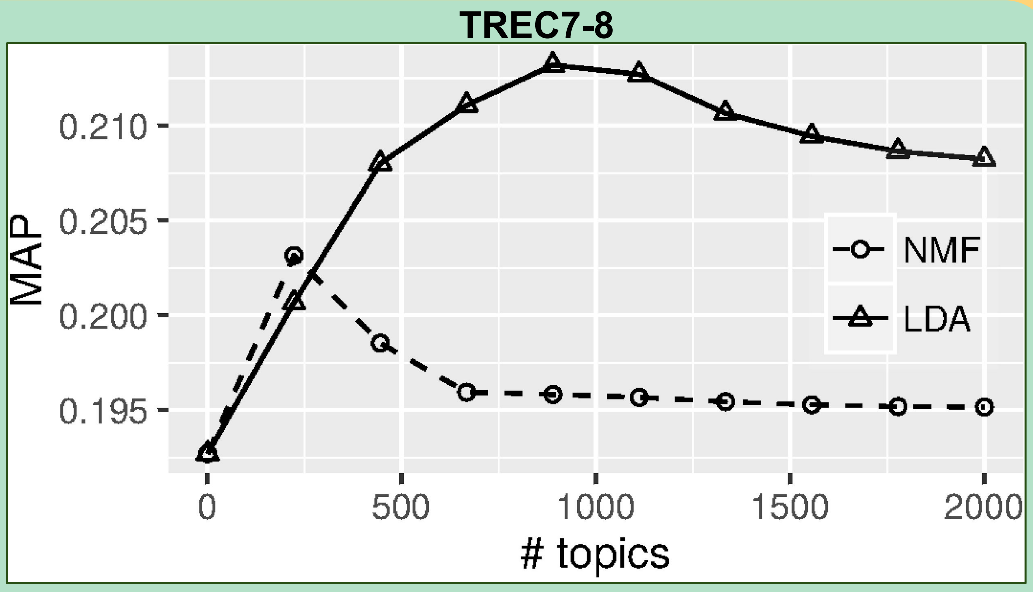
$$\min_{\mathbf{P}_{\mathrm{b}}, \mathbf{P}_{\mathrm{e}}} \frac{1}{2} \sum_i \sum_j [\mathbf{P}_{i,j} - (\mathbf{P}_{\mathrm{b}} \mathbf{P}_{\mathrm{e}})_{i,j}]^2$$

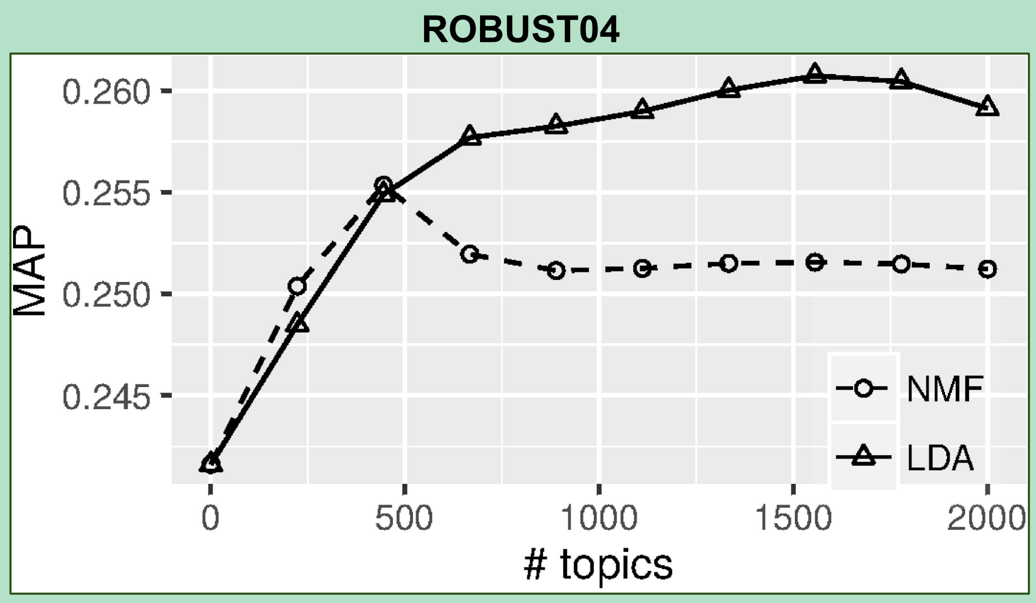
- Inner dimensions of dense matrices determine dimensionality of latent space
- **Document expansion LM**:

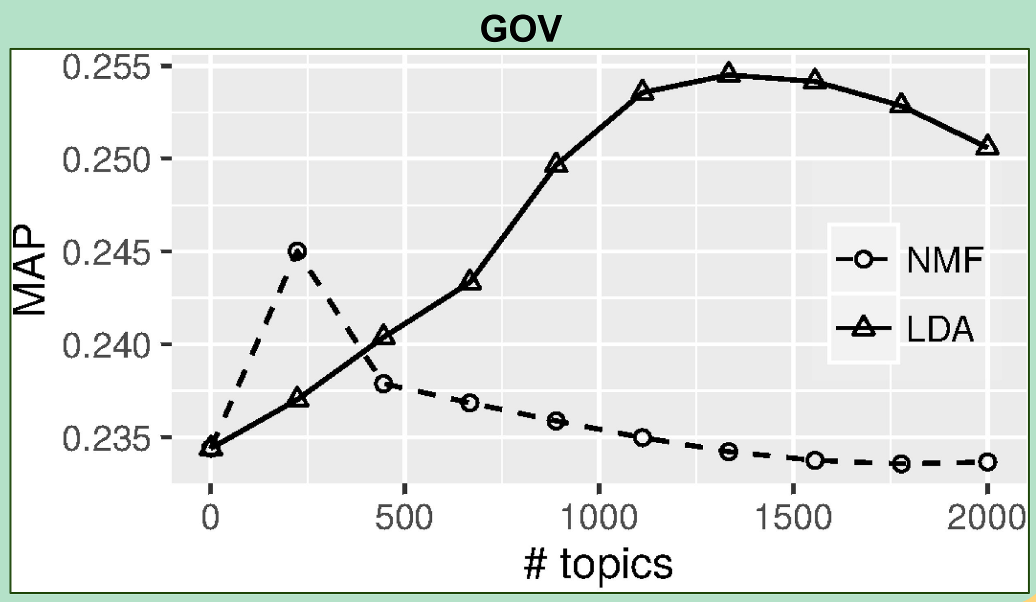
$$p_{\mathrm{nmf}}(w|d) = \sum_{z \in Z} p_{\mathrm{b}}(w|z) p_{\mathrm{e}}(z|d)$$

Collection	Method	All Queries MAP	All Queries NDCG@20	All Queries P@20	Difficult Queries MAP	Difficult Queries NDCG@20	Difficult Queries P@20
TREC7-8	QL-DIR	0.1927	0.4142	0.339	0.0205	0.1303	0.1326
	TM-CX	0.1963	0.4149	0.3603	0.0216	0.1346	0.1165
	TM-WE	*0.2084* ☆✙	*0.4155*	*0.3678*	*0.0434* ☆✙	*0.1431*	*0.151*
		(+8.1% / +6.2%)	*(+0.3% / +0.1%)*	*(+8.5% / +2.1%)*	*(+111.7% / +100.9%)*	*(+9.8% / +6.3%)*	*(+13.9% / +29.6%)*
	NMF	0.2035 ☆✙	0.4143	0.3655	0.0229 ☆	0.1317	0.1174
		(+5.6% / +3.7%)	(+0.0% / -0.1%)	(+7.8% / +1.4%)	(+11.7% / +6.0%)	(+1.1% / -2.2%)	(-11.5% / +0.8%)
	LDA	**0.2138** ☆✙	**0.4312**	**0.3745**	**0.0565** ☆✙	**0.1605**	**0.1652**
		(+10.9% / +8.9%)	**(+4.1% / +3.9%)**	**(+10.5% / +3.9%)**	**(+175.6% / +161.6%)**	**(+23.2% / +19.2%)**	**(+24.6% / +41.8%)**
ROBUST04	QL-DIR	0.2416	0.4065	0.3504	0.023	0.0997	0.0962
	TM-CX	0.2543	0.4104	0.3607	0.0353	0.1004	0.0972
	TM-WE	*0.2582* ☆✙	*0.4191*	*0.3634*	0.0345 ☆	*0.1026*	0.0979
		(+6.9% / +1.5%)	*(+3.1% / +2.1%)*	*(+3.7% / +0.7%)*	(+50.0% / -2.3%)	*(+2.9% / +2.2%)*	(+1.8% / +0.7%)
	NMF	0.2557 ☆✙	0.4188	**0.3641**	*0.0348* ☆	0.1012	*0.0983*
		(+5.8% / +0.6%)	(+3.0% / +2.0%)	**(+3.9% / +0.9%)**	*(+51.3% / -1.4%)*	(+1.5% / +0.8%)	*(+2.2% / +1.1%)*
	LDA	**0.2608** ☆✙	**0.4197**	0.3629	**0.0353** ☆	**0.1052**	**0.101**
		(+7.9% / +2.6%)	**(+3.2% / +2.3%)**	(+3.6% / +0.6%)	**(+53.5% / +0.0%)**	**(+5.5% / +4.8%)**	**(+5.0% / +3.9%)**
GOV	QL-DIR	0.2344	0.3942	0.5141	0.0183	0.1214	0.0353
	TM-CX	0.2449	0.4021	0.5191	0.0216	0.1225	0.0371
	TM-WE	*0.2515* ☆✙	*0.4101*	*0.5329*	*0.0276* ☆	*0.1278*	*0.0383*
		(+7.3% / +2.7%)	*(+4.0% / +2.0%)*	*(+3.7% / +2.6%)*	*(+50.8% /+27.8%)*	*(+5.2% / +4.3%)*	*(+8.9% / +3.2%)*
	NMF	0.2467 ☆	0.4082	0.5289	0.0232 ☆	0.1227	0.0379
		(+5.2% / +0.7%)	(+3.6% / +1.5%)	(+2.9% / +1.9%)	(+26.8% / +7.4%)	(+1.0% / +0.1%)	(+7.4% / +2.2%)
	LDA	**0.2539** ☆✙	**0.4131**	**0.5365**	**0.0296** ☆✙	**0.1327**	**0.0394**
		(+8.3% / +3.7%)	(+4.8% / +2.7%)	(+4.4% / +3.3%)	(+61.7% / +37.0%)	(+9.3% / +8.3%)	(+11.6% / +6.2%)

☆ and ✙ indicate statistically significant improvement in terms of MAP ($p<0.05$) using Wilcoxon signed rank test over the QL-DIR and TM-CX baselines, respectively. Best result for each metric is bolded, second best is italicized.







Conclusions:

1. We performed a comparative study of retrieval effectiveness of document expansion methods based on **different types of translation models** with the ones based on **dimensionality reduction techniques**, such as **topic models** and **matrix decomposition**, on publicly available collections of different size and type.
2. We found out that, although **LDA-based** document expansion generally outperforms document expansion methods based on **NMF** and **translation models**, its performance is comparable to document expansion using **translation model** estimated based on **word embeddings**.