Information Retrieval via Statistical Language Modeling

Matt Lease

Brown Laboratory for Linguistic Information Processing
Brown University
Providence, RI USA

BROWN

BLLIP

Outline

- Information Retrieval (IR) and the TF-IDF approach
- Statistical language modeling approach
  - Conceptual model & basic framework
  - Connection to TF-IDF
  - Simple extensions
- Evaluation: metrics, English data, and some results
- A few other directions, ideas, & questions

What is Information Retrieval (IR)?

- Finding information to satisfy a user's information need
  - Retrieving textual and/or non-textual information (e.g., images)
  - Variety of tasks: filtering, categorization, summarization, etc.
- For this talk, IR = ad hoc text retrieval
  - Given
    - A **collection** of text documents (or "docs")
    - A text **query** from a user
  - Goal: retrieve relevant documents from the collection
  - (Basically Google task without hyperlink structure or usage stats)
- This is a classic IR problem with a long history…

A Classic Approach: Vector Similarity & TF-IDF

- Queries and documents are vectors over collection vocabulary
  - Value of entry *i* = frequency of word w_i in the given query/document
  - Relevance defined by query/doc vector similarity (e.g. cosine)
- Refinement: weight words by TF * IDF
  - Term Frequency: relative importance of word in the query/doc
    - TF = frequency of word in query/doc
    - Length normalization: use normalized % rather than # of occurrences)
  - Inverse Document Frequency: specificity of the word in the collection
    - IDF = (# of docs in collection) / (# of docs the word occurs in)
- Pros: simple, fast, & effective in practice
- Cons: many little tweaks used with little theoretical justification
  - E.g. TF and IDF usually dampened

New Approach: Statistical Language Modeling

- Language Model (LM)
  - Defines probability distribution over possible word sequences (infinite)
  - Example model: unigram (bag-of-words) $P(w_{1:n} \mid \theta) = \prod_{i}^{n} P(w_i \mid \theta)$
  - Heavily used in speech recognition & machine translation
- Use in IR
  - Estimate a LM underlying each document (i.e. the θ above)
  - Document relevance ranked proportional to query likelihood under doc LMs
- Some Intuition (maybe…)
  - Often a query & relevant doc refer to the same information in different ways
  - However, if the doc were infinitely long (all possible paraphrases of the same information), presumably the query words would eventually appear in it
  - Thus, we assume query & doc are finite samples from some underlying LM which defines a probability distribution over all the ways to convey the same information
  - The challenge: estimating the "true" LM from the finite observed document

Smoothing

- Goal: Estimate "true" LM underlying an observed document
- Step 1: Maximum-Likelihood Estimation (MLE)
- Problem: MLE leaves no mass to unseen words (they're not likely, right?)
- Approach: "smooth" the MLE distribution (flatten peaks, fill valleys)
  - Robin Hood: rob the rich (i.e. seen words) to pay the poor (i.e. unseen words)
  - But… Who to rob and how much? Who to give to and how much?
- Heavily studied problem and many methods (C&G'98)
- For IR, usually interpolate doc LM with a LM for the entire collection
  - Collection LM estimated from many documents so more robust
  - To a rough approximation, collection can serve as held-out data to model differences between ML estimate and "truth"

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$

1. Bayes' rule

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$

1. Bayes' rule
2. Uniform doc prior

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$
4. $P(q_i \mid d) = (1-\lambda)P_{ML}(q_i \mid d) + \lambda P(q_i \mid C)$

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM
4. Linear smoothing

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$
4. $P(q_i \mid d) = (1-\lambda)P_{ML}(q_i \mid d) + \lambda P(q_i \mid C)$
5. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log P(q_i \mid d) + \sum_{i:c(q_i,d)=0} \log \lambda P(q_i \mid C)$

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM
4. Linear smoothing
5. Seen / unseen

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$
4. $P(q_i \mid d) = (1-\lambda)P_{ML}(q_i \mid d) + \lambda P(q_i \mid C)$
5. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log P(q_i \mid d) + \sum_{i:c(q_i,d)=0} \log \lambda P(q_i \mid C)$
6. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \mathrm{P}(w_i \mid C)} + \sum_i \log \lambda P(q_i \mid C)$

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM
4. Linear smoothing
5. Seen / unseen
6. Algebra

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$
4. $P(q_i \mid d) = (1-\lambda)P_{ML}(q_i \mid d) + \lambda P(q_i \mid C)$
5. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log P(q_i \mid d) + \sum_{i:c(q_i,d)=0} \log \lambda P(q_i \mid C)$
6. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \, \mathrm{P}(w_i \mid C)} + \sum_i \log \lambda P(q_i \mid C)$
7. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \, \mathrm{P}(w_i \mid C)} + n \log \lambda + \sum_i \log P(q_i \mid C)$

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM
4. Linear smoothing
5. Seen / unseen

6-7. Algebra

Formal Machinery and TF-IDF Comparison

1. $P(d \mid q) \propto P(q \mid d)P(d)$
2. $P(d \mid q) \propto P(q \mid d)$
3. $P(d \mid q_{1:n}) \propto \prod_i P(q_i \mid d) \propto \sum_i \log P(q_i \mid d)$
4. $P(q_i \mid d) = (1-\lambda)P_{ML}(q_i \mid d) + \lambda P(q_i \mid C)$
5. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log P(q_i \mid d) + \sum_{i:c(q_i,d)=0} \log \lambda P(q_i \mid C)$
6. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \mathrm{P}(w_i \mid C)} + \sum_i \log \lambda P(q_i \mid C)$
7. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \mathrm{P}(w_i \mid C)} + n \log \lambda + \sum_i \log P(q_i \mid C)$
8. $P(d \mid q_{i:n}) \propto \sum_{i:c(q_i,d)>0} \log \frac{P(q_i \mid d)}{\lambda \mathrm{P}(w_i \mid C)} + n \log \lambda$

~ Term Frequency

~ Length Normalization

~ Inverse Document Frequency

1. Bayes' rule
2. Uniform doc prior
3. Unigram LM
4. Linear smoothing
5. Seen / unseen
6-7. Algebra
8. Ignore constant

Outline

- IR and a classic approach to it: TF-IDF
- Statistical language modeling approach
  - Conceptual model & basic framework
  - Connection to TF-IDF
  - Simple extensions
- Evaluation: metrics, English data, and some results
- A few other directions, ideas, & questions

Document Expansion

- Brill: "There's no data like more data"
  - Could estimate a doc's "true" LM better if doc were longer (bigger sample)
- Problem: usually difficult/expensive to acquire the right kind of data
  - Pay someone to manually extend all docs in collection?!?
- Strategy: can often acquire "similar" data cheaply, and it often helps
- Here, simulate longer doc by mixing its LM with "similar" doc LMs
  - Set mixing weights proportional to similarity (norm. cosine) and doc length
- Interpolation with collection as document expansion
  - An alternative explanation for collection interpolation
  - one-size-fits-all vs. more specific smoothing

Query Expansion (Pseudo-Relevance Feedback)

- Intuition: address paraphrase mismatch by expanding query
- First idea: just add words to query
  - coarse granularity
- Instead reformulate our setup slightly
  - Estimate a LM for query as well as documents
  - Compute relevance as similarity between doc LM and the query LM
  - (we're back to traditional vector space model, only with better statistical theory and tools for weight estimation)
- But what words to add and how to weight them?
- Now a simple trick (semi-supervised learning)
  - Run query
  - Assume first N documents retrieved are relevant
  - Interpolate query LM with these N doc LMs
  - Re-run query
- Generally improves retrieval significantly, but sensitive to parameters
  - Parameter-free approach recently given in (Tao & Zhai, SIGIR'06)

Outline

- IR and a classic approach to it: TF-IDF
- Statistical language modeling approach
  - Conceptual model & basic framework
  - Connection to TF-IDF
  - Simple extensions
- Evaluation: metrics, English data, and some results
- A few other directions, ideas, & questions

Evaluation

- NIST TREC competitions have driven data/metrics for English IR
- Documents are manually annotated as relevant or not (binary)
  - Only subset of collections annotated (possible recall errors)
- Single query
  - Precision: # relevant docs retrieved / total # docs retrieved
  - Recall: # relevant docs retrieved / total # relevant docs
  - Given doc ranking, let P*i* be precision at the rank of *i*th relevant doc
    - E.g. If 1st, 2nd, & 5th ranked doc relevant: P1=1/1, P2=2/2, P3=3/5
    - For relevant docs not retrieved at all, assume P*i* = 0
  - Average Precision (AP) = P*i* averaged across all of the query's relevant docs
- Multiple queries
  - Mean Average Precision (MAP): AP averaged over all queries

English Queries

- NIST queries written at multiple levels of detail (fields)
  - TREC primarily had systems treat the *description* field as the input query
  - Research tends to use the *title* field as the input query (more like typical keyword query)
- NOTE: underlining *indicates new information not found in coarser fields of query*
- Example 1
  - **Title**: Antitrust Cases Pending
  - **Description**: Document discusses a pending antitrust case
  - **Narrative**:

    To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as well as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.
- Example 2
  - **Title**: Acquisitions
  - **Description**: Document discusses a currently proposed acquisition involving a U.S. company and a foreign company
  - **Narrative:**

    To be relevant, a document must discuss a currently proposed acquisition (which may or may not be identified by type, e.g., merger, buyout, leveraged buyout, hostile takeover, friendly acquisition). The suitor and target must be identified by name; the nationality of one of the companies must be identified as U.S. and the nationality of the other company must be identified as NOT U.S.
- Also some other fields that identify and define relevant terms

English Collections

	Size (megabytes)	# Docs	Median # Words/Doc	Mean # Words/Doc
Disk 1				
Wall Street Journal, 1987–1989	267	98,732	245	434.0
Associated Press newswire, 1989	254	84,678	446	473.9
Computer Selects articles, Ziff-Davis	242	75,180	200	473.0
Federal Register, 1989	260	25,960	391	1315.9
abstracts of U.S. DOE publications	184	226,087	111	120.4
Disk 2				
Wall Street Journal, 1990–1992 (WSJ)	242	74,520	301	508.4
Associated Press newswire (1988) (AP)	237	79,919	438	468.7
Computer Selects articles, Ziff-Davis (ZIFF)	175	56,920	182	451.9
Federal Register (1988) (FR88)	209	19,860	396	1378.1
Disk 3				
San Jose Mercury News, 1991	287	90,257	379	453.0
Associated Press newswire, 1990	237	78,321	451	478.4
Computer Selects articles, Ziff-Davis	345	161,021	122	295.4
U.S. patents, 1993	243	6,711	4445	5391.0
Disk 4				
the *Financial Times*, 1991–1994 (FT)	564	210,158	316	412.7
Federal Register, 1994 (FR94)	395	55,630	588	644.7
Congressional Record, 1993 (CR)	235	27,922	288	1373.5
Disk 5				
Foreign Broadcast Information Service (FBIS)	470	130,471	322	543.6
the *LA Times*	475	131,896	351	526.5

BROWN

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392
LM, Dirichlet smoothing (DS) with collection LM	0.2567

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392
LM, Dirichlet smoothing (DS) with collection LM	0.2567
LM, DS, document expansion (DE)	0.2671

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392
LM, Dirichlet smoothing (DS) with collection LM	0.2567
LM, DS, document expansion (DE)	0.2671
LM, DS, query expansion (QE)	0.2716

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392
LM, Dirichlet smoothing (DS) with collection LM	0.2567
LM, DS, document expansion (DE)	0.2671
LM, DS, query expansion (QE)	0.2716
LM, DS, DE+QE	0.2809

Some numbers

Allan et al. '99: *INQUERY and TREC-8*
Namba & Igata'99: *Fujitsu Laboratories TREC-8 Report*
Tao et al.'06: *Language Model Information Retrieval with Document Expansion*

- Collection: TREC 6-8 (~Disks 4-5, 0.5M docs)
- Queries: TREC 401-450 (~5K relevant docs)

Method	**MAP**
TREC-8 InQuery (TF-IDF based) system	0.2325
LM, linear interpolation with collection LM	0.2392
LM, Dirichlet smoothing (DS) with collection LM	0.2567
LM, DS, document expansion (DE)	0.2671
LM, DS, query expansion (QE)	0.2716
LM, DS, DE+QE	0.2809
TREC-8 Fujitsu system	0.2853

Outline

- IR and a classic approach to it: TF-IDF
- Statistical language modeling approach
  - Conceptual model & basic framework
  - Connection to TF-IDF
  - Simple extensions
- Evaluation: metrics, English data, and some results
- A few other directions, ideas, & questions

Xu et al'01: *Evaluating … cross-lingual information retrieval*
Lavrenko et al.'02: *Cross-lingual relevance models*

Task: Multilingual Document Retrieval

- Problem: relevant documents may not be in user's native language
- Ideal solution
  - User expresses query in native language
  - System retrieves relevant docs in any language
  - System translates relevant docs to native language (as needed)
- Track at CLEF'07 (Cross-Language Evaluation Forum)
  - English and Czech document collections (and others)
  - Queries can be issued in English or Czech (and a few other languages)
    - Also a mono-lingual Czech task, but not for English
- A few questions
  - How robust is LM approach cross-linguistically (e.g. mono-lingual IR in non-English)?
  - Do we observe the same trends between data characteristics and performance?
- Previous work in LM paradigm: Xu'01, Lavrenko'02, …?

Task: Cross-Language Speech Retrieval

- Another track at CLEF'07
  - "Documents" are conversations in English and Czech
  - Queries in English or Czech (and a few other languages)
- Chance to apply Brown's previous work in disfluency analysis and parsing of (English) conversational speech
  - Charniak and Johnson'01: detecting speech repairs and parsing
  - Johnson and Charniak'04: better repair detection
  - Lease, Johnson, and Charniak'06: detect fillers & repairs (better)

Interesting research: “Translation” Models

- Queries and docs are quite different written artifacts, so why not model this (in query likelihood)?
- Approach: define a “translation” model which generates queries from documents (e.g. noisy channel)
- Previous work
  - Berger & Lafferty’99, Jin et al. ‘02, Cao et al.’05, …?
- Results
  - improves over the basic LM methods
  - computationally expensive

Interesting research: Topic/Aspect Models

- Documents often span more than one topic, yet our model represents each document by exactly 1 LM
- Instead, could model doc as mixture of 1-to-N topic LMs
  - Generative process
    - pick length of document
    - for each word position
      - pick a topic
      - pick a word given topic
- Previous work: Hofmann'99, Blei et al'02, …?
- Results?

Summary

- Strengths
  - Solid statistical foundation
  - Lots of existing LM methodology to build on
  - Relatively simple models perform well empirically
- Challenges
  - Accurate as full-fledged traditional systems?
  - Efficient as traditional systems?
- What's next?
  - Many interesting directions to explore

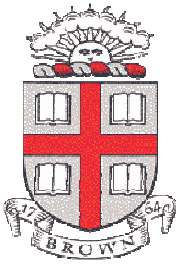
Acknowledgement

This talk draws on the following great tutorial:

ChengXiang Zhai. *Statistical Language Models for Information Retrieval.* SIGIR'06 Tutorial. Slides at http://sifaka.cs.uiuc.edu/lmir.

Zhai is also presenting a similar tutorial at HLT/NAACL'07

For more information…



Brown Laboratory for Linguistic Information Processing (BLLIP)
http://bllip.cs.brown.edu

Project is being undertaken at the
Institute of Formal and Applied Linguistics (ÚFAL)
at the Faculty of Mathematics and Physics, Charles University, Prague
http://ufal.mff.cuni.cz

Support for this work comes from the
National Science Foundation
Partnerships for International Research and Education (PIRE)