Estimating Topical Context by Diverging from External Resources

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

[α] LIA – University of Avignon

[β] LSIS – Aix-Marseille University

The New York Times

radiography risks

WIKIPEDIA The Free Encyclopedia

LDC GigaWord

WWW ClueWeb09-B (no spam)

① querying resources separately

$$P(t|\hat{\theta}_{\mathcal{R}}) \propto \sum_{D_F \in \mathcal{R}_Q} P(Q|\theta_{D_F}) \left(-\sum_{w \in t} P(w|\theta_{D_F}) \log P(w|\theta_{D_F}) \right)$$

② estimating a query model (or topical context) for each resource

③ computing KL divergence between each resource and target documents

$$D(\hat{\theta}_R||\theta_D) = -\sum_{t \in V} P(t|\hat{\theta}_{\mathcal{R}}) \log P(t|\theta_D)$$

④ document ranking, mixture of resource models (following [1])

$$s(Q,D) = \lambda \ \log P(Q|\theta_D) - (1-\lambda) \sum_{\mathcal{R} \in \mathcal{S}} \varphi_{\mathcal{R}} \cdot D(\hat{\theta}_{\mathcal{R}}||\theta_D)$$

Estimating Relevance Models using several sources of information was successfully studied in [1] and [2], but with a limited amount of sources and no exploration of relative performance improvements.

EXPERIMENTATIONS & RESULTS

	QL		RM3		MoRM		DfRes	
	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20
wt10g	0.2026	0.2429	0.2035	0.2449	$0.2339^{\alpha,\beta}$	$0.2833^{\alpha,\beta}$	$0.2463^{\alpha,\beta}$	$0.2954^{\alpha,\beta}$
robust	0.2461	0.3528	0.2727^{α}	0.3677	$0.2869^{\alpha,\beta}$	$0.3799^{\alpha,\beta}$	$0.3147^{\alpha,\beta,\gamma}$	$0.4024^{\alpha,\beta,\gamma}$

Table 1: Document retrieval results reported in terms of Mean Average Precision and Precision at 20 documents. We use a two sided paired wise t-test to determine significant differences over baselines. α, β and γ indicate statistical improvements over QL, RM3 and MoRM respectively, with $p < 0.05$.

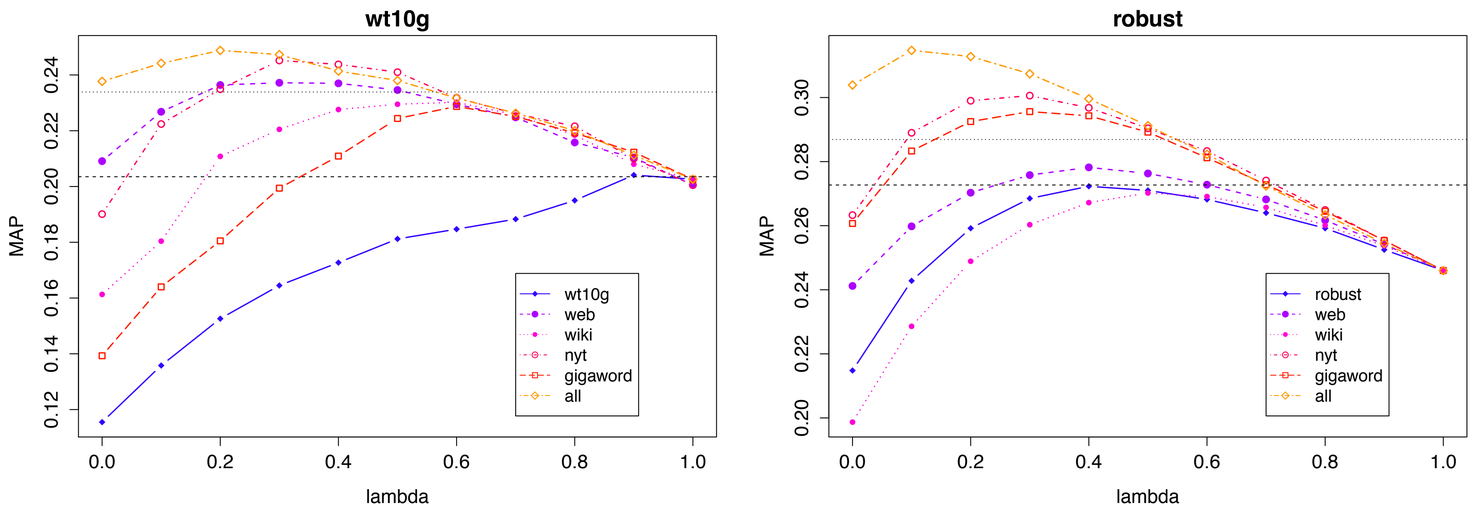

Figure 1: Retrieval performance (in MAP) as a function of the λ parameter. The DfRes results reported in Table 1 are depicted by curve "all", while all other curves correspond to DfRes with a single resource. Baselines are shown for reference: dashed lines represent RM3 and dotted lines represent MoRM.

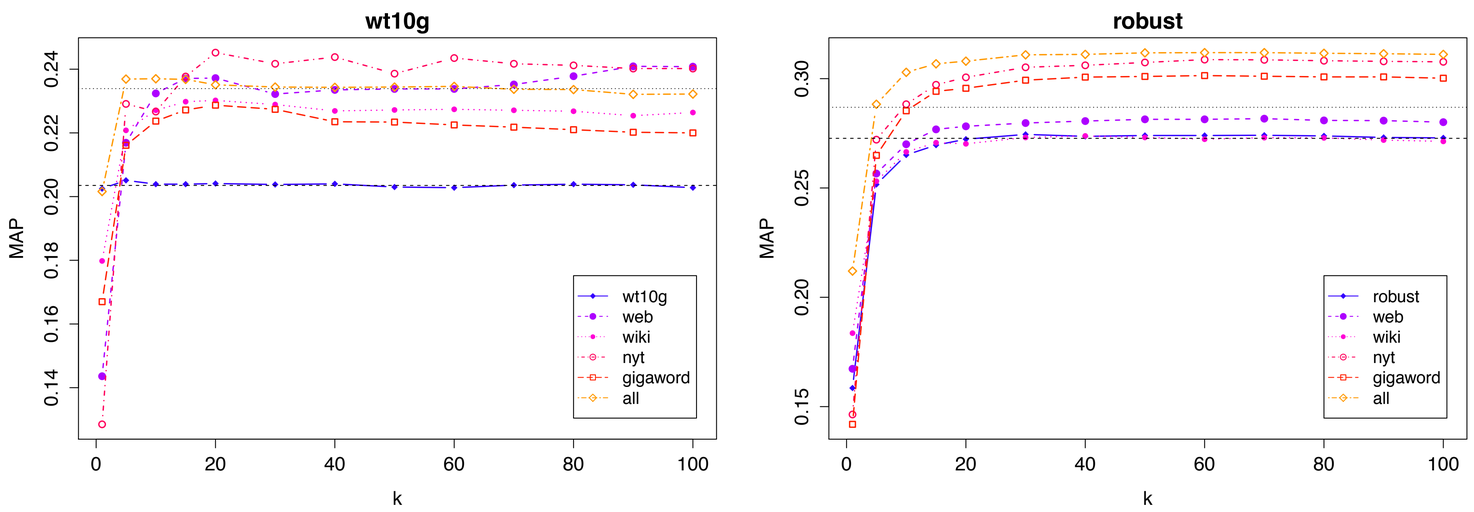

Figure 2: Retrieval performance (in MAP) as a function of the number of terms k used for estimating the resource language model. Legend is the same as in Figure 1.

Results show support for the principle of *polyrepresentation* [3].

Using only the estimated query model (i.e. setting λ = 0) achieves better results than all baselines.

No substantial differences between 15 and 100 terms.

	nyt	wiki	gigaword	web	robust	wt10g
wt10g	0.303	0.162	0.121	**0.313**	-	0.101
robust	**0.309**	0.076	0.281	0.149	0.185	-

Table 2: $\varphi_{\mathcal{R}}$ weights learned for resources on the two collections. We averaged weights over all queries.

Entropy allows to select multi-word terms.
Ex: « arteriovenous malformations »

FUTURE WORK

Iterative query modeling (adding a resource at a time if it is likely to improve performance).

REFERENCES

[1] M. Bendersky, D. Metzler, and W. B. Croft. Effective query formulation with multiple information sources. In *Proceedings of WSDM*, 2012.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, 2006.

[3] P. Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of SIGIR*, 1994.

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