Université de Montréal at TREC 2013: Experiments with Quantum Language Models in the Web Track

Alessandro Sordoni, Wei Yuan and Jian-Yun Nie
{sordonia,yuanwei,nie}@iro.umontreal.ca
DIRO, Université de Montréal
Montréal, H3C 3J7, Québec

1. INTRODUCTION

In TREC 2013, we focus on addressing the challenges posed by the Web track using our recently proposed Quantum Language Modeling (QLM) approach for IR [1]. QLM can be considered as a *dependence* model for IR for its capability of representing and integrating compound term dependencies into the scoring function. Among the main properties of the model, two of them make it stand out from the literature of existing dependence models (such as MRF [3]). First, QLM does not combine scores obtained from matching single terms and from matching compound dependencies, which makes it virtually *parameterless*. This is quite an appealing property for an IR system, especially when a new dataset such as ClueWeb12 is released and no previous training examples can be leveraged to fine-tune important parameters. The second peculiar feature of our model is its ability to automatically *fallback* onto the baseline bag-of-words score in the case that the required dependence relationship does not hold in the document. This is expected to bring improved *robustness* w.r.t. the baseline ranking. In the light of these considerations, the Web Track ad-hoc and robustness task seem the perfect testbeds for our model. In what follows we briefly review some of the theoretical background of QLM before delving into the description of the submitted runs and obtained results.

2. QUANTUM LANGUAGE MODELING

The main feature of QLM is the introduction of a brand new document and query representations generalizing the classical unigram language model representation. The adjective "quantum" is not meant to give an attractive look at the model but it denotes the mathematical inheritance of such a representation. Indeed, documents and queries are associated to a mathematical object which is well-known in quantum physics and is called *density matrix*. In [2], density matrices are shown to be the proper mathematical tool in order to generalize both vector space and language models for IR. From a linear algebra perspective, a density matrix ρ is symmetric, positive-semidefinite matrix of unitary trace, $\rho_q \in \mathcal{S}_+^n = \{\rho : \rho \in \mathbb{R}^{n\times n}, \rho = \rho^T, \rho \succeq 0, \mathrm{tr}(\rho) = 1\}$. It is a generalization of a classical unigram distribution in the sense that classical discrete probability weights can be arranged in the diagonal of the matrix. The off-diagonal entries (taking into account the symmetry) are additional parameters which do not appear in classical unigram models and are used to capture richer informations about text excerpts. The general idea is that density matrices spread a *generalized* probability measure μ onto the manifold of rank-one projectors $\mathcal{P}_1^n = \{uu^T : u \in \mathbb{R}^n, \|u\|_2 = 1\}$. The measure is called *generalized* because its integral over $\mathcal{P}^n$ does not sum to unity. However, if one have a set of projectors $\{P_1, \ldots, P_n\}$, $P_i \in \mathcal{P}_1^n$, for which $\sum_i P_i = I_n$, where I_n is the identity matrix in $\mathbb{R}^{n\times n}$, then $\sum_i \mu(P_i) = 1$. Hence, μ reduces to an ordinary probability measure over a complete set of projectors.

The core idea of the QLM approach is to *embed* in a projector any chosen syntactic expression one would like to take into account. This operation is defined by a mapping from a vocabulary of syntactic expressions (such as terms, bigrams or proximity features) to the set of projectors $\mathcal{P}_1^n$. Differently from the MRF model [3] and its successors [4], single terms or compound terms are not considered as atomic orthogonal entries in a *concept* vocabulary but automatically inherit the metric g of the embedding space $\mathbb{R}^{n\times n}$. For example, one can compute the similarity between two terms, or between a term and a given compound dependency by simply taking the trace of the corresponding projectors, i.e. $g(a, b) = \mathrm{tr}(P_a^T P_b) = \mathrm{tr}(P_a P_b)$. Given that each visible syntactic expression is embedded in a projector, one can efficiently represent a document as a sequence of projectors which, by assuming classical independence, can be turned into a *bag-of-projectors* representation, very similarly to [5]. By approximately maximizing the likelihood of the observed projectors, one can find the best estimator $\hat{\rho}$ for a given document model (or a query). Given the estimated document and query models, the ranking is done by computing the negative of a divergence on $\mathcal{S}_+^n$, namely the negative query to document von Neumann divergence. Given appropriate mappings, one can show that QLM is a proper generalization of the unigram LM approach to IR and reduces to unigram LM score when no compound concept is detected in a given document.

Overall, three desirable properties arise from such an approach: (1) ρ can be considered as a holistic document model capable of encoding occurrence information arising from compound and single terms in a principled way, (2) it efficiently removes any combination parameter (for example for combining single terms and compound term scores) from the scoring function [1] and (3) it falls back to a classical unigram LM score if any of the considered compound concepts is not detected in the document model.

3. QUERY EXPANSION WITH QLM

Query expansion with QLM was not proposed in the original QLM paper [1] but can be promptly introduced for the adhoc task addressed here. The idea is a straightforward

Run	ERR-IA@20	α-nDCG@20	P-IA@20	nDCG@20	ERR@20
TREC median	0.4675	0.5701	0.2880	0.1738	0.098
udemQlml1	0.4701	0.5531	0.3147	0.2286	0.1312
udemQlml1Fb	0.4115	0.4946	0.2966	0.2074	0.1144
udemQlml1FbWiki	**0.4799**	**0.5758**	**0.3425**	**0.2541**	**0.1515**

Table 1: Ad-hoc task.

generalization of query expansion in the classical LM framework: one smooths the original query model ρ_O with an expanded model ρ_L which is supposed to encode the *latent* aspects of the user information need and is simply obtained by selecting relevant terms in the top-K retrieved documents (for example using a Relevance Model [6]). The amount of smoothing is determined by a parameter λ as follows:

$$\rho_E = \lambda\, \rho_O + (1 - \lambda)\, \rho_L, \tag{1}$$

where ρ_E indicates the obtained expanded model. These operations are legit when manipulating density matrices as the set $\mathcal{S}_+^n$ is convex.

4. EXPERIMENTS

The english portion of the ClueWeb12 corpus (Category A) was indexed using the Indri toolkit[1]. All the parameters described next were chosen on the basis of preliminary experiments conducted upon the ClueWebB Web Track 10-11-12 queries. Both the index and the queries were stopped using the standard INQUERY stoplist and no stemming was performed. All the retrieval experiments were performed using our modified version of Indri, with a built-in version of QLM. As described in the original QLM paper [1], we decided to consider the powerset of query terms as useful compound concepts to capture. We consider that such compound concepts are expressed in a document if their component terms appear unordered in a window of adaptive length $L = l|\kappa|$, where $|\kappa|$ is the number of terms in the proximity expression. We set $l = 1$, which was found to optimize ERR@20 and NDCG@20 in preliminary experiments. For example, for the query `usda food pyramid`, we submit the following query expression to our modified version of Indri:

```
#q(usda food pyramid
   #uw2(usda food)
   #uw2(food pyramid)
   #uw2(usda pyramid)
   #uw3(usda food pyramid))
```

Notice that in QLM, no parameters are needed for the combination of unordered and single term scores. We further extended Indri's query language in order to run expanded queries. The modified syntax goes as follows:

```
#qweight(0.8 #q(usda food pyramid
             #uw2(usda food)
             #uw2(food pyramid)
             #uw2(usda pyramid)
             #uw3(usda food pyramid))
      0.2 #qweight( 0.8 health 0.2 nutrition ))
```

where `health` and `nutrition` are considered expansion terms with their respective probability weights.

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

In [1], it is shown that the QLM estimation process weakly suffers the metric divergence problem. Hence, we choose to avoid early-stopping and run the estimation algorithm until the relative change in the likelihood between iterations dropped below a threshold $\epsilon = 0.001$. Spam-filtering was applied on the entire ClueWeb12A corpus using the publicly available Waterloo Spam Ranking for the ClueWeb12 Dataset. We filter out the bottom 30% of the documents, as determined by the spam ranking. This threshold was found to optimize ERR@20 and NDCG@20 in our preliminary experiments with the ClueWebB queries. If compared to the standard TREC setting of filtering out the bottom 70% of the documents, our spam-filtering choice is more risk-inclined. However, we found that our model is quite robust to spam.

5. AD-HOC TASK

In this section we compare between the three runs submitted to the ad-hoc task of the TREC 2013 Web Track.

5.1 Description of the Runs

The description of the three runs is as follows:

- udemQlml1 is a "vanilla" run of QLM with the parameter settings described above. The purpose of this run was to evaluate the effectiveness of the retrieval approach on a single-pass batch retrieval setting.
- udemQlml1Fb performs query expansion using RM3 [7]. We considered the top $K = 10$ retrieved documents obtained by udemQlml1 and set the smoothing parameter $\lambda = 0.8$.
- udemQlml1FbWiki performs query expansion using expansion terms from Wikipedia pages. To this end, we indexed the 2009 Wikipedia dump and performed a run of QLM. We extracted expansion terms from the top $K = 5$ retrieved documents and set the smoothing parameter $\lambda = 0.6$.

5.2 Results

Table 1 compares the retrieval performance of these runs for the ad-hoc task. Our baseline run stands above median values for ERR@20 and nDCG@20. It is interesting to see that even if we do not explicitly favour diversity, our baseline run generally alignes to the median performance for diversity-oriented metrics and outperforms the median value for P-IA@20. The expansion from the top-K retrieved documents from the Web collection fails to improve performance due to the noisy nature of the retrieved set. This result is in-line with past results trying to apply RM3 on Web collections. However, in our preliminary experiments, such kind of expansion showed positive effects on the ClueWeb09B collection when used with the QLM approach. The expansion from Wikipedia pages has a significant positive impact

Alpha=0	ERR-IA@20	α-nDCG@20	P-IA@20	nDCG@20	ERR@20
TREC Median	0.1152	0.1189	0.0275	0.0057	0.0018
udemQlml1R	0.1178	0.1019	0.0542	0.0605	0.0349
udemQlml1FbR	0.1191	0.0950	0.0473	0.0209	0.0187
udemQlml1FbWikiR	**0.1276**	**0.1246**	**0.0820**	**0.0859**	**0.0552**

Alpha=1	ERR-IA@20	α-nDCG@20	P-IA@20	nDCG@20	ERR@20
TREC Median	0.0742	**0.0842**	-0.0066	-0.0306	-0.0223
udemQlml1R	0.0531	0.0432	0.0190	0.0292	0.009
udemQlml1FbR	**0.0843**	0.0524	0.0259	-0.0076	0.0008
udemQlml1FbWikiR	0.0678	0.0644	**0.0389**	**0.0522**	**0.0343**

Alpha=5	ERR-IA@20	α-nDCG@20	P-IA@20	nDCG@20	ERR@20
TREC Median	-0.0896	**-0.0543**	-0.1429	-0.1757	-0.1185
udemQlml1R	-0.2052	-0.1915	-0.1214	-0.0959	-0.0944
udemQlml1FbR	**-0.0549**	-0.1181	**-0.0600**	-0.1216	-0.0706
udemQlml1FbWikiR	-0.1712	-0.1764	-0.1336	**-0.0826**	**-0.0500**

Alpha=10	ERR-IA@20	α-nDCG@20	P-IA@20	nDCG@20	ERR@20
TREC Median	-0.2943	**-0.2275**	-0.3133	-0.3571	-0.2388
udemQlml1R	-0.5284	-0.4849	-0.2971	-0.2523	-0.2237
udemQlml1FbR	**-0.229**	-0.3312	**-0.1671**	-0.2641	-0.1599
udemQlml1FbWikiR	-0.4701	-0.4775	-0.3483	**-0.2510**	**-0.1544**

Table 2: Robustness Task with different values of α.

on the retrieval performance for all the retrieval metrics reported.

6. ROBUSTNESS TASK

For the robustness task, the same runs from the ad-hoc task were submitted (we renamed them by putting "R" to the ends). The only exception is that udemQlml1Fb was spam-filtered at 70% (i.e. udemQlml1FbR), which was the standard threshold used to obtain the LM baseline comparison run. Despite the various values of Alpha, all the three runs show to be robust in ERR@20 and nDCG@20, even though we did not specifically tune our methods with respect to robustness measures. When Alpha is greater than 1, however, all our runs suffer significant losses in α-nDCG. More experiments are needed to understand such behaviours. Another interesting finding is that spam-filtering at 70% is effective, which helps udemQlml1FbR standing above the median values of ERR-IA@20 and P-IA@20 as Alpha increases. We argue that if the same spam-filtering level was applied to the other two runs, even larger improvements with respect to median values could have been reported.

7. CONCLUSIONS

In TREC 2013, we participated to the Web track in order to test the effectiveness and efficiency of our novel ranking model. Results showed that a simple single-pass run by setting a very tight unordered window for capturing compound dependencies (1) outperforms median values in terms of ERR@20 and nDCG@20 and most importantly (2) offers a complexity comparable to a simple bag-of-word approach due to the limited window size. Overall, obtained results suggest that our model could be used as a powerful single-pass retrieval model or as a valuable additional feature in more complex learning to rank approaches.

8. REFERENCES

[1] A. Sordoni, J.-Y. Nie and Y. Bengio. Modeling term dependencies with quantum language models for IR. In *Proc. of SIGIR*, pages 653–662, 2013.

[2] A. Sordoni and J.-Y. Nie. A joint look to vector space and language models using density matrices. In *Proc. of QI*, 2013.

[3] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proc. of SIGIR*, pages 472–479, 2005.

[4] M. Bendersky and W. B. Croft. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proc. of SIGIR*, pages 941–950, 2012.

[5] S. Clinchant and F. Perronnin. Textual Similarity with a Bag-of-Embedded-Words Model. In *Proc. of ICTIR*, pages 25, 2013.

[6] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of SIGIR*, pages 120–127, 2001.

[7] Y. Lv and C. X. Zhai. A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proc. of CIKM*, pages 1895–1898, 2009.