Pseudo-relevance Feedback & Passage Retrieval

Ling573
NLP Systems & Applications
April 16, 2013

Major Issue in Retrieval

- All approaches operate on term matching
  - If a synonym, rather than original term, is used, approach can fail

Major Issue

- All approaches operate on term matching
  - If a synonym, rather than original term, is used, approach can fail
- Develop more robust techniques
  - Match "concept" rather than term

Major Issue

- All approaches operate on term matching
  - If a synonym, rather than original term, is used, approach can fail

- Develop more robust techniques
  - Match "concept" rather than term
    - Mapping techniques
      - Associate terms to concepts
        - Aspect models, stemming

Major Issue

- All approaches operate on term matching
  - If a synonym, rather than original term, is used, approach can fail
- Develop more robust techniques
  - Match "concept" rather than term
    - Mapping techniques
      - Associate terms to concepts
        - Aspect models, stemming
    - Expansion approaches
      - Add in related terms to enhance matching

Compression Techniques

- Reduce surface term variation to concepts

Compression Techniques

- Reduce surface term variation to concepts
- Stemming

Compression Techniques

- Reduce surface term variation to concepts
- Stemming
- Aspect models
  - Matrix representations typically very sparse

Compression Techniques

- Reduce surface term variation to concepts
- Stemming
- Aspect models
  - Matrix representations typically very sparse
  - Reduce dimensionality to small # key aspects
    - Mapping contextually similar terms together
    - Latent semantic analysis

Expansion Techniques

- Can apply to query or document

Expansion Techniques

- Can apply to query or document
- Thesaurus expansion
  - Use linguistic resource – thesaurus, WordNet – to add synonyms/related terms

Expansion Techniques

- Can apply to query or document
- Thesaurus expansion
  - Use linguistic resource – thesaurus, WordNet – to add synonyms/related terms
- Feedback expansion
  - Add terms that "should have appeared"

Expansion Techniques

- Can apply to query or document
- Thesaurus expansion
  - Use linguistic resource – thesaurus, WordNet – to add synonyms/related terms
- Feedback expansion
  - Add terms that "should have appeared"
    - User interaction
      - Direct or relevance feedback
    - Automatic pseudo relevance feedback

Query Refinement

- Typical queries very short, ambiguous
  - Cat: animal/Unix command

Query Refinement

- Typical queries very short, ambiguous
  - Cat: animal/Unix command
  - Add more terms to disambiguate, improve
- Relevance feedback

Query Refinement

- Typical queries very short, ambiguous
  - Cat: animal/Unix command
  - Add more terms to disambiguate, improve
- Relevance feedback
  - Retrieve with original queries
  - Present results
    - Ask user to tag relevant/non-relevant

Query Refinement

- Typical queries very short, ambiguous
  - Cat: animal/Unix command
  - Add more terms to disambiguate, improve
- Relevance feedback
  - Retrieve with original queries
  - Present results
    - Ask user to tag relevant/non-relevant
  - "push" toward relevant vectors, away from non-relevant
    - Vector intuition:
      - Add vectors from relevant documents
      - Subtract vector from non-relevant documents

Relevance Feedback

- Rocchio expansion formula

$$\vec{q}_{i+1} = \vec{q}_i + \frac{\beta}{R}\sum_{j=1}^{R}\vec{r}_j - \frac{\gamma}{S}\sum_{k=1}^{S}\vec{s}_k$$

  - $\beta + \gamma = 1$ (0.75,0.25);
    - Amount of 'push' in either direction
  - R: # rel docs, S: # non-rel docs
  - r: relevant document vectors
  - s: non-relevant document vectors
- Can significantly improve (though tricky to evaluate)

Collection-based Query Expansion

- Xu & Croft 97 (classic)
- Thesaurus expansion problematic:
  - Often ineffective
  - Issues:

Collection-based Query Expansion

- Xu & Croft 97 (classic)
- Thesaurus expansion problematic:
  - Often ineffective
  - Issues:
    - Coverage:
      - Many words – esp. NEs – missing from WordNet

Collection-based Query Expansion

- Xu & Croft 97 (classic)
- Thesaurus expansion problematic:
  - Often ineffective
  - Issues:
    - Coverage:
      - Many words – esp. NEs – missing from WordNet
    - Domain mismatch:
      - Fixed resources ‘general’ or derived from some domain
      - May not match current search collection
        - Cat/dog vs cat/more/ls

Collection-based Query Expansion

- Xu & Croft 97 (classic)
- Thesaurus expansion problematic:
  - Often ineffective
  - Issues:
    - Coverage:
      - Many words – esp. NEs – missing from WordNet
    - Domain mismatch:
      - Fixed resources ‘general’ or derived from some domain
      - May not match current search collection
        - Cat/dog vs cat/more/ls
  - Use collection-based evidence: global or local

Global Analysis

- Identifies word cooccurrence in whole collection
  - Applied to expand current query
  - Context can differentiate/group concepts

Global Analysis

- Identifies word cooccurrence in whole collection
  - Applied to expand current query
  - Context can differentiate/group concepts
- Create index of concepts:
  - Concepts = noun phrases (1-3 nouns long)

Global Analysis

- Identifies word cooccurrence in whole collection
  - Applied to expand current query
  - Context can differentiate/group concepts
- Create index of concepts:
  - Concepts = noun phrases (1-3 nouns long)
  - Representation: Context
    - Words in fixed length window, 1-3 sentences

Global Analysis

- Identifies word cooccurrence in whole collection
  - Applied to expand current query
  - Context can differentiate/group concepts
- Create index of concepts:
  - Concepts = noun phrases (1-3 nouns long)
  - Representation: Context
    - Words in fixed length window, 1-3 sentences
  - Concept identifies context word documents
- Use query to retrieve 30 highest ranked concepts
  - Add to query

Local Analysis

- Aka local feedback, pseudo-relevance feedback

Local Analysis

- Aka local feedback, pseudo-relevance feedback
- Use query to retrieve documents
  - Select informative terms from highly ranked documents
  - Add those terms to query

Local Analysis

- Aka local feedback, pseudo-relevance feedback
- Use query to retrieve documents
  - Select informative terms from highly ranked documents
  - Add those terms to query
- Specifically,
  - Add 50 most frequent terms,
  - 10 most frequent ‘phrases’ – bigrams w/o stopwords
  - Reweight terms

Local Context Analysis

- Mixes two previous approaches
  - Use query to retrieve top n passages (300 words)
  - Select top m ranked concepts (noun sequences)
  - Add to query and reweight

Local Context Analysis

- Mixes two previous approaches
  - Use query to retrieve top n passages (300 words)
  - Select top m ranked concepts (noun sequences)
  - Add to query and reweight
- Relatively efficient
- Applies local search constraints

Experimental Contrasts

- Improvements over baseline:
  - Local Context Analysis: +23.5% (relative)
  - Local Analysis: +20.5%
  - Global Analysis: +7.8%

Experimental Contrasts

- Improvements over baseline:
  - Local Context Analysis: +23.5% (relative)
  - Local Analysis: +20.5%
  - Global Analysis: +7.8%

- LCA is best and most stable across data sets
  - Better term selection that global analysis

Experimental Contrasts

- Improvements over baseline:
  - Local Context Analysis: +23.5% (relative)
  - Local Analysis: +20.5%
  - Global Analysis: +7.8%
- LCA is best and most stable across data sets
  - Better term selection that global analysis
- All approaches have fairly high variance
  - Help some queries, hurt others

Experimental Contrasts

- Improvements over baseline:
  - Local Context Analysis: +23.5% (relative)
  - Local Analysis: +20.5%
  - Global Analysis: +7.8%

- LCA is best and most stable across data sets
  - Better term selection that global analysis
- All approaches have fairly high variance
  - Help some queries, hurt others
- Also sensitive to # terms added, # documents

- Global Analysis

hypnosis	meditation	practitioners
dentists	antibodies	disorders
psychiatry	immunodeficiency-virus	anesthesia
susceptibility	therapists	dearth
atoms	van-dyke	self
confession	stare	proteins
katie	johns-hopkins-university	growing-acceptance
reflexes	voltage	ad-hoc
correlation	conde-nast	dynamics
ike	illnesses	hoffman

- Local Analysis

hypnot	hypnotiz	19960500
psychosomat	psychiatr	immun
mesmer	franz	suscept
austrian	dyck	psychiatrist
shesaid	tranc	professor
hallucin	18th	centur
hilgard	11th	unaccept
19820902	syndrom	exper
physician	told	patient

- LCA

hypnosis	brain-wave	ms.-burns
technique	pulse	reed
brain	ms.-olness	trance
hallucination	process	circuit
van-dyck	behavior	suggestion
case	spiegel	finding
hypnotizables	subject	van-dyke

What are the different techniques used to create self-induced hypnosis?

Passage Retrieval

- Documents: wrong unit for QA
  - Highly ranked documents
    - High weight terms in common with query
    - Not enough!

Passage Retrieval

- Documents: wrong unit for QA
  - Highly ranked documents
    - High weight terms in common with query
    - Not enough!
      - Matching terms scattered across document
      - Vs
      - Matching terms concentrated in short span of document
- Solution:
  - From ranked doc list, select and rerank shorter spans
  - Passage retrieval

Passage Ranking

- Goal: Select passages most likely to contain answer
- Factors in reranking:

Passage Ranking

- Goal: Select passages most likely to contain answer
- Factors in reranking:
  - Document rank

Passage Ranking

- Goal: Select passages most likely to contain answer
- Factors in reranking:
  - Document rank
  - Want answers!

Passage Ranking

- Goal: Select passages most likely to contain answer
- Factors in reranking:
  - Document rank
  - Want answers!
    - Answer type matching
      - Restricted Named Entity Recognition

Passage Ranking

- Goal: Select passages most likely to contain answer
- Factors in reranking:
  - Document rank
  - Want answers!
    - Answer type matching
      - Restricted Named Entity Recognition
  - Question match:
    - Question term overlap
    - **Span** overlap: N-gram, longest common sub-span
    - Query term **density:** short spans w/more qterms

Quantitative Evaluation of Passage Retrieval for QA

- Tellex et al.
- Compare alternative passage ranking approaches
  - 8 different strategies + voting ranker
- Assess interaction with document retrieval

Comparative IR Systems

- PRISE
  - Developed at NIST
  - Vector Space retrieval system
  - Optimized weighting scheme

Comparative IR Systems

- PRISE
  - Developed at NIST
  - Vector Space retrieval system
  - Optimized weighting scheme
- Lucene
  - Boolean + Vector Space retrieval
  - Results Boolean retrieval RANKED by tf-idf
    - Little control over hit list

Comparative IR Systems

- PRISE
  - Developed at NIST
  - Vector Space retrieval system
  - Optimized weighting scheme
- Lucene
  - Boolean + Vector Space retrieval
  - Results Boolean retrieval RANKED by tf-idf
    - Little control over hit list
- Oracle: NIST-provided list of relevant documents

Comparing Passage Retrieval

- Eight different systems used in QA
  - Units
  - Factors

Comparing Passage Retrieval

- Eight different systems used in QA
  - Units
  - Factors
- MITRE:
  - Simplest reasonable approach: baseline
  - Unit: sentence
  - Factor: Term overlap count

Comparing Passage Retrieval

- Eight different systems used in QA
  - Units
  - Factors
- MITRE:
  - Simplest reasonable approach: baseline
  - Unit: sentence
  - Factor: Term overlap count
- MITRE+stemming:
  - Factor: stemmed term overlap

Comparing Passage Retrieval

- Okapi bm25
  - Unit: fixed width sliding window
  - Factor: $$Score(q,d) = \sum_{i=1}^{N} idf(q_i) \frac{tf_{q_i,d}(k_1+1)}{tf_{q_i,d} + k_1(1-b+(b*\frac{|D|}{avgdl}))}$$
    - k1=2.0; b=0.75

Comparing Passage Retrieval

- Okapi bm25
  - Unit: fixed width sliding window
  - Factor: $$Score(q,d) = \sum_{i=1}^{N} idf(q_i) \frac{tf_{q_i,d}(k_1+1)}{tf_{q_i,d} + k_1(1-b+(b*\frac{|D|}{avgdl}))}$$
    - k1=2.0; b=0.75
- MultiText:
  - Unit: Window starting and ending with query term
  - Factor:
    - Sum of IDFs of matching query terms
    - Length based measure * Number of matching terms

Comparing Passage Retrieval

- IBM:
  - Fixed passage length
  - Sum of:
    - Matching words measure: Sum of idfs of overlap terms
    - Thesaurus match measure:
      - Sum of idfs of question wds with synonyms in document
    - Mis-match words measure:
      - Sum of idfs of questions wds NOT in document
    - Dispersion measure: # words b/t matching query terms
    - Cluster word measure: longest common substring

Comparing Passage Retrieval

- SiteQ:
  - Unit: n (=3) sentences
  - Factor: Match words by literal, stem, or WordNet syn
    - Sum of
      - Sum of idfs of matched terms
      - Density weight score * overlap count, where

Comparing Passage Retrieval

- SiteQ:
  - Unit: n (=3) sentences
  - Factor: Match words by literal, stem, or WordNet syn
    - Sum of
      - Sum of idfs of matched terms
      - Density weight score * overlap count, where

$$dw(q,d) = \frac{\sum_{j=1}^{k-1} \frac{idf(q_j) + idf(q_{j+1})}{\alpha \times dist(j, j+1)^2}}{k-1} \times overlap$$

Comparing Passage Retrieval

- Alicante:
  - Unit: n (= 6) sentences
  - Factor: non-length normalized cosine similarity

Comparing Passage Retrieval

- Alicante:
  - Unit: n (= 6) sentences
  - Factor: non-length normalized cosine similarity

- ISI:
  - Unit: sentence
  - Factors: weighted sum of
    - Proper name match, query term match, stemmed match

Experiments

- Retrieval:
  - PRISE:
    - Query: Verbatim quesiton
  - Lucene:
    - Query: Conjunctive boolean query (stopped)

Experiments

- Retrieval:
  - PRISE:
    - Query: Verbatim quesiton
  - Lucene:
    - Query: Conjunctive boolean query (stopped)
- Passage retrieval: 1000 word passages
  - Uses top 200 retrieved docs
  - Find best passage in each doc
  - Return up to 20 passages
    - Ignores original doc rank, retrieval score

Pattern Matching

- Litkowski pattern files:
  - Derived from NIST relevance judgments on systems
  - Format:
    - Qid answer_pattern doc_list
      - Passage where answer_pattern matches is correct
        - If it appears in one of the documents in the list

Pattern Matching

- Litkowski pattern files:
  - Derived from NIST relevance judgments on systems
  - Format:
    - Qid answer_pattern doc_list
      - Passage where answer_pattern matches is correct
        - If it appears in one of the documents in the list
- MRR scoring
  - Strict: Matching pattern in official document
  - Lenient: Matching pattern

Examples

- Example
  - Patterns
    - 1894 (190|249|416|440)(\s|\-)million(\s|\-)miles? APW19980705.0043 NYT19990923.0315 NYT19990923.0365 NYT20000131.0402 NYT19981212.0029
    - 1894 700-million-kilometer APW19980705.0043
    - 1894 416 - million - mile NYT19981211.0308
  - Ranked list of answer passages
    - 1894 0 APW19980601.0000 the casta way weas
    - 1894 0 APW19980601.0000 440 million miles
    - 1894 0 APW19980705.0043 440 million miles

Evaluation

- MRR
  - Strict: Matching pattern in official document
  - Lenient: Matching pattern
- Percentage of questions with NO correct answers

Evaluation

- MRR
  - Strict: Matching pattern in official document
  - Lenient: Matching pattern
- Percentage of questions with NO correct answers

	Strict				
	Lucene		PRISE		TREC
Algorithm	MRR	% Inc.	MRR	% Inc.	% Inc.
IBM	0.326	49.20%	0.331	39.60%	44.3%
ISI	0.329	48.80%	0.287	41.80%	41.7%
SiteQ	0.323	48.00%	0.358	40.40%	56.1%
MultiText	0.354	46.40%	0.325	41.60%	43.1%
Alicante	0.296	50.00%	0.321	42.60%	60.4%
bm25	0.312	48.80%	0.252	46.00%	n/a
stemmed MITRE	0.250	52.60%	0.242	58.60%	n/a
MITRE	0.271	49.40%	0.189	52.00%	n/a
Averages	0.309	49.15%	0.297	45.33%	n/a
Voting with IBM, ISI, SiteQ	0.350	39.80%	0.352	39.00%	n/a

Evaluation on Oracle Docs

Algorithm	# Incorrect	% Incorrect	MRR
IBM	31	7.18%	0.851
SiteQ	32	7.41%	0.859
ISI	37	8.56%	0.852
Alicante	39	9.03%	0.816
MultiText	44	10.19%	0.845
bm25	45	10.42%	0.810
MITRE	45	10.42%	0.800
stemmed MITRE	63	14.58%	0.762

Overall

- PRISE:
  - Higher recall, more correct answers

Overall

- PRISE:
  - Higher recall, more correct answers
- Lucene:
  - Higher precision, fewer correct, but higher MRR

Overall

- PRISE:
  - Higher recall, more correct answers
- Lucene:
  - Higher precision, fewer correct, but higher MRR
- Best systems:
  - IBM, ISI, SiteQ
  - Relatively insensitive to retrieval engine

Analysis

- Retrieval:
  - Boolean systems (e.g. Lucene) competitive, good MRR
    - Boolean systems usually worse on ad-hoc

Analysis

- Retrieval:
  - Boolean systems (e.g. Lucene) competitive, good MRR
    - Boolean systems usually worse on ad-hoc

- Passage retrieval:
  - Significant differences for PRISE, Oracle
  - Not significant for Lucene -> boost recall

Analysis

- Retrieval:
  - Boolean systems (e.g. Lucene) competitive, good MRR
    - Boolean systems usually worse on ad-hoc

- Passage retrieval:
  - Significant differences for PRISE, Oracle
  - Not significant for Lucene -> boost recall
- Techniques: Density-based scoring improves
  - Variants: proper name exact, cluster, density score

Error Analysis

- ‘What is an ulcer?’

Error Analysis

- ‘What is an ulcer?’
  - After stopping -> ‘ulcer’
  - Match doesn’t help

Error Analysis

- ‘What is an ulcer?’
  - After stopping -> ‘ulcer’
  - Match doesn’t help
  - Need question type!!
- Missing relations
  - ‘What is the highest dam?’
    - Passages match ‘highest’ and ‘dam’ – but not together
  - Include syntax?