Lecture 13: More uses of Language Models

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 13

What we'll learn in this lecture

- Comparing documents, corpora using LM approaches
- Generalization of $P(q|d)$ to same comparison model
- Relevance feedback under LM
- Relevance models
- Cross-lingual IR using LM techniques

Comparing documents

- In VSM, document similarity computed by distance in term space (cosine similarity)
- In LM, documents compared by similarity between probability distributions
- Several measures of dissimilarity between probability distributions available
- One is *Kullback-Leibler Divergence* (KL Divergence)

Kullback-Leibler divergence

- Let $p(x)$ and $q(x)$ be two prob dists over $\mathcal{X}$
- Then KL Divergence (relative entropy) $D(p\|q)$ defined as:

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \tag{1}$$

- Describes "mis-match" between distributions
  - E.g. if we develop optimal compression code based on $q()$, and use it to encode $p()$, $D(p\|q)$ is average extra bits per symbol
- Minimum value is 0, means identical distributions.
- Will give $+\infty$ if $q(x) = 0$, $p(x) > 0$ for any x.

KL Divergence applied

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \tag{2}$$

- Set $p = \theta_{d_1}$ as model of doc d_1, $q = \theta_{d_2}$ as model of doc 2
  - Will probably want some background smoothing
- KL Divergence applicable to any models
  - E.g. for doc d and corpus C, $D_{KL}(\theta_d\|\theta_c)$
- Note: not symmetric
  - Mutual information, $I(X;Y) = D_{KL}\{P(X,Y)\|P(X)P(Y)\}$, a symmetric alternative
  - KL divergence more appropriate where natural assymmetry (as doc to corpus)
    - MI blows up if $p(x) = 0$, $q(x) > 0$
    - KL divergence doesn't

KL Divergence as retrieval metric

Could use KL-Divergence as retrieval metric:

$$R(Q, D) = -KL(\theta_Q \| \theta_D) \tag{3}$$

In fact, this rank-equivalent to regular LM if

$$p(w|\theta_Q) = \frac{c(w, Q)}{|Q|} \tag{4}$$

i.e. if we use MLE for query model. (Neat, huh?)

Relevance feedback

- Query expanded with feedback from query results:
  - Automatically take top docs as relevant (PRF)
  - User specifies relevant documents (TRF)
- In VSM / Rocchio,
  - Query modelled as pseudo-document
  - Expanded by averaging with mean of feedback documents
  - Supports arbitrary weighting of feedback terms

Relevance feedback in LM4IR

- In LM4IR, query is example utterance generated by language model
- No straightforward way of weighting query terms
- So expansion only by literally adding terms to query
- Can't just add all terms from expansion documents to query
- How to select terms to add?
- Ratio models:
  - Select terms with high probability in feedback documents
  - ... low probability in collection
- Still unpleasantly heuristic

Relevance feedback with KL Divergence

- Want method that
  - Supported weights in expanded query
  - Provides mechanism for calculating weights
- This is provided by the KL Divergence framework
- Interpolate query model θ_Q with feedback model $\theta_{\mathcal{F}}$:

$$\theta_{Q'} = (1 - \alpha)\theta_Q + \alpha\theta_{\mathcal{F}} \tag{5}$$

- Then calculate:

$$R(D, Q; \mathcal{F}) = -D(\theta_{Q'} \| \theta_D) \tag{6}$$

- Efficiency gained by only retaining high-score terms in $M_{Q'}$
- Now we need to estimate $\theta_{\mathcal{F}}$ from feedback documents $\mathcal{F} = \{d_1, d_2, \ldots, d_n\}$

Estimating feedback model: unmixed

Follow the development in Zhai and Lafferty (CIKM, 2001)

- Want to find model $\theta_{\mathcal{F}}$ that generated (relevant parts of) $\mathcal{F}$
- Assume unigram. Then:

$$P(\mathcal{F}|\theta) = \prod_i \prod_w P(w|\theta)^{c(w,d_i)} \tag{7}$$

  where w iterates over words, i over feedback documents
- Find θ that maximizes (7) (for MLE)
- This is not (quite)[1] $\frac{c(w,\mathcal{F})}{\|\mathcal{F}\|}$, unless $|\mathcal{F}| = 1$
- However, not all of feeback documents relevant
- . . . so (7) not appropriate

[1] I think. Tell me if I'm wrong.

Estimating feedback model: mixture model

- Assume instead that words in $\mathcal{F}$ come from mixture of two models:
  - Relevance feedback model $\theta_{\mathcal{F}}$
  - Background (corpus) model C
- Therefore:

$$P(\mathcal{F}|\theta) = \prod_i \prod_w ((1-\lambda)P(w|\theta) + \lambda P(w|C))^{c(w,d_i)} \tag{8}$$

- Fix λ, solve for θ that maximizes (8)
  - Using EM algorithm (see Zhai and Lafferty for details)
- That θ is the value plugged in for $\theta_{\mathcal{F}}$ in:

$$\theta_{Q'} = (1-\alpha)\theta_Q + \alpha\theta_{\mathcal{F}} \tag{9}$$

- Finally, score using KL divergence

Mixture model: interpretation

$$P(\mathcal{F}|\theta) = \prod_i \prod_w ((1-\lambda)P(w|\theta) + \lambda P(w|C))^{c(w,d_i)} \qquad (8)$$

- Estimating θ on (8) dampens weight of coll-frequent terms
- If term w is frequent in feedback documents ($c(w, \mathcal{F})$ high):
  - if w is frequent in collection ($c(w, C)$ high)
    - then $c(w, \mathcal{F})$ largely explained by $c(w, C)$
    - and $P(w, \theta)$ doesn't have to be high
  - if w is rare in collection ($c(w, C)$ low)
    - then $c(w, \mathcal{F})$ not explained by $c(w, C)$
    - and $P(w, \theta)$ must be high
- Note λ must be fixed (i.e. externally tuned)
- Trying to optimize (8) for both λ and θ sets $\lambda = 0$, $P(w|\theta) \approx \frac{c(w,\mathcal{F})}{\|\mathcal{F}\|}$ (why?)
- Seems a Bayesian approach is possible (project for brave?)

Mixture model: practical effectiveness

$$P(\mathcal{F}|\theta) = \prod_i \prod_w ((1-\lambda)P(w|\theta) + \lambda P(w|C))^{c(w,d_i)} \tag{8}$$

- Zhai and Lafferty (CIKM 2001) find PRF with mixture model improves over plain LM
- Consider another feedback model (minimize divergence from feedback model), similar effectiveness
- LM+PRF beats TF*IDF+Rocchio
- λ not too sensitive, as long as not very high (gives very bad performance)

Relevance model

$$
\begin{aligned}
R(Q, D; \mathcal{F}) &= -KL(\theta_{Q'} \| \theta_D) \\
&= -KL(\{(1-\alpha)\theta_Q + \alpha\theta_{\mathcal{F}}\} \| \theta_D) \\
&\approx P(R = r | Q, D) \\
P(w|\theta_{Q'}) &= (1-\alpha)\frac{c(w,q)}{|q|} + \alpha P(w|\theta_{\mathcal{F}}) \\
&\approx P(w|\theta_R)
\end{aligned}
$$

- Query model expanded with relevance feedback, θ'_Q
- …an approximation to *relevance model*

Alternative relevance model

Lavrenko and Croft (2001), give similar (simpler) relevance model:

$$
\begin{aligned}
P(w|q;\mathcal{F}) &\propto \sum_{F\in\mathcal{F}} P(w|F)\prod_{i}^{|q|} P(q_i|F) \\
P(\{w,q_i\}|F) &= \lambda\left(\frac{c(w,F)}{|F|}\right) + (1-\lambda)P(w)
\end{aligned}
$$

(They also present a more robust, unequal sampling method)

Cross-lingual IR

- Query in language L_Q (say, English)
- Search over documents in language L_S (say, Chinese)
- Could be done by translating query, or documents
- But can be done directly
- ... using relevance LM to bridge gap

Relevance model in CLIR

- Assume parallel corpora $\mathcal{M}_E$, $\mathcal{M}_C$, with $\{(M_E, M_C)\}$ pairs of parallel documents
- Assume target corpus is $\mathcal{T}_C \neq \mathcal{M}_C$.
- Issue query q against $\mathcal{M}_E$.
- Retrieve top n docs $\mathcal{F}_E$, fetch parallel docs $\mathcal{F}_C$
- Estimate:

$$P(w_C|\theta_{q_E;\mathcal{F}}) = \sum_{\{F_E,F_C\}\in\mathcal{F}} P(w_C|F_C) \prod_i^{|q|} P(q_i|F_E) \qquad (10)$$

- Apply (10) to each word in each doc in $\mathcal{T}_C$ to calc rel score
- Achieves 90–95% of effectiveness of monolingual IR

Looking back and forward

Back

- Language models (from queries, documents, document sets, corpora) comparing using KL divergence (or Mutual Information)
- KL divergence of query from document a generalization of language model approach
- Relevance feedback in LM can be done by interpolated query and feedback models
  - Feedback model itself mixed with background model
- Relevance feedback methods used to create relevance model
- Relevance model can be applied to perform cross-lingual IR

Looking back and forward

Forward

- Language models with relevance feedback similar to Naive Bayes classification
- Relevance models a supervised version of topic models

Further reading

- Lafferty and Zhai, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval", SIGIR 2001.
- Zhai and Lafferty, "Model-based Feedback in the Language Modeling Approach to Information Retrieval", CIKM 2001.
- Lafferty and Zhai, "Probabilistic Relevance Models Based on Document and Query Generation", LMIR 2003.
- Zhai, "Statistical Language Models for Information Retrieval: A Critical Review", FnTIR, 2008.
- Lavrenko and Croft, "Relevance-Based Language Models", SIGIR 2001.
- Lavrenko, Choquette, and Croft, "Cross-Lingual Relevance Models", SIGIR 2002.