A Concept Language Model for Ad-hoc Retrieval

Bin Zou, Vasileios Lampos, Shangsong Liang,
Zhaochun Ren, Emine Yilmaz, Ingemar Cox
Department of Computer Science, University College London, United Kingdom
{bin.zou.14, v.lampos, shangsong.liang, zhaochen.ren, emine.yilmaz, i.cox}@ucl.ac.uk

ABSTRACT

We propose an extension to language models for information retrieval. Typically, language models estimate the probability of a document generating the query, where the query is considered as a set of independent search terms. We extend this approach by considering the concepts implied by both the query and words in the document. The model combines the probability of the document generating the concept embodied by the query, and the traditional language model probability of the document generating the query terms. We use a word embedding space to express concepts. The similarity between two vectors in this space is estimated using a weighted cosine distance. The weighting significantly enhances the discrimination between vectors. We evaluate our model on benchmark datasets (TREC 6–8) and empirically demonstrate it outperforms state-of-the-art baselines.

Keywords

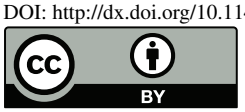
Language model; Ad-hoc retrieval; Word embeddings

1. INTRODUCTION

A core task in information retrieval is to judge whether documents are relevant to a given query. The traditional Query Likelihood (QL) language model makes the assumption that both query and documents are bag-of-words and retrieves documents according to the likelihood of observing a query given the document's language model [5]. In the case where a query term is not present in a document, smoothing strategies are applied based on the statistical distribution in the overall collection. Improvements to these smoothing strategies incorporate topic modelling techniques. For early topic models, document specific language models were produced by projecting both queries and documents to the same latent semantic space such that different words that are semantically close can be easily identified [2, 8]. However, these approaches rely on a predefined number of topics.

More recently, distributed representations of words, or *word embeddings*, have been used to capture various latent language characteristics, such as syntax, topics, semantics and spelling [3]. Language models have incorporated word embeddings in order to improve retrieval precision [1, 4, 6, 9].

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
DOI: http://dx.doi.org/10.1145/3041021.3054209



In this paper, we propose a Concept Language Model (CLM) that uses word embeddings. The model considers both (i) the probability of the concept embodied by the query, c_q given the concept(s) embodied in the document, c_d, together with (ii) the traditional language model probability of the document d generating the query q. This latter probability also incorporates a word embedding that serves the function of term expansion.

2. CONCEPT LANGUAGE MODEL

At the highest level, CLM is formulated as

$$\hat{p}(q|d) = (1-\beta)\hat{p}(c_q|c_d) + \beta \prod_{t \in q} \hat{p}(t|d)\,, \tag{1}$$

where the probability of a document d generating a query q is a weighted combination of (i) the probability that the concept, c_q embodied by q, is generated by the concept(s), c_d, of the document, and (ii) the probability, $\hat{p}(t|d)$, of the document generating the individual terms, t. The parameter β controls the relative weight of the two components. We now describe each component in detail. Note that the $\hat{}$ symbol on p is used to stress that the model is estimated.

A query q consists of a number of terms, $t_1 \dots t_n$. These terms have associated concepts $c_{t_1} \dots c_{t_n}$. Assuming independence of terms and concepts, the traditional QL model considers the probability of a document generating each term. If the term is absent from the document, smoothing based on the collection statistics is used. More recently, by incorporating the concept implied by the query term and inferred from the embedding space, a form of term (query) expansion can also be achieved [1, 4, 6, 9]. We assume that the concept(s) embodied within a document is the sum of the concepts expressed by the individual words in the document. Thus, the probability of d generating term t is based not only on the empirical frequency of t in d, but also on the probability, $\hat{p}(c_t|c_d)$, that the corresponding concept, c_t, is generated by the document concept c_d. Under the word independence assumption, this latter probability is approximated by the normalized sum of the individual probabilities, $\hat{p}(c_t|c_w)$, of each concept, c_w, implied by each word in d, generating concept c_t. Thus, using Dirichlet smoothing, we have

$$\hat{p}(t|d) = \frac{|d|}{|d|+\mu}\left(\frac{\mathrm{tf}_d(t)}{|d|} + \frac{1}{|d|}\sum_{w \in d}\hat{p}(c_t|c_w)\right) + \frac{\mu}{|d|+\mu}\frac{\mathrm{tf}_{\mathcal{D}}(t)}{|\mathcal{D}|}\,,$$

where tf_d and $\mathrm{tf}_{\mathcal{D}}$ are the term frequencies in the document and collection respectively. The product of the individual probabilities of a document generating each term provides a final probability of a document generating the query. We supplement this with the first term in Eq. 1. We assume that the concept embodied by the query is the product of the concepts embodied by the individual terms, i.e. c_q is equivalent to $c_{t_1} \times \dots \times c_{t_n}$. This is simply a generalization of term independence to concept independence. Similarly, we assume that the concept(s) embodied by a document is the sum of the concepts embodied by the individual words

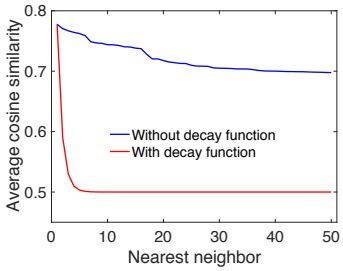


Figure 1: The average cosine similarity of 100 terms between their concepts and their nearest neighbours' concepts in descending order without (blue) and with (red) decay function.

in the document. Then for each word w in d, we estimate the probability, $\hat{p}(c_q|c_w)$ of this concept, c_w generating the query concept. To do so, we first determine the $\hat{p}(c_q|c_w) = \hat{p}(c_{t_1} \dots c_{t_n}|c_w) = \hat{p}(c_{t_1}|c_w) \dots \hat{p}(c_{t_n}|c_w)$. The sum over all words provides the final probability c_d generating c_q. Thus,

$$\hat{p}(c_q|c_d) = \frac{1}{|d|} \sum_{w \in d} \hat{p}(c_q|c_w) = \frac{1}{|d|} \sum_{w \in d} \left(\prod_{t \in q} \hat{p}(c_t|c_w) \right) .$$

To estimate the probability $\hat{p}(c_t|c_w)$, we built an embedding space as discussed shortly. The probability, $\hat{p}(c_t|c_w)$, is then given by

$$\hat{p}(c_t|c_w) = \frac{\texttt{sim}(c_t, c_w)}{\sum_{t' \in N_t} \texttt{sim}(c_{t'}, c_w)} ,$$

where the denominator serves to normalise the probability value between 0 and 1, and N_t is a neighbourhood of the closest words to c_t. The similarity function

$$\texttt{sim}(c_t, c_w) = \frac{\cos(c_t, c_w)}{\theta^{r(c_t, c_w)}} ,$$

is the cosine similarity between two vectors c_t and c_w in the embedding space it is transformed to the interval $[0, 1]$ via $(x + 1)/2$ to avoid negative sub-scores. The denominator is a decay function, the purpose of which is described next.

The cosine function is often used to capture the semantic relationship between embedding vectors, but it cannot be directly used in ad-hoc retrieval. This is because, as shown in Figure 1, there are no substantial differences between the cosine similarity of the most similar term and the 50th term. To enhance discrimination, we define a monotonic decay function. Given t and the collection vocabulary V, we rank all words in V based on their cosine similarities to t. We denote the rank of c_w with respect to c_t as $r(c_t, c_w)$. The term $\theta > 1$ is a constant decay factor. In practice, we only consider the 50 nearest neighbours of query term t in V and denote them as N_t. Since we have a fixed vocabulary, most computations such as cosine similarities can be pre-computed. According to the red line in Figure 1, the similarities between t and words in V can be easily discriminated. Note that we also tried a sigmoid function, but it provided worse performance.

3. EXPERIMENTS

To evaluate CLM, we work with the TREC collection (Disks 4-5), used in TREC 6, 7 and 8 for the ad-hoc retrieval tracks, which contains 150 queries. To build the index of the collection, we apply tokenization, stemming, and stop-word removal. We evaluate performance using MAP and precision at 10 and 20. Statistical significance of observed differences between two comparisons is assessed using a two-tailed paired t-test and the significance level is set to $p < 0.05$.

CLM is compared to 5 baselines: the traditional QL model [5], LLM [8], BM25 [7], GLM [1] and EQE [9]. Word embeddings are trained using word2vec [3] on TREC collections. The skip-gram model with negative sampling is employed. We set the window size to 10, and the dimension of embedding vectors to 300. The Dirichlet smoothing parameter μ is set to 1,500 and the decay factor θ is set to 3. The smoothing parameter β is set to 0.7. Note that optimal parameters are chosen via 2-fold cross validation.

Table 1: Comparing CLM against the baselines.

Dataset	Metric	QL	LLM	BM25	GLM	EQE	CLM
TREC 6	MAP	0.213	0.219	0.214	0.228	0.234	**0.249**
	P@10	0.400	0.398	0.403	0.409	0.411	**0.424**
	P@20	0.330	0.333	0.335	0.337	0.342	**0.351**
TREC 7	MAP	0.177	0.164	0.171	0.195	0.198	**0.209**
	P@10	0.400	0.385	0.390	0.418	0.416	**0.425**
	P@20	0.321	0.313	0.316	0.342	0.344	**0.349**
TREC 8	MAP	0.232	0.242	0.243	0.250	0.251	**0.266**
	P@10	0.429	0.436	0.435	0.442	0.446	**0.457**
	P@20	0.382	0.397	0.387	0.406	0.412	**0.422**

The performance of CLM and the baselines is shown in Table 1. In terms of MAP, CLM statistically significantly outperforms all the baselines on all the datasets. In addition, as can be seen in the table, CLM which integrates a decay function outperforms the models, i.e. GLM and EQE, that do not use one. This demonstrates the effectiveness of the proposed decay function in word embedding language models. The CLM achieves an improvement of 7.60% over GLM and 5.98% over EQE, in terms of MAP.

4. DISCUSSION AND CONCLUSION

In this paper, we proposed a concept language model using word embeddings. The model estimates the probability of a document generating the individual terms in the query, and the probability of the document's concepts generating the query concept. These two probabilities are weighted and summed. The CLM requires estimating the probability that a concept c_w implied by word w in the document generates a concept c_t implied by a term t in the query. This probability is estimated based on the cosine distance between the two vectors c_w and c_t in the embedding space normalized by a decay function that aimes to make the probabilities more discriminatory. The CLM was evaluated on the ad-hoc tasks of TREC 6, 7 and 8, and was shown to significantly outperform state-of-the-art baselines, according to MAP. Future work will focus on learning task-specific embedding vectors. We will also consider combining our model with pseudo-relevance feedback.

5. REFERENCES

[1] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word Embedding Based Generalized Language Model for Information Retrieval. In *Proc. SIGIR*, pages 795–798, 2015.
[2] X. Liu and W. Croft. Cluster-based Retrieval using Language Models. In *Proc. SIGIR*, pages 186–193, 2004.
[3] T. Mikolov and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. NIPS*, pages 3111–3119, 2013.
[4] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving Document Ranking with Dual Word Embeddings. In *Proc. WWW*, pages 83–84, 2016.
[5] J. Ponte and W. Croft. A Language Modeling Approach to Information Retrieval. In *Proc. SIGIR*, pages 275–281, 1998.
[6] N. Rekabsaz, M. Lupu, A. Hanbury, and G. Zuccon. Generalizing Translation Models in the Probabilistic Relevance Framework. In *Proc. CIKM*, pages 711–720, 2016.
[7] S. Robertson and H. Zaragoza. *The Probabilistic Relevance Framework: BM25 and Beyond*. 2009.
[8] X. Wei and W. Croft. LDA-based Document Models for Ad-Hoc Retrieval. In *Proc. SIGIR*, pages 178–185, 2006.
[9] H. Zamani and W. Croft. Embedding-based Query Language Models. In *Proc. ICTIR*, pages 147–156, 2016.