Information Retrieval and Web Search
Question Answering Systems
Summarization
Machine Translation

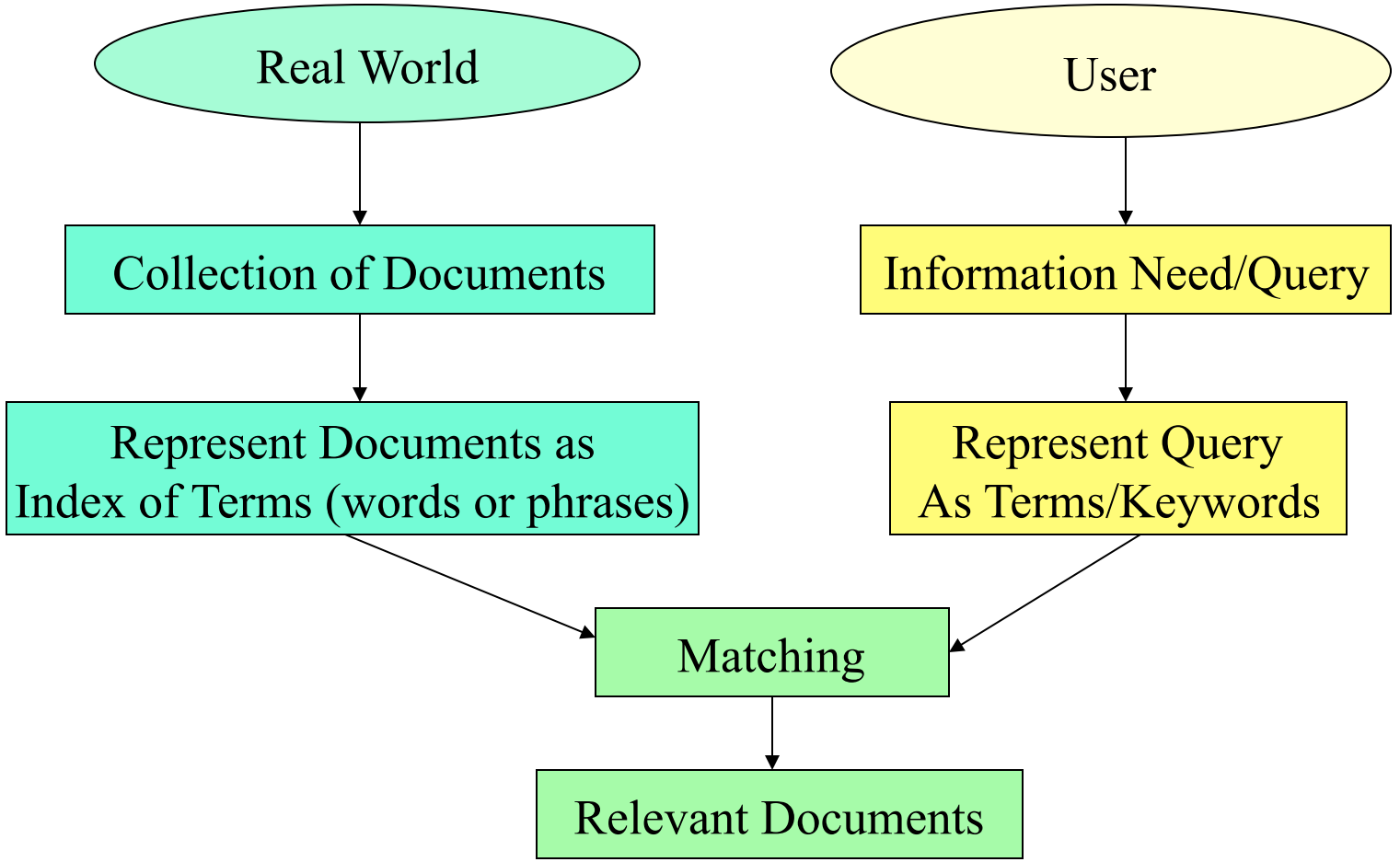
What is Information Retrieval

- Gerard Salton, 1968:
  Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information
- Manning, Raghavan and Schutze, 2008:
  Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)
    - "Document" is the generic term for an information holder (book, chapter, article, webpage, etc)

What is tough about IR?

- One issue is how to **represent documents** so others might retrieve them
  - Need to match the text of the document with the query
  - In full, free-text systems this is an issue because documents and queries are expressed in language
    - and language is synonymous and polysemous
    - methods for solving the language issue are difficult
  - Sometimes called the **vocabulary gap** or mismatch
- Given the retrieval of some documents how to decide which ones are **most relevant** to the user's query
  - Most often implemented as a **ranking** of the resulting documents

Typical Information Retrieval System



IRS Issues

- What is relevance?
  - whether the user considers the document to be “on topic” or to satisfy the information need
- To improve relevance, researchers propose retrieval models and test how well they work
  - including a ranking algorithm to order the list of retrieved documents
- Evaluation – how well do the retrieval models work?
  - Precision – the proportion of the retrieved documents that are relevant
    - accuracy, or trying not to retrieve non-relevant documents
  - Recall – the proportion of relevant documents that are retrieved
    - coverage, trying to get as much good stuff as possible

IRS Research

- Traditional IR System research assumes that a user is interested in finding out information on a particular topic
- TREC collections and research experiments at NIST/Text REtrieval Conference (TREC): http://trec.nist.gov
  - build IR systems with different retrieval models
  - test against a standard collection of newswire documents
  - human evaluators judge relevant documents
  - report system evaluations in terms of precision and recall
  - Example type of query:

    I am interested in all documents that discuss oil reserves and current attempts to find new reserves, particularly those that discuss the international financial aspects of the oil production process.

Information Needs

- Other branches of research focus on the user and whether the user's underlying information seeking is satisfied
- Early theories by Belkin, Oddy, etc.
  - Functions of the retrieval system to model the user's information need in an interactive retrieval session:
    - Characterize User
    - Get initial information need
    - Develop need context
    - Formulate information need
    - Conduct search for documents
    - Evaluate results
    - Feedback from user

Constructing the Index

- Process documents and identify terms to be indexed
  - Terms are often just the words
    - Usually stemming is applied and stop words removed
  - Sometimes basic noun phrases are also added, particularly proper names
- Compute weights of terms, depending on model definition
- Build index, a giant dictionary mapping terms to documents
  - For each term,
    - keep a list of documents that it occurs in
    - weights

Models

- Vector Space Models
  - Widely used weights known as TF/IDF (term frequency / inverted document frequency)
    - TF – frequency of the term in the document (normalized by document length)
      - Intuition: more frequently occurring terms are more important
    - IDF – invert the document frequency, the number of documents in the collection that the term occurs in
      - Intuition: terms occurring in all documents are less important to distinguish which ones are relevant to the query
- Other models include
  - Probabilistic models
  - Language models
  - Boolean models

Queries and matching

- Natural language queries are converted to terms, usually called keywords
  - In web search, typical queries are keywords already
- Query terms are used to retrieve documents from the index
- Model defines how to match query terms to documents, using the weights, and usually resulting in a score for each document
- Documents are returned in order of relevance

Evaluation

- Human judgments as to whether returned documents are relevant to the query

	Relevant	**Non-Relevant**
Retrieved	a (true positives)	b (false positives)
Non-Retrieved	c (false negatives)	d (true negatives)

Precision = a / a + b

Recall = a / a + c

Another Evaluation Measure

- The F-measure is a combination of recall and precision, averaged using the harmonic mean
  - Let P be precision and R be recall

    $F = (\beta^2 + 1) PR / (\beta^2) P + R$

  - Typically, the measure is used for $\beta = 1$, giving equal weight to precision and recall

    $F_{\beta=1} = 2 PR / P + R$

Evaluating Ranked Retrieval Results

- Precision-Recall Curves
  - Given the top k ranked documents, compute precision and recall
  - Plot precision vs. recall giving "sawtoothed" curve
    - or give precision vs. recall at 11 positions of recall
- Average Precision
  - Compute the average of precision over a sequence of recall levels
- Interpolated Average Precision
  - Interpolated precision at a recall level r is defined as the highest precision found at any recall level r' >= r.

Sample “Recall Level Precision Averages” Table

Recall Level Precision Averages	
Recall	Precision
0.00	0.5857
0.10	0.3927
0.20	0.3252
0.30	0.2799
0.40	0.2521
0.50	0.2131
0.60	0.1776
0.70	0.1395
0.80	0.0885
0.90	0.0415
1.00	0.0118
Average precision over all relevant docs	
Non-interpolated	**0.2109**

Text Retrieval Conference (TREC)

- Co-sponsored by the National Institute of Standards and Technology (NIST) & the Defense Advanced Research Projects Agency (DARPA)
  - Begun in 1992
- Purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
  - Provides document collections, queries and human judges
  - Main IR track was called the "Ad-Hoc Retrieval Track"
- Has grown in the number of participating systems and the number of tracks each year.
  - Tracks have included cross-language retrieval, filtering, question answering, interactive, web, novelty, video, …

Improving Retrieval

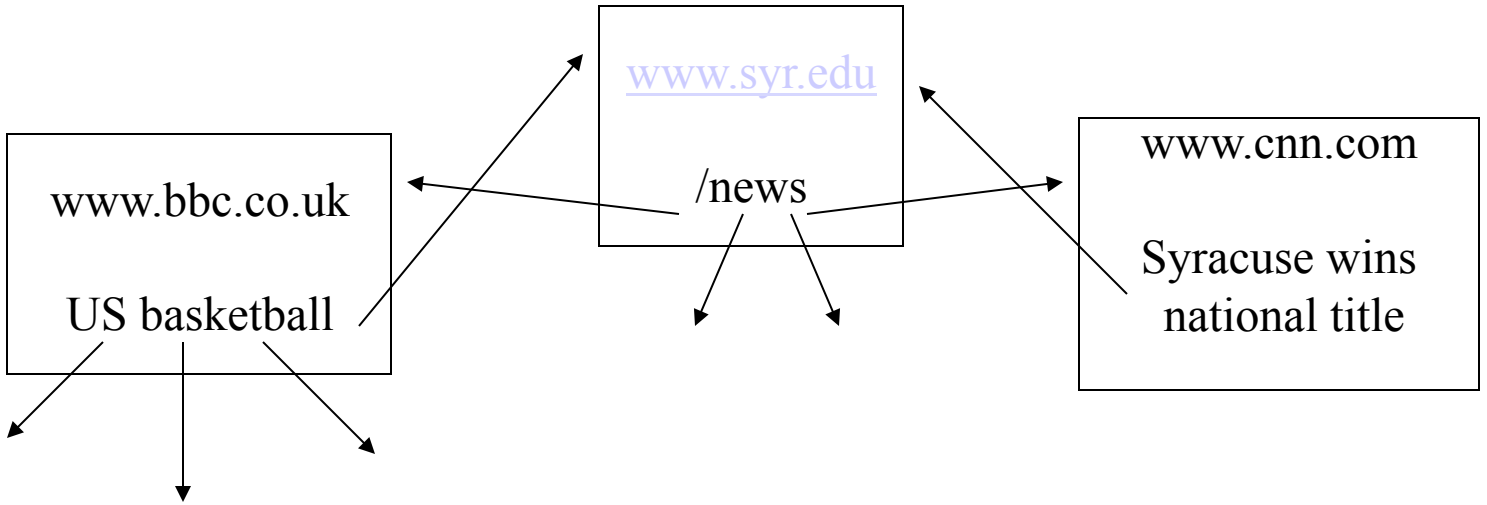
- Query expansion, adding semantically similar words or context words
  - Results are mixed
- Relevance Feedback
  - One technique consistently shown to improve retrieval
  - Human relevance feedback – after human has selected a few really relevant documents, add terms from those documents to the query
  - Pseudo-relevance feedback
    - Perform one retrieval and assume that the top n documents are relevant
    - Use those documents to add terms to the query

Web Search

- With the advent of the Web, basic IR was applied to this scenario of linked documents world wide
- Why/How would IR be different on the Web
  - Web behaves large, IR databases behave small
  - Databases are fielded at predictable times, sites update whenever
  - Collection frequencies needed for Inverse Document Frequency (IDF) are so impermanent
  - Quality control of documents on the web is not present
  - Links may not be permanent
  - No such thing as a complete inverted file for the entire web – many hidden pages (Deep Web)
  - Impact of pay for ranking
  - Web search companies do not reveal all of what they do with documents and queries

Web Graph

- View the collection of static web pages as a graph with "hyperlinks" between them
- Hyperlink in HTML, given by the anchor tag, will give the URL of another web page
  - in-degree is the number of links coming to a page from other pages
  - out-degree is the number of links on the page



Web Crawling

- In order to build an index of documents for web search, the web crawler, or spider, has to locate documents
- Required Features:
  - Robustness – it must not get stuck in dead ends or loops
  - Politeness – it must not overwhelm any web server with too fast or too many requests
    - web servers set politeness policies
- Desired Features
  - Quality – should try to give "useful" pages priority
  - Freshness – should obtain updated pages so that the web index has a fairly current version of the web page
  - Performance and efficiency, scalability, operate in a distributed fashion

Document Parsing

- Find content and process into tokens for traditional use in IR indexing
  - content may be text in-between tags
  - image tags may have text attributes to describe the image
  - may discard javascript and other computational elements
  - may even try to discard "noisy" text in the form of web site navigation, standard copyright notices, etc.
    - one technique is to observe that real content text has fewer tags per token than non-content text
- Keywords may be added to the document that don't appear directly in the content
  - metadata tags may have keywords
  - special weights may be added for tokens appearing in header tags
  - anchor text from other pages (see next slide)
- Look at on-line example "view page source"

Anchor Text

- Sometimes the text content of a web page does not contain generally descriptive words for that page
  - example: home page for IBM does not contain the word “computer”
  - home page for Yahoo does not contain the word “portal”
- Generally descriptive words may be found in anchor text of links, or even near it, that occur in other pages

   Big Blue

  example of a large computing firm

   Big Blue

  an example of a large computing firm is
   here
  - typically, we disregard anchor text words such as “click” and “here”

Link Analysis for Document Ranking

- Link analysis can be viewed as a development of citation analysis for the web
  - Bibliographic citation analysis used book and article references
  - Bibliometric analysis of bibliographic citation links
    - Web examples: Web of Science from ISI / Citeseer
- The intuition behind link analysis is that a hyperlink from page A to page B represents an endorsement of page B, by the creator of page A.
  - not true for some links, such as links to adminstrative notices on corporate websites - “internal” links are typically discounted.
- Two major algorithms, PageRank and HITS, that give scoring weights for web pages
  - such weights are combined with other weights from content tokens and many other ranking criteria

Additional Criteria for Ranking

- Popularity – what are the current topics of the day?
- Click-through results – statistics about which pages users click-on after getting ranked results can inform ranking algorithms to improve later rankings
- Context – keep track of the user's interests
  - What do other users like this one like?
- Learning to rank – use machine learning on ranked relevance results to improve rankings

Question Answering (QA)

- IR assumes that the user wants an entire document, QA assumes that the user wants a small, focused text result that answers a question
- Answers may not only be phrased with different vocabulary, but they may be implied by other statements
  Q: *What year did Marco Polo travel to Asia?*
  A: *Marco polo divulged the truth after returning in 1292 from his travels, which included several months on Sumatra.*
- QA systems must apply more NLP analysis to text in order to find the answers to questions
- The document collections of QA systems range in size from the Web to a targeted collection of documents such as a company's product documents

Typical QA System

- QA systems apply a two-step strategy
    - The document collection of a QA systems is too large to apply the (time-consuming) NLP processing to all the documents
  - First, use IR to retrieve a set of relevant documents
  - Then process those documents with NLP techniques
  - Identify answers in the documents

Factoid Questions

- *Where is Belize located?*
- *What type of bridge is the Golden Gate Bridge?*
- *What is the population of the Bahamas?*
- *How far away is the moon?*
- *What is Francis Scott Key best known for?*
- *What state has the most Indians?*
- *Who invented the paper clip?*
- *How many dogs pull a sled in the Iditarod?*
- *Where did bocci originate?*
- *Who invented the electric guitar?*
- *Name a flying mammal?*
- *How many hexagons are on a soccer ball?*
- *Who is the leader of India?*

TREC Question Answering Track

- Goal: encourage research into systems that return actual answers
  - From 1999 – 2004 in various forms
- Questions were short and fact based
  - From Encarta and Excite search logs
- Extract or construct answer from set of documents
- For each question
  - Supply up to five ranked answer submissions
  - With the most likely answer ranked first
  - Answer strings were evaluated by NIST's human assessors for correctness
- Evaluation: Mean-Reciprocal Rank (MRR) – score given is the reciprocal of the rank of the first correct answer

Approaches

- Question Classification
  - Expected answer type
    - "Who founded Virgin Airlines?" Expected type: Person
    - "What Canadian city has the largest population?" City
  - Question types
    - Definition questions, special forms such as birth and death dates
- Query Reformulation
  - Rephrase query as a partial answer "Virgin Airlines was founded by X" and try to find semantically similar sentences
- Separate document into passages and rank them
  - Number of words semantically similar to query words
  - Contains expected answer type
  - Proximity of query words and semantically similar words to expected answer types

Approaches

- Answer pattern approaches
  - Define patterns of sentences to answer the question
  - Use parsed sentences to match those patterms
    - Patterns include subject/verb forms
    - Appositions
    - Many others
- The Web approach
  - Since there are potentially millions of documents on the Web that could answer this question, assume that there is one that answers the question in the same form and vocabulary as the question
  - Search the web to find the answer in the right form

More Complex Questions

- Many questions go beyond asking for simple facts
  - Comparison questions
  - Questions that involve deeper understanding of issues
- These questions cannot typically be answered by a simple phrase or sentence
- Approaches focus on identifying an answer passage

CNLP KAAS for AIDE

- Knowledge Acquisition & Access System is a domain-specific QA
  - National Aeronautic & Space Agency funding
- Collection
  - **Textbooks, technical papers / reports, websites**
  - **Pre-selected for relevance and pedagogical value**
- Community of users focused on specific tasks
  - For undergrad students from 2 universities majoring in Aerospace Engineering
  - Students using AIDE for a course can ask questions while working in teams or on own
- Where QA system must function:
  - In **real** time, not batch mode
  - On **real** users' **real** world questions
  - With **real**, not surrogate assessments of relevance
- Used in a collaborative learning environment

Sample Questions from Real Users

- *How difficult is it to mold and shape graphite-epoxies compared with alloys or ceramics that may be used for thermal protective applications?*
- *How dsose the shuttle fly?*
- *Do welding sites yield any structural weaknesses that could be a threat for failure?*
- *Are Thermal Protection systems of spacecrafts commonly composed of one panel or a collection of smaller tiles?*
- *How can the aerogels be used in insulation of holes in TPS?*

- Two-stage QA model
  - **1st – passage retrieval using expanded query representation**
  - **2nd – selection of answer-providing passages based on generic + specialized entities & relations**

Summarization

- *Text summarization is the process of distilling the most important information from a text to produce an abridged version for a particular task and user*
  - Definition adapted from Mani and Maybury 1999
- Types of summaries in current research:
  - Outlines of any document
  - Abstracts of a scientific article
  - Headlines of a news article
  - Snippets summarizing a Web page or a search engine results page
  - Action items or other summaries of a business meeting
  - Summaries of email threads
  - Compressed sentences for simplified or clarified text
  - Single-document vs. multi-document summarization

Typical approaches to general problem

- Currently, a true re-phrasing overall summary is not yet achievable. Most systems primarily select sentences from a document and do some rephrasing
  - **Content Selection**
    - Identify the sentences of clauses to extract
  - **Information Ordering**
    - How to order the selected units
  - **Sentence Realization**
    - Perform cleanup on the extracted units so that they are fluent in their new context.

Content Selection

- Simple approach is to select sentences that have more informative words according to saliency defined from a topic signature of the document
- Centroid-based summarization uses log-likelihood ratios for words, computing the probability of observing the word in the input more often than in the background corpus
- Other centrality methods try to rank the sentences according to a centrality score
- Methods based on rhetorical parsing use coherence relations to identify satellite and nuclues sentences
- Machine learning methods use features based on
  - Position, cue phrases, word informativeness, sentence length, cohesion (computing lexical chains of the document)

Summarization Evaluation

- Extrinsic (task-based) evaluation: humans are asked to rate the summaries according to how well they are enabled to perform a specific task
- Intrinsic (task-independent) evaluation
  - Human judgments to rate the summaries
  - ROUGE
    - Humans generate summaries for a document collection
    - System-generated summaries are rated according to how close they come to the human-generated summary
    - Measures have included unigram overlap, bigram overlap, and longest common subsequence
  - Pyramid method
    - Humans identify "units of meaning" and then an overlap measure is computed

Machine Translation

- Translating text from one language to another is a task challenging even for humans to try to fully capture the style and nuanced meaning of the original
- While research focuses on trying to produce the fully-automatic, high-quality translation, there are many tasks for which a rough translation is sufficient
- The differences between languages include systematic differences that can be modeled in some way and idiosyncratic and lexical differences that must be dealt with one by one.

Why MT is hard

- Given the Japanese phrase
  fukaku hansei shite orimasu
- If this is translated to English as
  we apologize
  it is not faithful to the original meaning
- But if we translate it as
  we are deeply reflecting (on our past behavior, and what we did wrong, and how to avoid the problem next time)

  the translation is not fluent.

  Example from Jurafsky and Martin text.

Classical MT

- In this line of MT research, approaches can be classified according to the level of unit of translation
  - See the Vauquois triangle
  - Direct translation uses a word translation approach
  - Syntactic and semantic transfer approaches use syntactic phrase and semantic units, respectively, as the unit of translation

Statistical Approaches

- Build probabilistic models of faithfulness and fluency and combine the models to produce the most probable translation.
- Modeled as a noisy channel "pretend that the foreign input F is a corrupted version of the target language output E and the task is to discover the hidden sentence E that generated the observed sentence F."
- Requires three components
  - Language model to compute P(E)
  - Translation model to compute P(F|E)
  - Decoder, which is given F and produces the most probable E
    - Usually phrase-based

Alignment and Parallel Corpora

- All translation models are based on probabilities of word alignment
- Word alignment models are automatically trained from parallel corpora
- Hansard corpora work best for this
  - Translations of official government documents
    - Canadian parliament documents for French, English and a variety of native American languages
    - United Nations proceedings documents
  - Literary parallel corpora are not as suitable because of the stronger presence of literary devices, such as metaphor

MT Evaluation

- Human raters can evaluate along the two dimensions of fluency and fidelity (and there are several individual metrics for each of these dimensions)
- BLEU automatic evaluation system
  - Evaluation corpus contains human generated translations
  - Metrics evaluate how closely the system-generated translations correspond to the human ones