Generating Query Substitutions

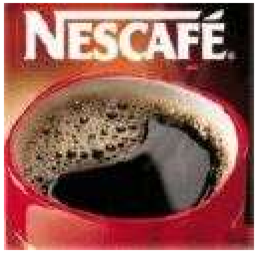
Rosie Jones

Benjamin Rey, Omid Madani and Wiley Greiner

Yahoo! Research

Substitutes

→

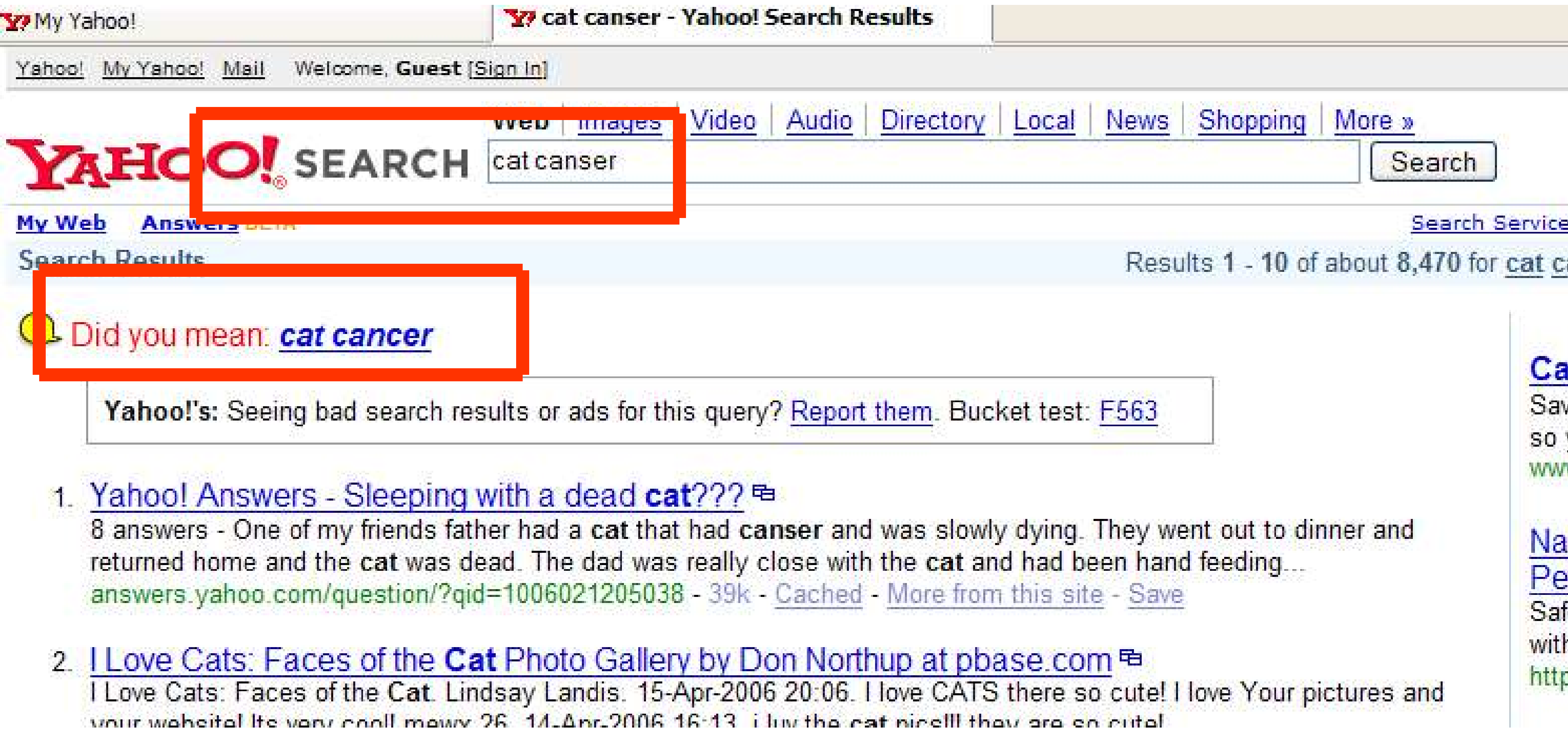


→

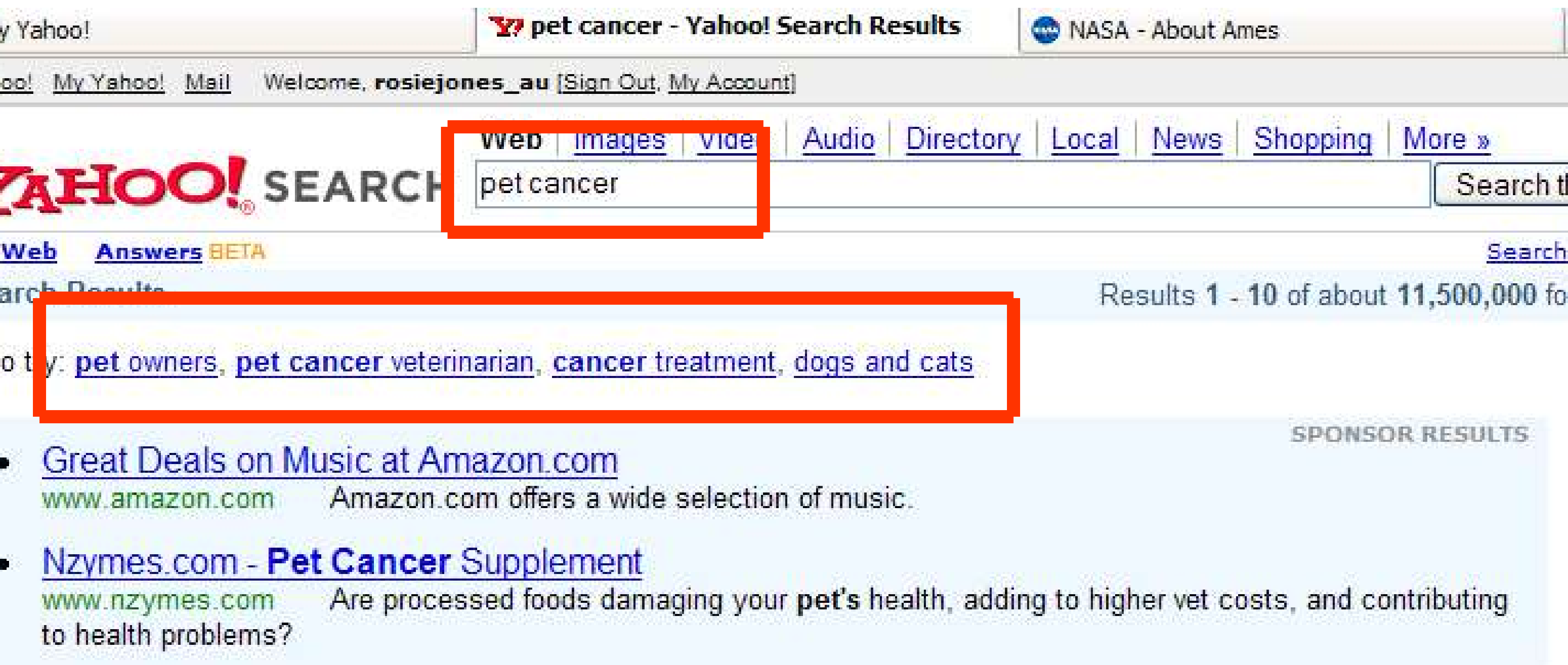
→

→

Query Substitutions



Query Substitutions



Functions of Rewriting

- Enhance meaning
  - Spell correction
  - Corpus-appropriate terminology
    - Cat cancer → feline cancer
- Change meaning
  - Narrow
    - [lexical entailment: fruit → apple]
  - Broaden
    - [alternatives, common interests]
    - Conference proceedings → textbooks

Trying to Find Nathan Welsh, who lives and works in Edinburgh

- nathan welsh edinburg scotland
- nathan welsh edinburgh scotland
- financial consultants edinburg scotland
- financial consultants edinburgh scotland
- financial consultants
- nathan welsh 16-18 pennwell place edinburgh
- nathan welsh 16-18 pennywell place edinburgh
- international phone directory
- white pages
- edinburgh scotland phone directory
- edinburgh scotland uk
- nathan welsh investment consultant edinburg
- nathan welsh investment consultant edinburgh
- investment consultants edinburgh scotland
- nathan welsh
- kansas virginia
- herndon virginia

Spell correction
Name →profession
Spell correction
Delete terms, generalize
Try second approach, using his address
Spell correction
Try looking up addresses
rephrase
specialization
Generalize to location
Switch to new topic

Half of Query Pairs are Related

Type	Example	%
non-rewrite	mic amps -> create taxi	53.2%
insertions	game codes -> video game codes	9.1%
substitutions	john wayne bust -> john wayne statue	8.7%
deletions	skateboarding pics → skateboarding	5.0%
spell correction	real eastate -> real estate	7.0%
mixture	huston's restaurant -> houston's	6.2%
specialization	jobs -> marine employment	4.6%
generalization	gm reabtes -> show me all the current auto rebates	3.2%
other	thansgiving -> dia de acconde gracias	2.4%

Substitutions are repeated

- car insurance → auto insurance
  - 5086 times in a sample
- car insurance → car insurance quotes
  - 4826 times
- car insurance → geico [brand of car insurance]
  - 2613 times
- car insurance → progressive auto insurance
  - 1677 times
- car insurance → carinsurance
  - 428 times

Different Users, Different Days

Statistical Test to Find Significant Rewrites

Test whether

$$p(q2 \mid q1) >> p(q2)$$

P(britney spears|brittney spears) >> P(britney spears)

8% >> 0.01%

Log likelihood ratio test (GLRT) gives distributed score

About 90% of query pairs are related after filtering with LLR > 100

Many Types of Substitutable Rewrites

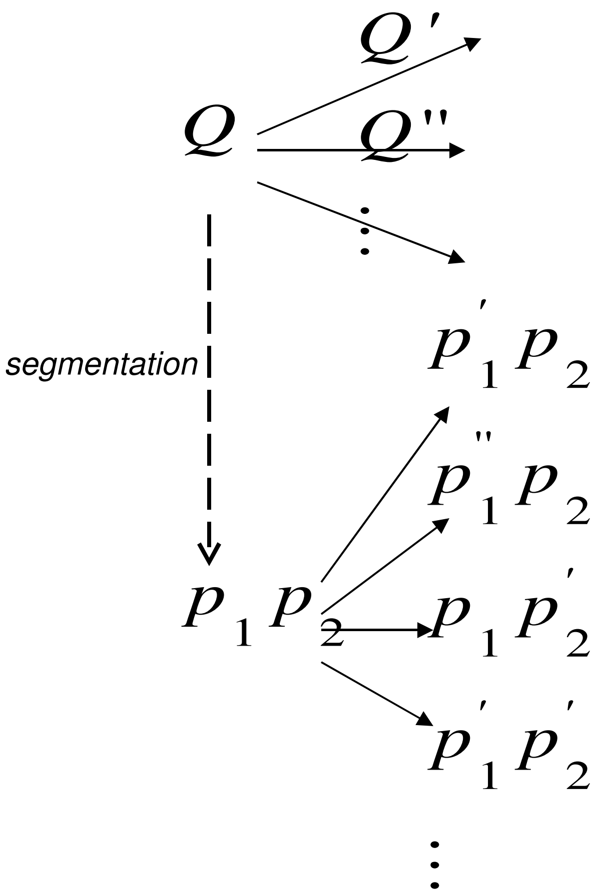
dog -> dogs	9185	pluralization
dog -> cat	5942	both instances of 'pet‘
dog -> dog breeds	5567	generalization
dog -> dog pictures	5292	more specific
dog -> 80	2420	random junk in query processing
dog -> pets	1719	generalization -- hypernym
dog -> puppy	1553	specification -- hyponym
dog -> dog picture	1416	more specific
dog -> animals	1363	generalization -- hypernym
dog -> pet	920	generalization -- hypernym

Defining Categories of Relatedness for Sponsored Search

1- Precise Match	A near-certain match. *E.g.: automotive insurance - automobile insurance;*
2- Approximate Match	A probable, but inexact match with user intent. E.g.*: apple music player - ipod shuffle*
3- Marginal Match	A distant, but plausible match to a related topic. E.g.: *glasses - contact lenses*
4- Mismatch	A clear mismatch.

We will call {1,2} Precise and {1,2,3} Broad

Increase Tail Coverage with Query Segmentation



- Query segmented using high mutual information terms
- Most frequent queries: replace whole query
- Infrequent queries: replace constituent phrases

Generating Query Substitutions

- Q1 -> {q2,q3,q4,q5,q6}
- "catholic baby names" -> {christian baby names, christian baby boy names, catholic names, …}
- All are statistically relevant (log likelihood ratio on successive queries)

Find a model to

- rank substitutions, to be able to pick the best ones

$$score\left(Q-\text{¿}u''_1 u_2\right) < score\left(Q-\text{¿}Q''\right) < \ldots$$

- associate a probability of correctness

$$P\left(Q-\text{¿}Q' \text{ is correct} \,|\, score\left(Q-\text{¿}Q'\right)\right)$$

Train/Test Data

- Sample 1000 queries (q1)
- Select a single substitution for each (q2)
- Manually label the <q1,q2> pairs
- Learn to score <q1,q2> pairs
- Order by score
- Assess Precision/Recall
  - Precise task {1,2} vs {3,4}
  - Broad task {1,2,3} vs {4}

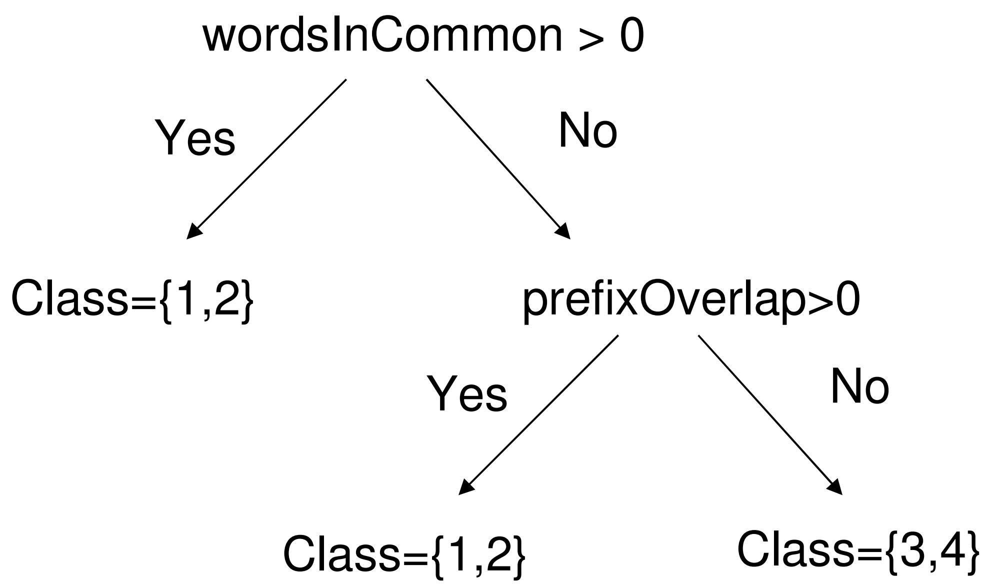
Predicting High Quality Query Suggestions

- Used labels to fit model
- Tried 37 features for model:
  - Lexical features including
    - Levenshtein character edit distance
    - Prefix overlap
    - Porter-stem
    - Jaccard score on words
  - Statistical features including
    - Probability of rewrite
    - Frequency of rewrite
  - Other
    - Number of substitutions (numSubst)
      - Whole query = 0
      - Replace one phrase = 1
      - Replace two phrases = 2
    - Query length, existence of sponsored results…

Baseline: Most suggestions broadly related

	Precise {1,2}	Broad {1,2,3}
Random Suggestion	55%	-
Maximize LLR Minimize numSubst	66%	87.5%

Simple Decision Tree



Interpretation of the decision tree:

- substitution must have at least 1 word in common with initial query
- the beginning of the query should stay unchanged

Linear Regression Model

Regression: continuous output in [1,4]

$$LMScore = intercept + \sum_{f=features} w_f . f$$

Classification:

If(*LMScore* < *T*) then *Good*, else *Bad*

For each T, we have a precision and a recall

Evaluation*:*

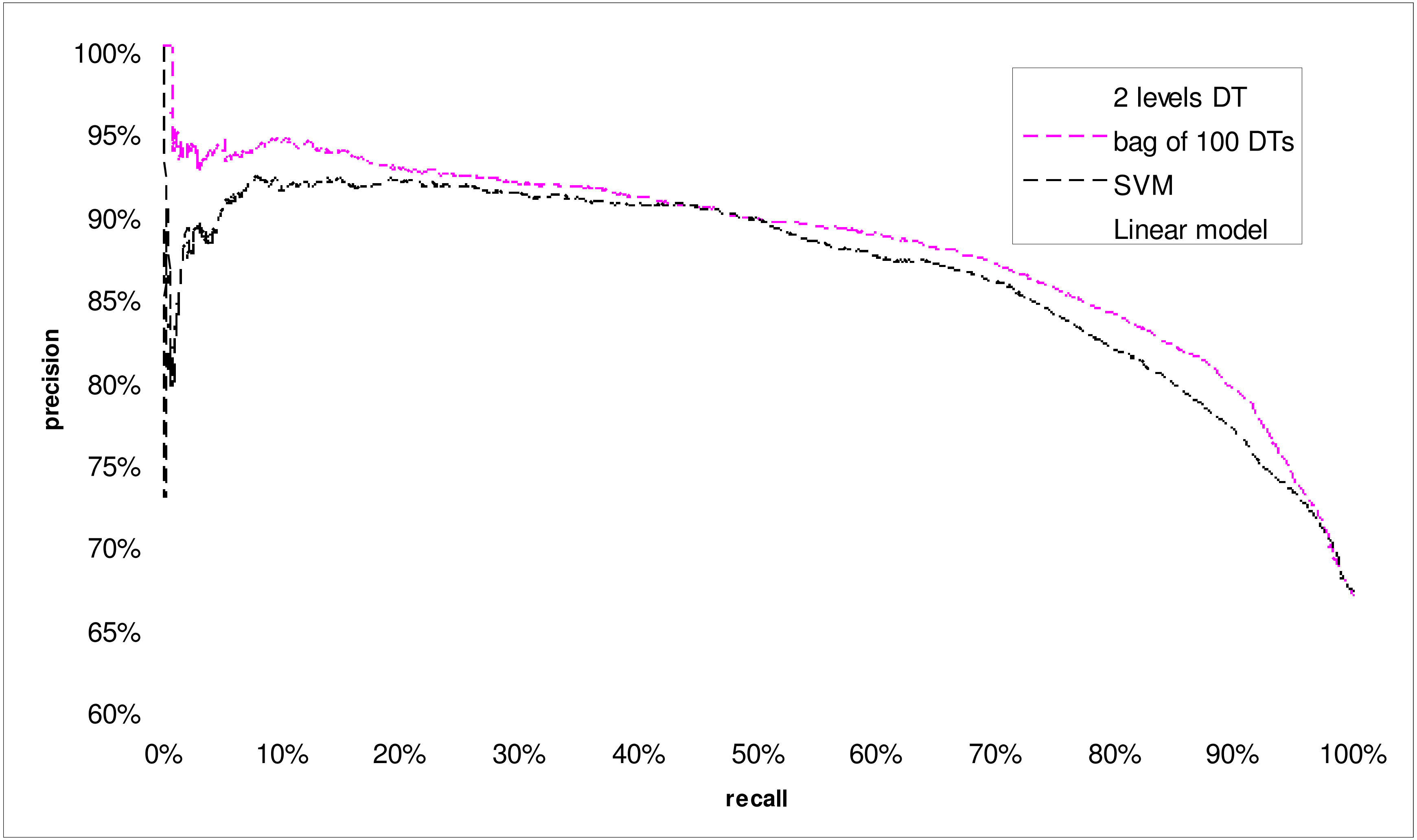
Average precision / recall on 100 times 10-fold cross validation

Learned Function

$$f(q_1, q_2) = 0.74 + 1.88 \times editDist(q_1, q_2)$$
$$+ 0.71 \times wordDist(q_1, q_2)$$
$$+ 0.36 \times numSubst(q_1, q_2)$$

- Outputs continuous score [1..4]
- Like decision tree
  - Prefer few edits
  - Prefer few word changes
  - Prefer whole-query or few phrase changes
- Normalize output to a probability of correctness using sigmoid fit

SVM, Bags of Trees, Linear Model Give Similar Trade-offs



Results Reranking Query-Suggestion Pairs

	Breakeven	Max-F1	Av. Prec
Baseline	0.66	0.66	0.66
2 level DT	0.71	0.83	0.71
SVM	0.81	0.83	0.86
Linear Model	**0.80**	**0.84**	**0.87**
Bag of 100 DTs	**0.83**	**0.84**	**0.88**

Generate Best Candidate on New Sample: Precision at 100% Recall

	Precise {1,2}	Broad {1,2,3}
Random Suggestion	55%	-
Maximize LLR Minimize numSubst	66%	87.5%
Linear Regression	74%	87.5%

Example Query Substitutions

Initial Query	Substitution	Hand-label	Alg. Prob
anne klien watches	anne klein watches	1	92%
sea world san diego	sea world san diego tickets	2	90%
restaurants in washington dc	restaurants in washington	2	89%
nash county	wilson county	3	66%
frank sinatra birth certificate	elvis presley birth	4	17%

Applications

- Sponsored Search
- Query Expansion
- Assisted Search
- Lexical entailment

Other Sources of Phrase Similarity

Data Source	Cooccurence	Distributional Similarity
Queries	Pet | dogs	dog | food pet | food
Query Sequence	Pets $\rightarrow$ dogs	dogs $\rightarrow$ petshop pets $\rightarrow$ petshop
Sentences	"Pets such as dogs"	"pets like their owners" "dogs like their owners"
Documents	"… pets … dogs …"	"dogs .. owners .. food" "pets .. owners .. food"

Sample Query Substitution Types

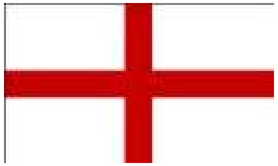
- meaning of dreams → interpretation of dreams (synonym)
- furniture etegere → furniture etagere (spelling correction)
- venetian hotel → venetian hotel las vegas (expansion)
- delta employment credit union → decu (acronym)
- lyrics finder → mp3 finder (related term)
- national car rental → alamo car rental (related brand)
- amanda peet → saving silverman (actress in)

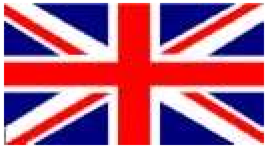
Future Work

- Add other sources of similarity
- Try additional features
- IR experiments

Substitutes for Edinburgh Castle

- edinburgh castle scotland (0.86)
- edinburgh (0.79)
- stirling castle (0.76)
- dudley castle (0.56)





Substitutes for Mars Bar

- mars candy bar (0.83)
- mars candy (0.77)

- Substitutes for “deep fried mars bar”
  – ?