Are Semantically Coherent Topic Models Useful for Ad Hoc Information Retrieval?

Romain Deveaud[α] – Eric SanJuan[α] – Patrice Bellot[β]

[α] LIA – University of Avignon

[β] LSIS – Aix-Marseille University

opening adoption records

TREC 2004 Robust track, query #679

① querying target collection for retrieving query-oriented documents

② selecting the top-N pseudo-relevant feedback documents

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) \prod_{t \in Q} P(t|\theta_D)$$

③ estimating relevance model with feedback documents

④ smoothing the original query model with the estimated relevance model

$$P(w|\theta_Q) = \lambda\, P(w|\tilde{\theta}_Q) + (1-\lambda) P(w|\hat{\theta}_Q)$$

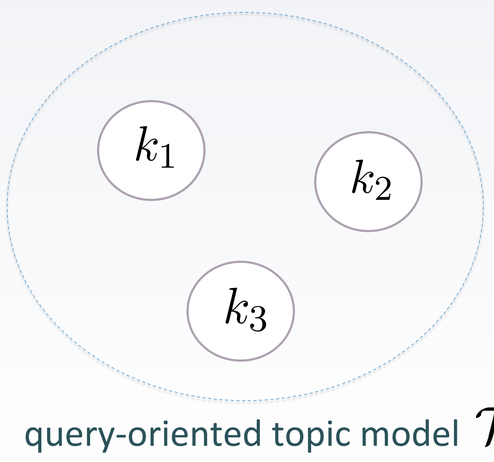
TRADITIONAL RELEVANCE MODEL [1]

PROPOSITION: TOPIC-DRIVEN RELEVANCE MODEL (TDRM)

③ topic modeling (Latent Dirichlet Allocation)

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} \left(P(\theta_D) P(w|\theta_D) P_{TM}(w|D) \prod_{t \in Q} P(t|\theta_D) \right)$$

④ estimating TDRM using a query-oriented topic model



query-oriented topic model $\mathcal{T}_\Theta$

$$P_{TM}(w|D) = \sum_{k \in \mathcal{T}_\Theta} \phi_{k,w} \cdot \theta_{D,k}$$

SEMANTIC COHERENCE & RETRIEVAL EVALUATION

Several measures of topic model coherence emerged from the NLP community [2].
We choose the PMI-based measure, which was demonstrated robust and effective:

$$c(\mathcal{T}_\Theta^K) = \frac{1}{K} \sum_{i=1}^{K} \sum_{(w,w') \in k_i} \log \frac{P(w,w') + \epsilon}{P(w)P(w')}$$

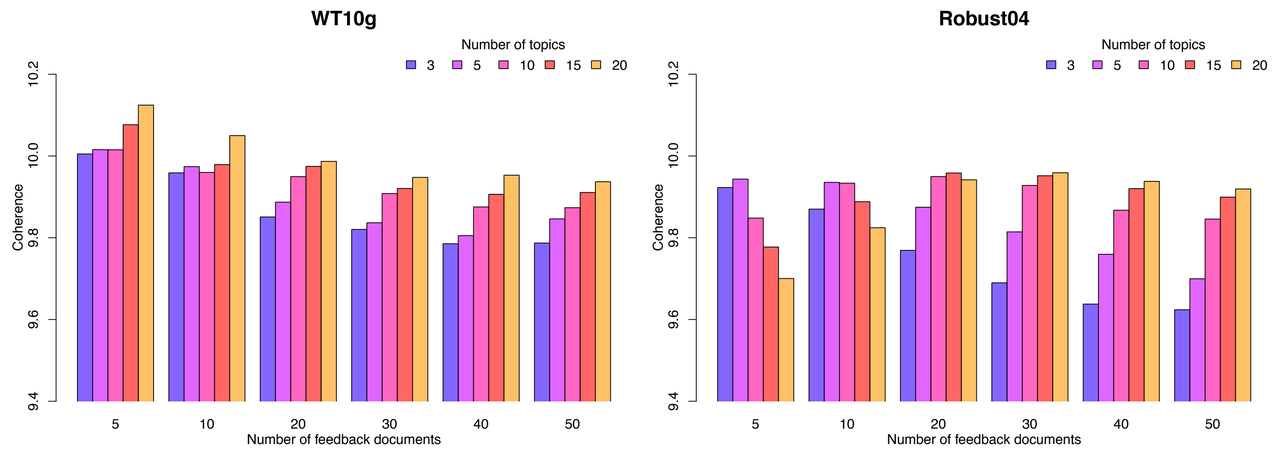


Figure 1: Semantic coherence of the topic models for different values of K, in function of the number N of feedback documents.

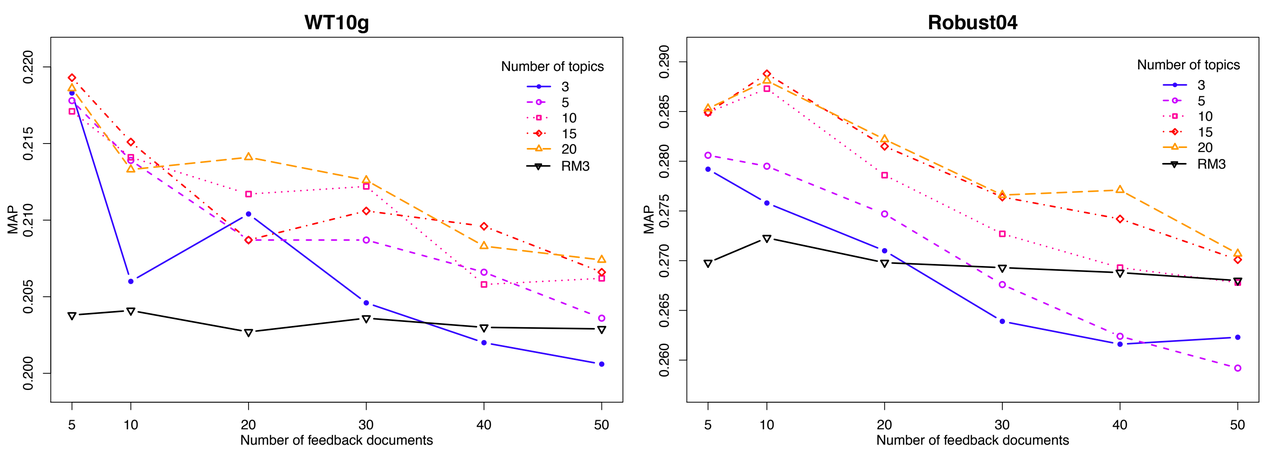


Figure 2: Retrieval performance in terms of Mean Average Precision (MAP) of the TDRM approach. Each line represent a different number of topics K, and the performance are reported in function the number N of feedback documents. The black, plain line represents the RM3 baseline.

More coherent topic models achieve **better retrieval results** when used within the proposed TDRM framework.

However using the **most coherent** topic models do not reach the retrieval performance of the best performing methods.

REFERENCES

[1] V. Lavrenko and W. B. Croft. Relevance-Based Language Models. In *Proceedings of SIGIR*, 2001.

[2] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of EMNLP*, 2012.

Contact: romain.deveaud@gmail.com

This work was supported by the French Agency for Scientific Research (Agence Nationale de la Recherche) under CAAS project (ANR 2010 CORD 001 02).