Reading : Bayesian Query-Focused Summarization

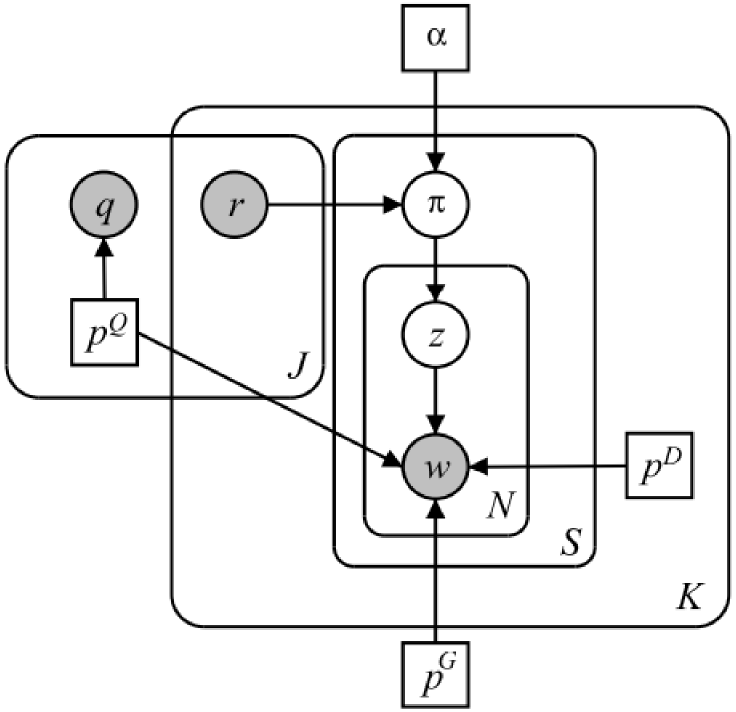
Yi-Chen Lin

Apr 29, 2013

What is summarization?

- Given a (set of) document(s), find a sentence to summarize the content(s)
- Query-focused summarization
  - $Q_{1:J}$: queries
    - title, summary, narrative, concepts (key words)
    - title provide much information
  - $D_{1:K}$: documents
  - $r_{J \times K}$: relevance judgments

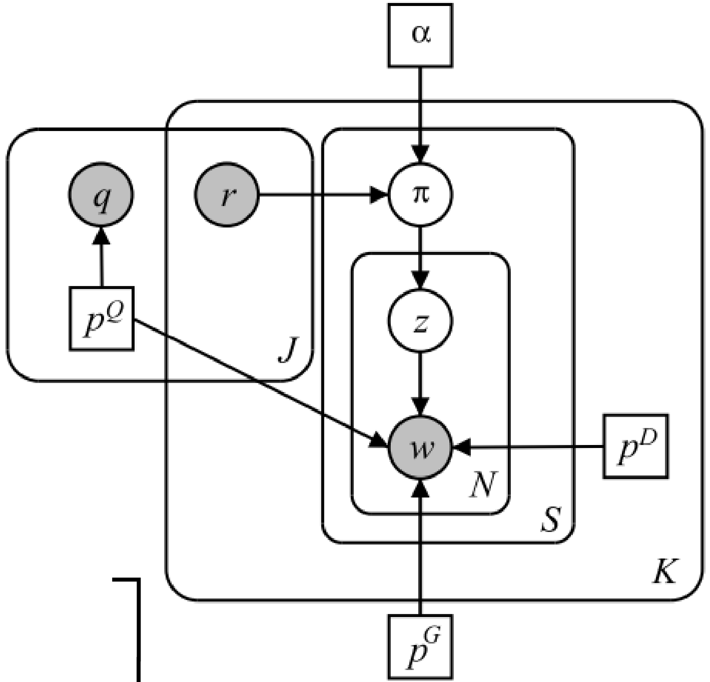
BayeSum



- Given: q, r, w
- $z_{k,s,n} \sim Mult(z|\pi_{k,s})$
  - distribution indicator
- $\pi_{k,s} \sim Dir(\pi_{k,s}|\alpha) r_k(\pi_{k,s})$
  - $r_k(\pi)$
    - Constraint for limiting words to generate only from specific distribution (pq_j, pd_k, pG)

BayeSum

- $$p(q_{1:J}, r, d_{1:K}) = \left[\prod_j \prod_n \mathrm{p}^{q_j}(q_{jn})\right] \times \left[\prod_k \prod_s \int_\Delta \mathrm{p}(\pi_{ks} \mid \alpha, r) d\pi_{ks} \times \prod_n \sum_{z_{ksn}} p(z_{ksn} \mid \pi_{ks}) p(w_{ksn} \mid z_{ksn})\right]$$



Expectation Propagation

- Used to construct tractable approximation probability distribution in the form of

$$p(x) \propto \prod_i \psi_i(x)$$

  - Most of exponential family are in this form

- Pros/Cons
  - Global approximation, better than ADF
  - But may not converge

Expectation Propagation

- Use $q(x; \theta)$ to approximate $p(x) \propto \prod_i \psi_i(x)$
- Assumed Density Filter approximation
  - $q(x; \theta^i) \cong \prod_j^i \psi_j(x)$
  - $\theta^i = \text{argmin}_\theta KL(\psi_i(x)\, q(x; \theta^{i-1}) \,||\, q(x; \theta))$
  - The order of approximation matters

EP: the algorithm(1)

- Initialization
  - $\forall$i, $m_i(x)=1$, $q(x;\theta) \propto \prod_i m_i(x)$
  - $m_i(x) = \dfrac{q(x;\theta^*)}{q(x;\theta^{-i})}$
    - An update ratio of q according to $\psi_i(x)$

EP: the algorithm(2)

- Iteration
  1. Choose some i
  2. $q(x;\theta^{-i}) \propto \dfrac{q(x;\theta)}{\mathrm{m}_i(x)}$ is normalized
  3. $\theta^* = \mathrm{argmin}_\theta \mathrm{KL}(\psi_i(x)\, q(x;\theta^{-i}) \,||\, q(x;\theta))$
  4. $\mathrm{m}_i(x) = \dfrac{q(x;\theta^*)}{q(x;\theta^{-i})}$

Measurement of the experiment

- Precision
  - {relevant & retrieved} / {retrieved}
- R-prevision
  - Precision at **R**-th position in the ranking of results for a query that has **R** relevant documents
- P@n:
  - Precision which only cares about the top n ranked results
  - (P@2)
- MAP: mean average precision
  - $\{ \Sigma_q \text{AveP}(q) \} / Q$
  - AveP(q) = average P@k for a query q, where $k \in \{k\text{: rank k doc is relevant}\}$
- MRR: mean reciprocal rank
  - $\{ \Sigma_q 1/\text{rank}_q \} / Q$, where rank_q is the rank of correct result of query q

Experiment Result

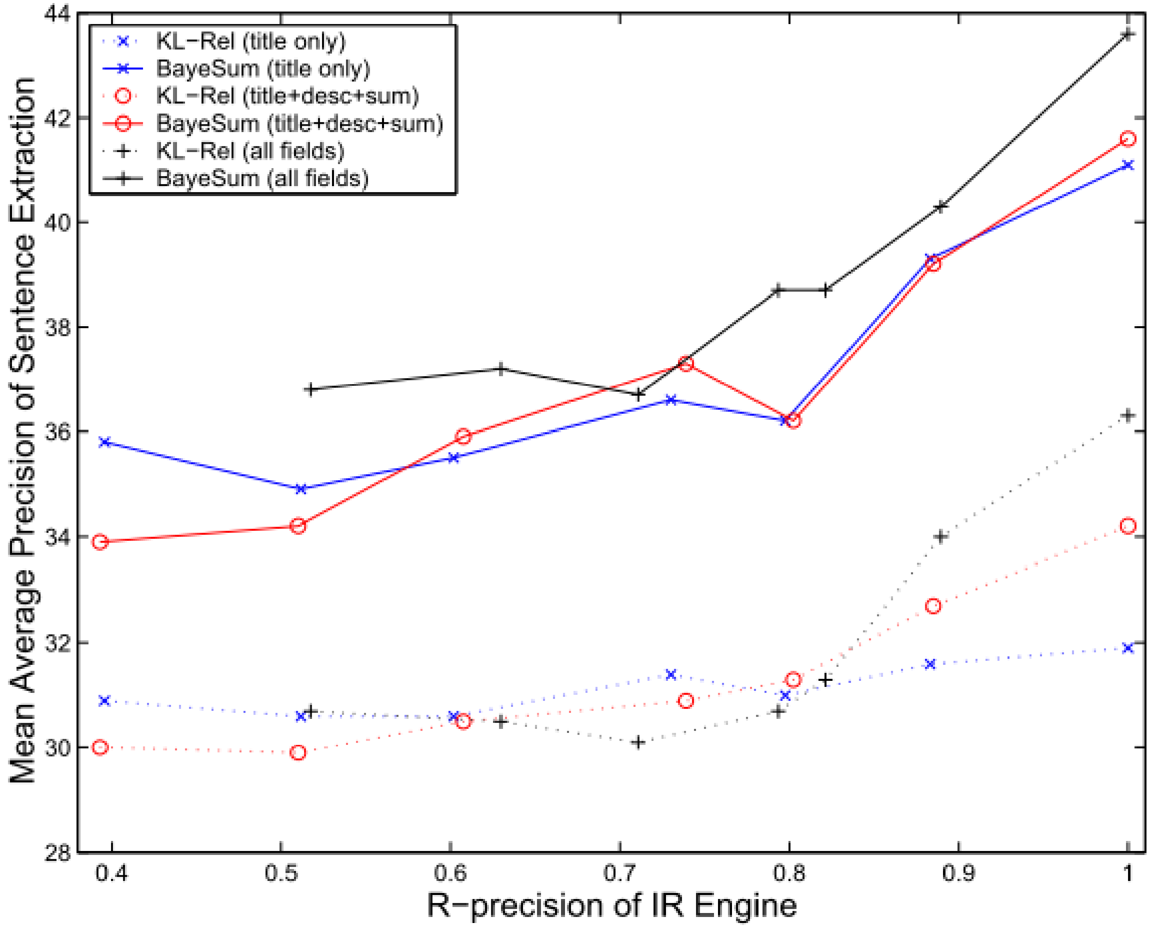
	MAP	MRR	P@2
RANDOM	19.9	37.3	16.6
POSITION	24.8	41.6	19.9
JACCARD	17.9	29.3	16.7
COSINE	29.6	50.3	23.7
KL	36.6	64.1	27.6
KL+REL	36.3	62.9	29.2
BAYESUM	44.1	70.8	33.6

Table 1: Empirical results for the baseline models as well as BAYESUM, when all query fields are used.

		MAP	MRR	P@2
POSITION		24.8	41.6	19.9
Title	KL	19.9	32.6	17.8
	KL-Rel	31.9	53.8	26.1
	BAYESUM	41.1	65.7	31.6
+Description	KL	31.5	58.3	24.1
	KL-Rel	32.6	55.0	26.2
	BAYESUM	40.9	66.9	31.0
+Summary	KL	31.6	56.9	23.8
	KL-Rel	34.2	48.5	27.0
	BAYESUM	42.0	67.8	31.8
+Concepts	KL	36.7	64.2	27.6
	KL-Rel	36.3	62.9	29.2
	BAYESUM	44.1	70.8	33.6
No Query	BAYESUM	39.4	64.7	30.4

Table 2: Empirical results for the position-based model, the KL-based models and BAYESUM, with different inputs.

Noisy Relevance Judgments



Discussion

- Any guideline to choose a language model for IR tasks?
  - $p_d(q)$, $p_q(d)$, $KL(p_q \,||\, p_d)$