Parsimonious Relevance Models

Edgar Meij Wouter Weerkamp Krisztian Balog Maarten de Rijke
ISLA, University of Amsterdam

Idea

Apply **parsimonization** to re-estimate the term probabilities assigned by **relevance models** in order to make them more sparse. This removes more general terms and emphasizes content-bearing terms.

Approach

Relevance Models

Relevance modeling assumes that query and relevant documents are both samples from an underlying term distribution: the relevance model. To determine the parameters of this model, a set of (pseudo-)relevant documents is used to sample from. We use Lavrenko and Croft's model 2:

$$P_{RM}(t|Q) \propto P(t) \prod_{q \in Q} \sum_{D \in \mathcal{D}_Q} P(q|D) \frac{P(t|D)P(D)}{P(t)}$$

Parametrizing documents

In a language modeling setting, document models are generally estimated using a maximum likelihood approach (ML) and smoothed with a large background model, such as the collection:

$$P(t|D) = \frac{1}{2}\left(\frac{\#(t, D_i)}{\sum_{t'} c(t', D)} + P(t)\right)$$

Parsimonization

Maximum likelihood estimates may emphasize general terms or terms related to the domain of interest that are not pertinent to the query. To assign more probability mass to query-specific terms, we use an EM algorithm to iteratively adjust the term probabilities for each (pseudo-)relevant document. Doing so enables us to emphasize document-specific terms.

E-step: $$e_t = \#(t, D) \frac{\gamma P(t|D)}{(1-\gamma)P(t) + \gamma P(t|D)}$$

M-step: $$P(t|D) = \frac{e_t}{\sum_{t'} e_{t'}}$$

Query modeling

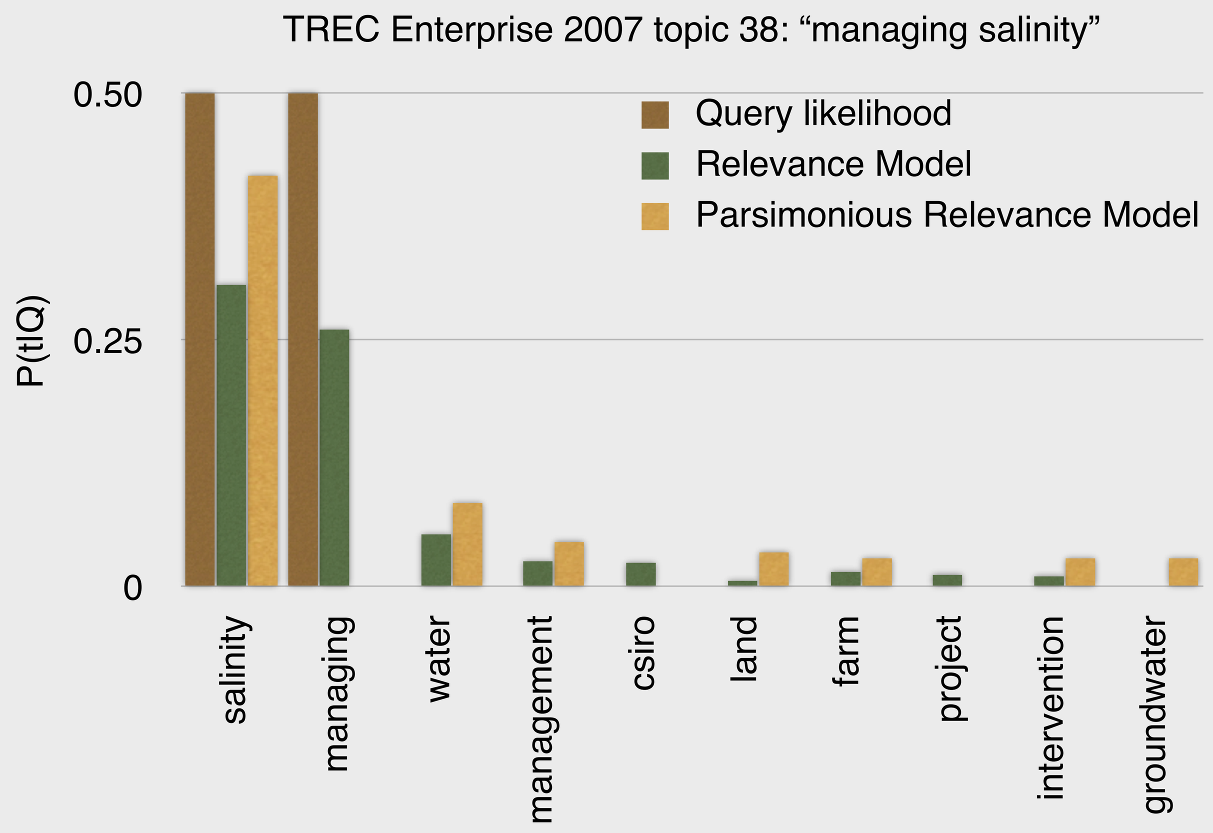
In order to apply our model during retrieval we use it for blind relevance feedback, by interpolating the initial query with the found terms:

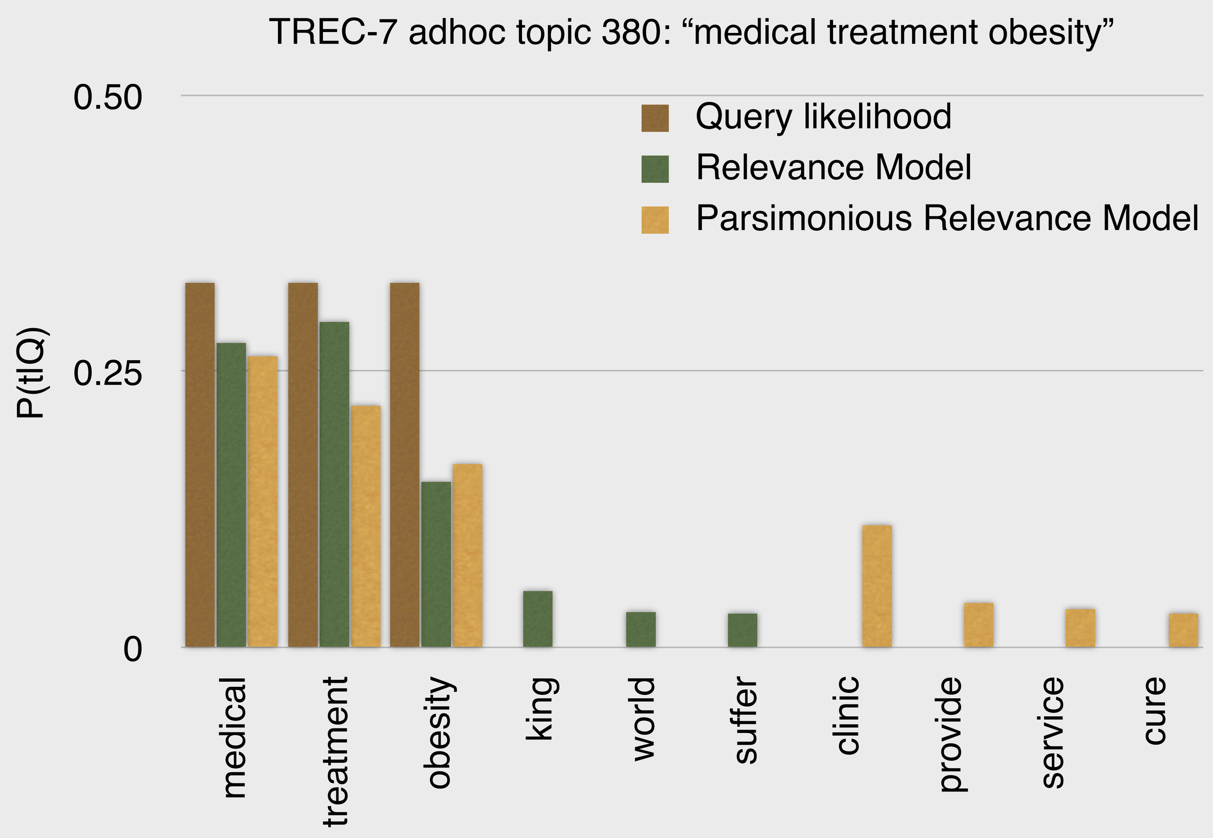
$$P(t|Q) = (1-\lambda)P_{ML}(t|Q) + \lambda P_{RM}(t|Q)$$

Evaluation

We use test collections from six different TREC tracks for our evaluation; **TREC-7 adhoc**, **TREC Robust 2004** (both use TREC disks 4 and 5, minus the congressional record), **TREC Blog 2006**, **TREC Blog 2007**, **TREC Genomics 2007**, and **TREC Enterprise 2007** (document search task, using relevance level 1).

Examples





Introducing new terms

The parsimonious relevance model emphasizes new terms that have a relatively high probability given the (pseudo-) relevant documents with respect to the background model, which are pertinent to the query; "groundwater," "clinic," "provide," "service," and "cure" in these examples.

Removing general terms

Our approach strongly reduces the probability mass of more general terms which occur frequently in the background model; "managing," "medical," and "treatment" in these examples. This also automatically removes stopwords.

Results

The parsimonious relevance model (PRM) is able to outperform a query-likelihood baseline (QL) in terms of **mean average precision** on all evaluated collections. On most collections, it **significantly** outperforms the relevance model based on ML estimates (RM). Our approach **improves early precision** on all collections but one, both compared to the baseline and to relevance models. The statistically significant improvements on TREC Robust 2004 are especially interesting, since this collection is known for its difficulty at handling relevance feedback.

Test collection	Run	MAP	P@10	MRR
TREC-7	QL	0.1642	0.3760	**0.6295**
	RM	0.1747	0.3640	0.5618
	PRM	**0.2091**†/‡	**0.4120**†/‡	0.5662‡
TREC Robust 2004	QL	0.2247	0.3968	0.6098
	RM	0.2430†	0.4056	0.6050
	PRM	**0.2689**†/‡	**0.4289**†/‡	**0.6115**†/‡
TREC Blog 2006	QL	0.3213	**0.6720**	**0.7236**
	RM	0.3313	0.6380	0.6983
	PRM	**0.3379**‡	0.6700‡	0.7206
TREC Blog 2007	QL	0.4327	0.6820	0.7558
	RM	0.4371	0.6780	0.6929
	PRM	**0.4571**‡	**0.7280**‡	**0.7629**
TREC Genomics 2007	QL	0.2695	0.4306	0.6098
	RM	0.2828	0.4389	0.5732
	PRM	**0.2850**	**0.4528**	**0.6196**
TREC Enterprise 2007	QL	0.3552	0.7100	0.8583
	RM	0.4227†	0.6940	0.8304
	PRM	**0.4433**†	**0.7400**‡	**0.8597**

Results per test collection for the baseline query-likelihood run (QL), relevance models (RM), and parsimonious relevance models (PRM) (best results are marked in boldface). †/‡ indicates a statistically significant difference as compared to the baseline or to the RM run respectively, using a two-tailed paired t-test at $p < 0.01$.

Conclusion

We use parsimonious language models to re-estimate term probabilities assigned by relevance models. Parsimonious relevance models (i) **improve retrieval effectiveness in terms of MAP** on all collections, (ii) **significantly outperform their non-parsimonized counterparts** on most measures, and (iii) have a **precision enhancing effect**, unlike other blind relevance feedback methods.

University of Amsterdam
vl-e virtual laboratory for e-science
NWO Netherlands Organisation for Scientific Research
MultiMatch