

Text Technologies

Relevance-based Models

Victor Lavrenko

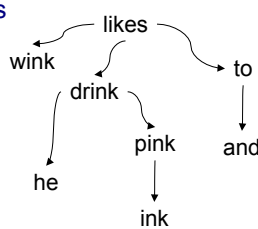
Copyright © 2011 Victor Lavrenko

Modeling word dependence

- Classical model assumes all words independent
 - blatantly false, made by almost all retrieval models
 - the most widely criticized assumption behind IR models
 - should be able to do better, right?
- Tree dependence model (van Rijsbergen, 1977)

- structure dependencies as maximum spanning tree
 - edge weights: mutual information between words
- each word depends on its parent (and R)
 - total # parameters: twice that of BIR

$$\begin{aligned}
 &P(\text{"he likes to wink and drink pink ink"}) \\
 &= P(\text{likes}) * P(\text{to|likes}) * P(\text{wink|likes}) \\
 &* P(\text{and|to}) * P(\text{drink|likes}) * P(\text{he|drink}) \\
 &* P(\text{pink|drink}) * P(\text{ink|pink})
 \end{aligned}$$



Copyright © 2011 Victor Lavrenko

Classical Probabilistic model: BIR

- Probability Ranking Principle: best possible ranking

$$P(R=1|D) = \prod_{w \in D} \frac{p_w(1-q_w)}{q_w(1-p_w)} = \prod_{w \in D \cap Q} \frac{N - N_w + 0.5}{N_w + 0.5}$$

- Assumptions:
 - A0: relevance for document in isolation
 - A1: words absent or present (can't model frequency)
 - A2: all words mutually independent (given relevance)
 - A3: empty document equally likely for R=0,1
 - A4: non-query words cancel out
 - A5: query words: relevant class doesn't matter
 - A6: non-relevant class ~ collection as a whole

make it fast:
D/Q overlap

estimate p_w, q_w
w/out relevance

- How can we improve the model?

- relax particularly bad assumptions

Copyright © 2011 Victor Lavrenko

Do dependence models work?

- Many similar attempts since the original
 - dozens published results, probably hundreds of attempts
 - many different ways to model dependence
 - never consistent improvement: always "promising results"
- Why? It works in other fields.
 - independence (unigram) would be a silly choice for ASR, MT
 - need to handle surface form of the string (is output grammatical?)
 - in IR we are already dealing with well-formed strings
 - pointless to waste probability mass on grammaticality
 - BIR doesn't really assume independence
 - necessary condition significantly weaker than independence

Copyright © 2011 Victor Lavrenko

BIR doesn't assume independence

$$\frac{P_{R=1}(\vec{d})}{P_{R=0}(\vec{d})} = \prod_w \underbrace{\frac{P_1(d_w)}{P_0(d_w)}}_{\text{independence}} \times \underbrace{\frac{k_1(w)}{k_0(w)}}_{\text{will not affect ranking if}} = \prod_w \underbrace{\frac{P_1(d_w | d_{\pi(w)})}{P_0(d_w | d_{\pi(w)})}}_{\text{1st order dependence}}$$

$$k_r(w) = \frac{P_r(d_w, d_{\pi(w)})}{P_r(d_w)P_r(d_{\pi(w)})}$$

$$\sum_w \log \frac{P_1(d_w, d_{\pi(w)})}{P_1(d_w)P_1(d_{\pi(w)})} \sim \sum_w \log \frac{P_0(d_w, d_{\pi(w)})}{P_0(d_w)P_0(d_{\pi(w)})}$$

aggregate dependence between word and parent in the relevant class aggregate dependence in the non-relevant class

Sufficient condition: **proportional interdependence**

the **total** amount of interdependence among **all** words in a document is approximately the same under R=1 and R=0

Copyright © 2011 Victor Lavrenko

Meaning of independence

- Independence:
 - seeing "subprime" doesn't affect chances of seeing "loan"
- Linked Dependence:
 - seeing "subprime" increases chance of seeing "loan"
 - by the same amount under R=1 and R=0
 - reasonable... unless topic is financial crisis
- Proportional Interdependence:
 - "subprime" increases chance of "loan"
 - can be more co-dependent in relevant class
 - as long as offset by other word sets under R=0
 - "world cup" more co-dependent in non-relevant class

Copyright © 2011 Victor Lavrenko

Classical Probabilistic model: BIR

- Probability Ranking Principle: best possible ranking

Assumptions: $P(R=1|D) = \prod_{w \in D} \frac{p_w(1-q_w)}{q_w(1-p_w)} = \prod_{w \in D \cap Q} \frac{N - N_w + 0.5}{N_w + 0.5}$

- A0: relevance for document in isolation
 - A1: words absent or present (can't model frequency)
 - A2: all words mutually independent (given relevance)
 - A3: empty document equally likely for R=0,1
 - A4: non-query words cancel out
 - A5: query words: relevant class doesn't matter
 - A6: non-relevant class ~ collection as a whole
- make it fast: D/Q overlap
- estimate p_w, q_w w/out relevance

- How can we improve the model?

- relax particularly bad assumptions

Copyright © 2011 Victor Lavrenko

Modeling word frequencies

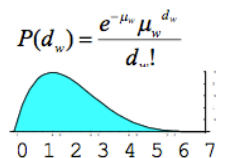
- Want to model TF (empirically useful) $P(R=1|D) = \prod_{w \in D} \frac{P(D_w | R=1)}{P(D_w | R=0)}$
 - A1': assume $D_w = d_w \dots$ # times word w occurs in document D
 - estimate $P(d_w|R)$: e.g. "obama" occurs 5 times in a rel. doc
 - naive: estimate probability for every outcome:
 - $P(D_w=0), P(D_w=1), P(D_w=2), P(D_w=3) \dots$ for R=1 and R=0
 - many outcomes \rightarrow many parameters (BIR had only one p_w)
 - want "smoothness": $d_w=5$ similar to $d_w=6$, but not $d_w=1$

- parametric model: assume $d_w \sim$ Poisson

- single parameter $\mu_w \dots$ expected frequency

- problem: Poisson a poor fit to observations

- does not capture bursty nature of words



Copyright © 2011 Victor Lavrenko

Two-Poisson model [Harter]

- Idea: words generated by a mixture of two Poissons
 - “elite” words for a document: occur unusually frequently
 - “non-elite” words – occur as expected by chance
 - document is a mixture: $P(d_w) = P(E=1) \frac{e^{-\mu_{1,w}} \mu_{1,w}^{d_w}}{d_w!} + P(E=0) \frac{e^{-\mu_{0,w}} \mu_{0,w}^{d_w}}{d_w!}$
 - estimate $m_{0,w}, m_{1,w}, P(E=1)$ by fitting to data (max. likelihood)
- Problem: need probabilities conditioned on relevance
 - “eliteness” not the same as relevance
 - Robertson and Sparck Jones: condition eliteness on $R=0, R=1$
 - final form has too many parameters, and no data to fit them...

- BM25: an “approximation” to conditioned 2-Poisson

$$\frac{p_w(d_w)q_w(0)}{q_w(d_w)p_w(0)} \approx \exp\left(\underbrace{\frac{d_w(1+k)}{d_w + k((1-b) + b \cdot dl / avg.dl)}_{\text{squashed TF, k,b: parameters}} \cdot \log \frac{N - N_w + 0.5}{N_w + 0.5}}_{\text{IDF}}\right)$$

Copyright © 2011 Victor Lavrenko

Example: BM25

- documents: $D_1 = \text{“a b c b d”}$, $D_2 = \text{“b e f b”}$, $D_3 = \text{“b g c d”}$, $D_4 = \text{“b d e”}$, $D_5 = \text{“a b e g”}$, $D_6 = \text{“b g h h”}$
- query: $Q = \text{“a c h”}$, assume $k = 1, b = 0.5$

word:	a	b	c	d	e	f	g	h	
$N(w)$:	2	6	2	3	3	1	3	1	$N = 6$
$N - N_w / N_w$:	$4.5 / 2.5$	$0.5 / 6.5$	$4.5 / 2.5$	$3.5 / 3.5$	$3.5 / 3.5$	$5.5 / 1.5$	$3.5 / 3.5$	$5.5 / 1.5$	

$$\log \frac{P(D_1 | R = 1)}{P(D_1 | R = 0)} \approx 2 \times \left(\frac{1 \cdot (1 + 1)}{1 + 1 \cdot (0.5 + \frac{0.5 \cdot 5}{4})} \cdot \log \frac{4.5}{2.5} \right)$$

$$\log \frac{P(D_6 | R = 1)}{P(D_6 | R = 0)} \approx \left(\frac{2 \cdot (1 + 1)}{2 + 1 \cdot (0.5 + \frac{0.5 \cdot 4}{4})} \cdot \log \frac{5.5}{1.5} \right)$$

BM25: an intuitive view

Copyright © 2011 Victor Lavrenko

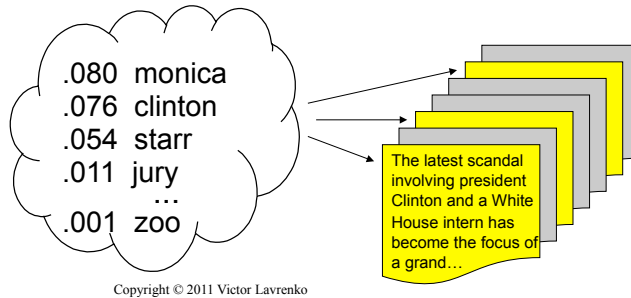
Classical Probabilistic model: BIR

- Probability Ranking Principle: best possible ranking
 - Assumptions:

$$P(R=1 | D) = \prod_{w \in D} \frac{p_w(1 - q_w)}{q_w(1 - p_w)} = \prod_{w \in D \cap Q} \frac{N - N_w + 0.5}{N_w + 0.5}$$
 - A0: relevance for document in isolation
 - A1: words absent or present (can't model frequency)
 - A2: all words mutually independent (given relevance)
 - A3: empty document equally likely for $R=0, 1$
 - A4: non-query words cancel out
 - A5: query words: relevant class doesn't matter
 - A6: non-relevant class ~ collection as a whole
 - How can we improve the model?
 - relax particularly bad assumptions
- make it fast: D/Q overlap
estimate p_w, q_w w/out relevance

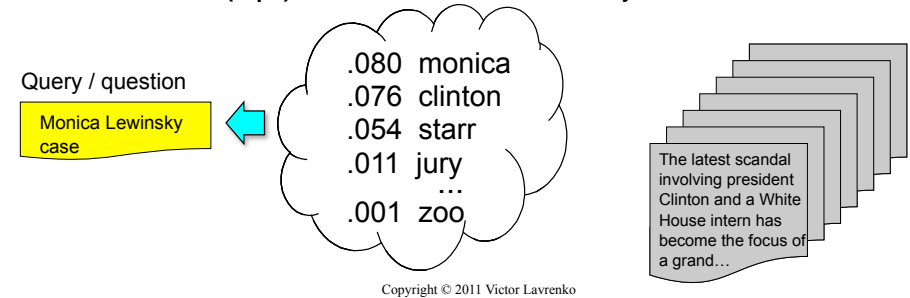
Relevance Model

- probabilistic model of the relevant class
 - $P(\mathbf{w}|\mathbf{R})$... chance of seeing word \mathbf{w} in a random relevant doc.
 - relevant documents = random samples from $P(\mathbf{w}|\mathbf{R})$
 - different event space: $P(\mathbf{w}|\mathbf{R})$ is a multinomial urn
 - document = sequence of observations (words) drawn from that urn
 - allows repetitions \rightarrow way to model frequencies



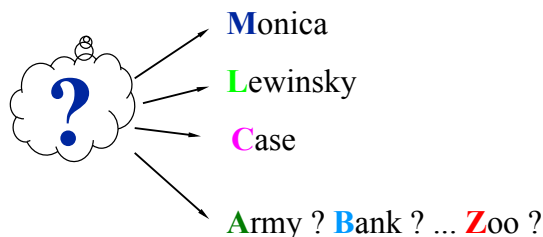
Estimating a Relevance Model

- Easy if we had relevant examples:
 - count how many times each word occurs in the relevant set
- But we don't have examples of relevant documents
- Assume query is a random sample from $P(\mathbf{w}|\mathbf{R})$
 - sample too small to estimate $P(\mathbf{w}|\mathbf{R})$ by counting
 - assume $P(\mathbf{w}|\mathbf{R})$ has a certain form, use Bayesian estimation



Estimation: the sampling game

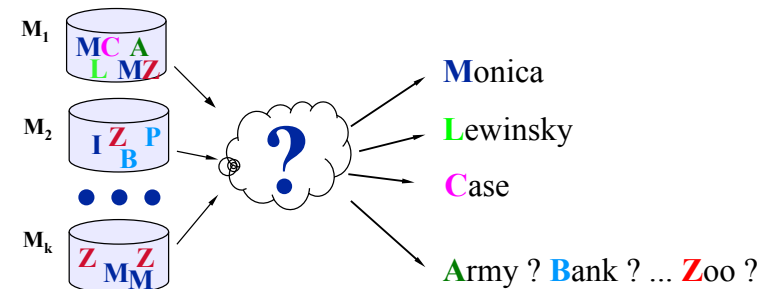
- Assume our query is a sample from the $P(\mathbf{w}|\mathbf{R})$
- Play a sampling game:
 - unknown distribution, sample 3 times, get: *Monica, Lewinsky, Case*
 - what's the chance we get a word *Zoo* if we sample one more time?



$$P(\text{Zoo}|\mathbf{R}) \sim P(\text{Zoo} \text{ Monica Lewinsky Case})$$

Estimation: non-parametric form

assume a universe $\{M_1 \dots M_k\}$ of candidate models
 estimate the probability $P(\text{Zoo} | \text{Monica, Lewinsky, Case})$ over that universe



$$\begin{aligned} P(\text{Zoo}|\mathbf{R}) &\sim P(\text{Zoo}|\text{Monica Lewinsky Case}) \\ &= P(\text{Z}|M_1)P(M_1|MLC) + P(\text{Z}|M_2)P(M_2|MLC) + \dots \\ &= \sum_M P(\text{Zoo}|M)P(M|\text{Monica Lewinsky Case}) \end{aligned}$$

Estimation: candidate models

- $M_1 \dots M_K$ = documents in our collection
 - each doc = example of how words co-occur with other words
- Examples of Relevance Models
 - showing top 10 from probability distribution over entire vocabulary
 - each distribution estimated from 2-3 words only
 - using a collection of ~64k news stories from 1998 (TDT2)

"Monica Lewinsky Case"		"Israeli Palestinian Raids"		"Rats in Space"		"John Glenn"		"Unabomber"	
$P(w Q)$	w	$P(w Q)$	w	$P(w Q)$	w	$P(w Q)$	w	$P(w Q)$	w
0.041	lewinsky	0.077	palestinian	0.062	rat	0.032	glenn	0.046	kaczynski
0.038	monica	0.055	israel	0.030	space	0.030	space	0.046	unabomber
0.027	jury	0.034	jerusalem	0.020	shuttle	0.026	john	0.019	ted
0.026	grand	0.033	protest	0.018	columbia	0.016	senate	0.017	judge
0.019	confidant	0.027	raid	0.014	brain	0.015	shuttle	0.016	trial
0.016	talk	0.012	find	0.012	mission	0.011	seventy	0.013	say
0.015	case	0.011	clash	0.012	two	0.011	america	0.012	theodore
0.014	president	0.010	bank	0.011	seven	0.011	old	0.012	today
0.013	clinton	0.010	west	0.010	system	0.010	october	0.011	decide
0.010	starr	0.010	troop	0.010	nervous	0.010	say	0.011	guilty

Copyright © 2011 Victor Lavrenko

Estimation summary

- Relevance model: $P(w | R) = P(w | q_1 \dots q_k) = \frac{P(w, q_1 \dots q_k)}{P(q_1 \dots q_k)}$
- Joint probability for set of words: $P(w, q_1 \dots q_k) = \sum_D P(D) P(w | D) \prod_{i=1}^k P(q_i | D)$
 - from DeFinetti's theorem
- Uniform prior over models: $P(D) = \frac{1}{N}$
- Document model $P(w|D)$
 - Bayesian estimate based on frequency with Dirichlet prior $P(w | D) = \frac{\#(w, D) + \mu P(w | C)}{|D| + \mu}$
- Dirichlet prior $P(w|C)$
 - maximum-likelihood estimate: $P(w | C) = \frac{\#(w, C)}{|C|}$
 - based on collection counts

Copyright © 2011 Victor Lavrenko

Ranking with Relevance Models

- $P(w|R)$... model of the relevant class
 - rank using odds ratio $\frac{P(d|R)}{P(d|N)} = \prod_{w \in d} \frac{P(w|R)}{P(w|N)} \leq \left(\frac{P(w^*|R)}{P(w^*|N)} \right)^{|d|}$
 - $P(w|N)$ as before
 - perfect document = single "killer feature" w^*
 - will favor documents with few highly-discriminative words
- Better approach: Kullback-Leibler divergence
 - negative KL between document & relevance model $-\sum_w P(w|R) \log \frac{P(w|R)}{P(w|D)}$
 - perfect document = same distribution as relevance model
 - will favor documents with many relevant words
- One of the most successful retrieval algorithms

Copyright © 2011 Victor Lavrenko

Extensions of Relevance Models

- general model for estimating a joint distribution
- what if we have paired English-Chinese docs?
 - compute: $P(\text{Environment Protection} | \text{Environment Protection})$
 - very successful approach to cross-language retrieval
- what if we have paired images w. annotations?

$P(\text{Cumberland} | \text{Fort Cumberland})$

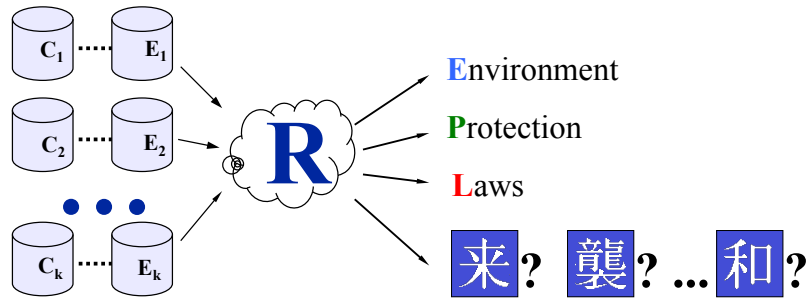
$P(\text{Tiger} | \text{Image of Tiger})$

- state-of-the-art approach for auto-annotating images

Copyright © 2011 Victor Lavrenko

Cross-lingual Relevance Models

probability of seeing 来 in the relevant class R



$$\begin{aligned}
 P(\text{来} | R) &\sim P(\text{来} | \text{Environment Protection Laws}) \\
 &= P(\text{来} | C_1)P(E_1 | \text{E P L}) + P(\text{来} | C_2)P(E_2 | \text{E P L}) + \dots \\
 &= \sum_k P(\text{来} | C_k)P(E_k | \text{Environment Protection Laws})
 \end{aligned}$$

Copyright © 2011 Victor Lavrenko

Summary

- Probability Ranking Principle
 - ranking by $P(R=1|D)$ is optimal
- Classical probabilistic model
 - words: binary events (relaxed in the 2-Poisson model)
 - words assumed independent (not accurate)
 - numerous attempts to model dependence, most without success
- Formal, interpretable model
 - explicit, elegant model of relevance (if observable)
 - very problematic if relevance not observable
 - authors resort to heuristics, develop BM25
- Relevance models:
 - estimate $P(w|R)$ from a very short sample (2-3 words)
 - state-of-the-art for search, cross-language, images

Copyright © 2011 Victor Lavrenko