Compression-Based Document Length Prior for Language Models

Javier Parapar[1], David E. Losada[2], Álvaro Barreiro[1]

javierparapar@udc.es, david.losada@usc.es, barreiro@udc.es

[1]Information Retrieval Lab, Computer Science Department, University of A Coruña

[2]Dept. of Electronics and Computer Science, University of Santiago de Compostela

Abstract

The inclusion of document length factors has been a major topic in the development of retrieval models. We believe that current models can be further improved by more refined estimations of the document's scope. In this poster we present a new document length prior that uses the size of the compressed document. This new prior is introduced in the context of Language Modeling (LM) with Dirichlet smoothing. The evaluation performed on several collections shows significant improvements in effectiveness.

1. Introduction

- Doc length (DL) is important in IR, e.g. BM25, pivoted vector space models or Language Models with Dirichlet smoothing, incorporate some form of DL correction. DL corrections are based on rough estimations of the doc's contents (e.g. byte size or term count).
- **Claim:** The size of the compressed doc is appropriate to estimate the doc's scope.
- **Example:** Two docs (d_1, d_2) with the same size, if the compressed size of d_1 is much smaller than the compressed size of $d_2 \Rightarrow d_1$ is more verbose than d_2.
- **Proposal:** Use the size of the compressed document to define a doc prior in LM with Dirichlet smoothing.

2. Compression-Based Prior

Probability of a doc given a query:

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \quad (1)$$

Dirichlet smoothing:

$$P(q|d) = \prod_{i=1}^{n} \frac{tf(q_i,d) + \mu \cdot P(q_i|C)}{|d| + \mu} \quad (2)$$

n: # query terms, $tf(q_i,d)$: raw term frequency of q_i in d, $|d|$: # doc terms, μ: smoothing parameter, $P(q_i|C)$: probability of q_i occurring in the collection C.

$\log P(q|d)$, reduces to :

$$\sum_{i:tf(q_i,d)>0} \log\left(1 + \frac{tf(q_i,d)}{\mu P(q_i|C)}\right) + n \log \frac{\mu}{|d| + \mu} \quad (3)$$

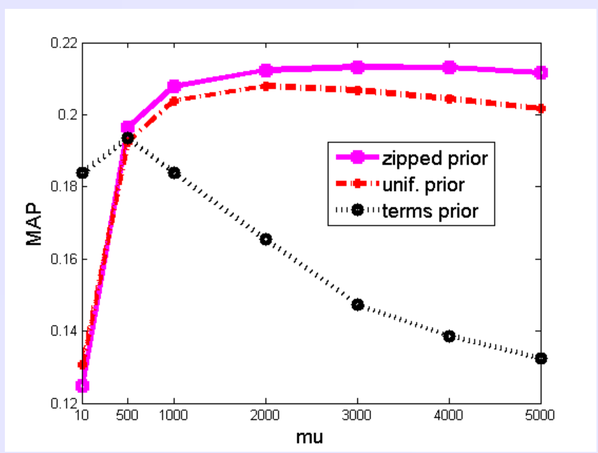
We compare the following non-uniform priors:

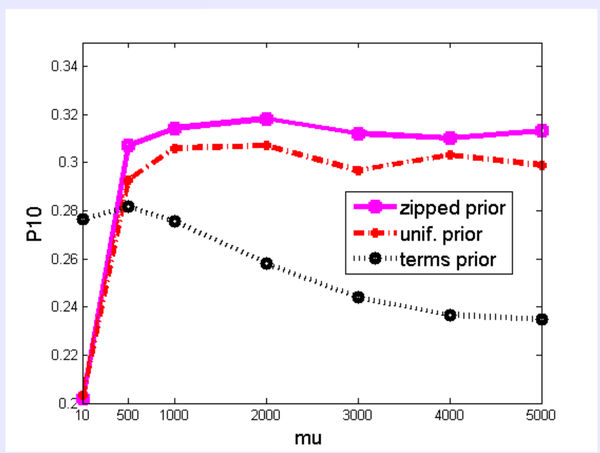
$$\text{terms prior: } P(d) = \frac{|d|}{\sum_{d_i \in C} |d_i|}, \quad \text{zipped prior: } P(d) = \frac{com(d)}{\sum_{d_i \in C} com(d_i)} \quad (4)$$

$com(d)$: size (bytes) of the compressed doc (zipped) divided by the original size (bytes) of the doc.

3. Experiments and Results

- TREC-5, 6 & 8 (50 qs each), WT10g (100 qs).
- Porter stemmer + stopword removal.
- Short queries (title only).
- Reported results for WT10g but the same trends hold in every collection (see Table 1 in the Procs.).





Terms prior is worse than standard Dirichlet (Uniform prior). Zipped prior is the best. The improvement is consistent across smoothing levels.

4. Conclusions

The novel compression-based prior improves significantly the LM Dirichlet model.

The performance improvements are robust across different parameter settings and test collections.

References

[1] S. Robertson, S. Walker, S. Jones, M. HancockBeaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC-3*, 109–127, 1995.

[2] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. SIGIR-96*, 21–29, 1996.

[3] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

[4] D. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11:109–138, 2008.

SIGIR 2009, ACM 32nd Conference on Research and Development in Information Retrieval. July 19-23, 2009. Boston, Massachusetts.