Challenges in Information Retrieval and Language Modeling

Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002

James Allan (editor), Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft (editor), Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, ChengXiang Zhai

Summary

Information retrieval (IR) research has reached a point where it is appropriate to assess progress and to define a research agenda for the next five to ten years. This report summarizes a discussion of IR research challenges that took place at a recent workshop.

The attendees of the workshop considered information retrieval research in a range of areas chosen to give broad coverage of topic areas that engage information retrieval researchers. Those areas are retrieval models, cross-lingual retrieval, Web search, user modeling, filtering, topic detection and tracking, classification, summarization, question answering, metasearch, distributed retrieval, multimedia retrieval, information extraction, as well as testbed requirements for future work. The potential use of language modeling techniques in these areas was also discussed.

The workshop identified major challenges within each of those areas. The following are recurring themes that ran throughout:

- *User and context sensitive retrieval*
- *Multi-lingual and multi-media issues*
- *Better target tasks*
- *Improved objective evaluations*
- *Substantially more labeled data*
- *Greater variety of data sources*
- *Improved formal models*

Contextual retrieval and global information access were identified as particularly important long-term challenges.

1. Introduction

What is information retrieval? Early definitions, dating from the 1960's, emphasize the very general nature of the task. For example, in Salton's classic textbook[*]:

> Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

In that textbook, information retrieval is assumed to also include database systems and question answering systems, and information is construed to mean documents, references, text passages, or facts.

Over the 1970's and 1980's, much of the research in IR was focused on document retrieval, and the emphasis on this task in the Text Retrieval Conference (TREC) evaluations of the 1990's has further reinforced the view that IR is synonymous with document retrieval. Web search engines are, of course, the most common example of this type of IR system.

The enormous increase in the amount of online text available and the demand for access to different types of information have, however, led to a renewed interest in a broad range of IR-related areas that go beyond simple document retrieval, such as question answering, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), software engineering, chemical and biological informatics, text structuring, text mining, and genomics. Salton's general definition is even more applicable now than it has been in the past.

---

[*] G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill, New York, 1968.

One common theme that has been used to distinguish IR-related research from research in database systems is that the information that is retrieved is derived from "unstructured" data sources. In the past, this distinction has been very clear, but if marked-up text is regarded as "semi-structured" and in the domain of database systems, then the boundary between the two areas becomes less obvious. Given the number of papers in recent database conferences on nearest-neighbor and similarity search, distributed search, Web search, and information extraction, it seems apparent that IR and database systems now have many common interests.

The huge success of Web search engines such as Google might lead some to question the need for extensive IR research. There are a number of possible answers to this question, but here are some major points:

- *Web search and IR are not equivalent*. As mentioned previously, IR encompasses many types of information access. Web search is only part (although an important part) of this spectrum of information systems.

- *Web queries do not represent all information needs*. A broad range of information access technologies are being created to address the diversity of information needs of people in different contexts. If we focus only on the current mix of queries in Web search engine logs, many of those information needs will not be addressed.

- *Web search engines are effective for some types of queries in some contexts*. Retrieval experiments in the TREC environment, and commercial success, demonstrate that, for a very popular type of query (find the right home page), retrieving the pages containing all the query words and then ranking them according to other features based on links, anchor text, URLs, and HTML tags is very effective. For other types of queries, and in other environments (e.g. corporate), this approach to ranking is less successful.

These factors, plus the resurgence of interest in formal, statistical methods for language-related tasks such as IR, make this an appropriate time to reassess and more clearly define the IR research agenda. To respond to this need, a group of IR and language technology researchers met at the University of Massachusetts Amherst to discuss IR research challenges. This report summarizes those discussions. The workshop was the second in a series funded by ARDA[*] that is focused on the language modeling approach to IR. For this workshop, the first priority was to identify the research challenges overall and the second was to discuss how language modeling could help to address these challenges.

The discussion of IR challenges is broken down by topic. In each topic area, there is a short overview of the area followed by a list of "near-term" challenges (3-5 year timeframe). In some cases, there is also a discussion of how language modeling could be used and what resources would be required for progress. The topic areas are: retrieval models; cross-lingual retrieval; Web search; user modeling; filtering, TDT, and classification; summarization; question answering; metasearch and distributed retrieval; multimedia retrieval; information extraction; and testbeds.

Before discussing this detailed list of research issues, we first consider some possible longer-term challenges for a ten-year timeframe and beyond. This type of challenge, although necessarily quite abstract, is valuable for illustrating general research directions and emphasizing the importance of collaborations.

---

[*] Advanced Research Development Agency. http://www.ic-arda.org/

2. Long-Term Challenges

In the discussions of longer-term challenges at the workshop, two main themes emerged. These were global information access and contextual retrieval.
A definition of the first challenge is:

Global information access: *Satisfy human information needs through natural, efficient inter-action with an automated system that leverages world-wide structured and unstructured data in any language.*

On the World Wide Web alone, digital information is being accumulated at a phenomenal rate and in a vast number of languages. By some estimates, the number of web pages written in Mandarin Chinese will overtake those written in English in the not too distant future. A "grand" challenge for IR will be to develop massively distributed, multi-lingual retrieval systems. Such systems would take as input an information need, encoded in any language, and return relevant results, encoded in any language. The design of such a system would probably borrow from techniques in distributed retrieval, data fusion, and cross-lingual IR, but it is unlikely that simply combining known techniques from these fields would yield such a system. What would the architecture of such a system be? What data sets could be used to develop and test such a system?

Wireless devices are being rapidly developed which can spontaneously form ad hoc peer-to-peer networks. A world in which such devices are ubiquitous gives rise to a number of challenges for IR. For example, one might fancifully imagine flying to an unknown city for a business trip, renting a car, opening one's wireless PDA, and connecting to an ad hoc peer-to-peer network in order to submit requests for information such as: (1) background on local points of interest, (2) reviews of local restaurants, and (3) "dossiers" on people who are likely to be encountered.. How would such information needs be described? (Free text, structured queries, etc.) To whom would such requests be sent? How would the results obtained be combined and presented to the user?

A definition of the second challenge is:

Contextual retrieval: *Combine search technologies and knowledge about query and user context into a single framework in order to provide the most "appropriate" answer for a user's information needs.*

In general, interactions with Web search engines could be characterized as "one size fits all". This means that all queries are treated as simple Web queries where the aim is to locate useful home pages and the burden is placed on the user to scan and navigate the retrieved material to find the answers. There is no representation of user preferences, search context, or the task context.

Despite some recent attention to this problem, little progress has been made due to the difficulty of capturing and representing knowledge about users, context, and tasks in a general Web search environment. Future search engines should be able to use context and query features to infer characteristics of the information need such as query type, answer type, and answer level, and use these characteristics in retrieval models to rank potential answers such as sentences, passages, documents, or combinations of documents.

An example of contextual retrieval would be a context aware, transmedia IR system. With such a system, if a user enters a query such as "Taj Mahal", and if the user is at a desktop computer and has spent time earlier planning a conference trip to India (reading emails on the trip; examining travel web pages; placing entries in a diary), then the system will be aware of this context and will be more inclined to retrieve pictures and videos of the Indian mausoleum, while music from the jazz band with the same name would be less likely to be retrieved. The ranked output would contain images indexed by the query text, and images similar in visual content to the retrieved text-tagged images. The high bandwidth of the machine's connection would also be taken into account when choosing images or videos to be delivered. Indeed, the bandwidth may even be a factor in the retrieval algorithm. On the other hand, if a user were starting out on a long car trip (the system being aware of this context from diaries and location sensors) and the user has often

requested audio in such a situation, then the music from the jazz band will be more likely to be retrieved. It is this context awareness, making use of knowledge about the user, the current task, the history and future plans, location and device, and the fact that retrieval can be of any kind of digital information, regardless of query type or mode, which makes this a long-term challenge.

Other types of challenges can involve contextual information. Examples are:

1. Developing intelligent classification algorithms that will be able to unobtrusively elicit user feedback, combine it with contextual and historical evidence, and produce effective structured annotation of new data.

2. Summarizing content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user, given a task and the user profile.

In addition to these two major themes, it is clear that many of the major research opportunities of the future will involve collaborations between fields. For example, IR and database collaborations will be crucial for the development of probabilistic models for integrating unstructured, semi-structured and structured data, for the design of effective distributed, heterogeneous information systems, and possibly for systems that detect new and interesting information in sensor-based streams. Collaborations with natural language processing and machine learning researchers will be needed for advanced question answering systems, and, of course, there are many exciting opportunities in collaborations with the sciences (e.g. bio-, geo-, astro-, ...).

3. Topic Discussions

3.1 Retrieval Models

Formal retrieval models have formed the basis of IR research since the early 1960's. A number of different models have been developed to describe aspects of the retrieval task: document content and structure, inter-document linkage, queries, users;

their information needs and the context in which the retrieval task is embedded. The reliance on formal retrieval models is one of the great strengths of IR research.

Information retrieval encompasses a broad range of complex information seeking tasks and current retrieval models capture only a small part of that complexity. Current models of text, for example, deal with relatively simple aspects of language (words, phrases, names) and do not capture aspects of linguistic structure that may be important to specific retrieval tasks. Similarly, current models of the user are very weak or, often, missing from current retrieval models. Current retrieval models generally focus on a narrow range of retrieval tasks (e.g., ad hoc retrieval and filtering) while ignoring other important tasks (e.g., browsing, known-item retrieval, questions answering, summarization).

Research aimed at developing more comprehensive retrieval models is critical. The goal of a single comprehensive model of retrieval is beyond our current level of understanding, but models that better describe individual tasks or a range of related tasks are possible and necessary to progress in the field.

Near-Term Challenges

1. Models that incorporate the evolving information needs of users performing realistic tasks and the utility of information as well as topicality (relevance is more than topicality).

2. More advanced models that, for example, estimate translation models or compute sub-topic models through clustering, currently require significantly more computation than is practical for deployment in many applications. Significant advances in the underlying estimation and inference algorithms will make probabilistic models much more attractive for use in large scale systems.

3. Generalization of current techniques to information sources that may be significantly different from today's document collections
4. Models and tools for incorporating multiple sources of evidence (text, queries, relevance

judgments, user context, annotations etc.).

5. Intrinsic and extrinsic evaluation techniques that measure the performance of retrieval and related technologies in the context of larger tasks, and based principled error analysis to go beyond 'what' and address 'why.'

6. Full elucidation of the relationships between the various modeling approaches.

Resource Requirements

1. More refined taxonomy of search tasks and task models.

2. Data, lots of data (documents, queries, relevance judgments, user behavior, task-related data).

Language Modeling

Retrieval models that incorporate language modeling techniques have produced promising results over the last four years. Simple language models have been shown to incorporate document and collection statistics in a more systematic way than earlier tf.idf based techniques. Language models work as well as the classical models using tf.idf, but further improvements are likely to require a broad range of techniques in addition to language modeling. The essence of the language modeling approach, which is shared with more classical probabilistic approaches to IR, is that probabilistic modeling is taken to be the primary scientific tool. At present, this appears to be the most promising framework for advancing information retrieval to meet future challenges presented by more diverse data sources and advanced retrieval tasks

3.2 Cross-Lingual Information Retrieval (CLIR)

Though initially the Web was dominated by English speakers, now less than half of existing web pages are in English. Accessing information in a host of languages is clearly important for many uses.

In monolingual retrieval, the queries are in the same language as the collection being accessed. The purpose of CLIR is to support queries in one language against a collection in other languages.

CLIR has very recently achieved one major milestone: cross-lingual document retrieval performs essentially as accurately as monolingual retrieval. This has been demonstrated in formal evaluations in the TREC in 2000, where English queries were run against a Chinese corpus, and 2001, where English queries were against an Arabic corpus and CLEF[*] 2001, with French queries against an English corpus.

Near-Term Challenges

1. *Effective user functionality.* The technology has been evaluated formally for accuracy in returning lists of documents ranked by estimated relevance to the user's need. The next steps for effective user functionality are to incorporate effective user feedback about their information need and to provide readable translations of (parts of) the retrieved documents to support document selection. Systems should also provide better support for query formulation and reformulation based on some set of intermediate search results.

2. *New, more complex applications.* We can apply the technology to the next level of cross-lingual challenges. Cross-lingual (factoid) question answering would allow posing factoid questions (e.g., asking who, what organization, when, or where) in English and receive answers in English, based on documents (in English or another language) containing the answer. Cross-lingual gisting/summarization would provide short English summaries of documents in another language.

3. *Languages with sparse data.* The technology has been developed and proved in the context of languages with many speakers, e.g., English, Spanish, French, German, Chinese, and Arabic. One challenge now is developing ways to very quickly (a few weeks) and inexpensively (a few person-weeks) find/create data for languages where resources are minimal today (for example, Farsi, Pashto, Hindi, or any of a large number of

---

[*] Cross-Language Evaluation Forum held in Europe http://www.clef-campaign.org/

5

possibilities).

4. *Massive improvements in monolingual retrieval based on learning semantic relationships from parallel and comparable corpora.* The lessons learned from CLIR suggest new ways to approach monolingual retrieval. Given the successes to date in CLIR, any improvements in monolingual retrieval should generate comparable improvements across languages and vice versa

5. *Merging retrieval result lists from databases in multiple languages.* This problem is related to the meta-search problem, but has the additional challenge that the statistics underlying the ranking functions are built on different vocabularies and thus not directly comparable.

6. *More tightly integrated models for CLIR.* Most current approaches to CLIR are based on a shallow integration between translation models and retrieval models, where independence is assumed for both components. A promising direction seems to be to explore models with less strict independence assumptions, enabling the use of context information for translation directly.

7. *More training data.* Language modeling depends on training data. Today there are several hundreds of pairs of query and relevant documents. For factoid questions, there are roughly 1000 queries with correct answers. It is generally agreed that developing those materials over the last 10 years has spurred much productive research and led to major improvements in IR. A new challenge is to develop ways to collect a hundred times more training data for language modeling—e.g., 100,000 pairs of query and relevant documents/answers rather than a few hundred.

Resource Requirements

Major contributors to progress in CLIR have been formal evaluations and data that have been made available through TREC, CLEF, and the NII-NACSIS Test Collection for IR Systems (NCCTR).

For the new challenges above, the following seem crucial:

1. An effective, inexpensive way to collect far more (100 times more) queries and results (answers or relevant documents)

2. Appropriate resources for evaluations of CLIR with low density languages

3. Testbeds (see Testbeds section) for delivering effective functionality to users

Language Modeling

In addition to the language techniques used for monolingual retrieval, CLIR has capitalized on a probabilistic translation model—e.g., a mixture model of words from general language usage, and the product of probabilities of a term coming from a document and of that term being translated to a query term. The techniques use a mix of manual bilingual dictionaries and statistical bilingual dictionaries automatically derived from corpora of translated documents. It is interesting that machine translation systems for the document language are not a prerequisite.

3.3 Web Search

Search engines for the Web are one of the most publicly visible realizations of information retrieval technology. For some search tasks (e.g., home page finding), systems such as Google provide highly accurate results. However, the Web contains more than just home pages and searchers are interested in more than just finding single pages. Significant challenges remain for improving general Web search.

Near-Term Challenges

1. *Web Structure.* The challenges are to define the actual document collection that we call the Web and to understand how the uniqueness of the Web collection structure affects retrieval methodology.

   - What is a "document" that can be retrieved and deemed as relevant to a user's information need? It could be a single Web page, a group

of related Web pages (a Web site), a path through the Web, or a portal to a hidden database (e.g., Amazon). A different view or different retrievable "documents" might imply different retrieval algorithms. For example, if we take a whole Web site as a document, then we would need to measure the relevance of the whole Web site, rather than a single page.

- What is the boundary of the Web collection? Should the collection include dynamically generated web pages? How is the global, public web different from corporate intranets and extranets? The differences may imply different approaches to IR applications on each.

- How does the hypertext structure, which distinguishes the Web from a traditional document collection, affect the evaluation methodology and retrieval models? For example, the independent relevance assumption is clearly violated due to the links; a page pointing to a relevant document can be regarded partially relevant, even if it is not relevant by itself.

2. *Crawling and Indexing.* The challenge is to develop an architecture for information access that can ensure freshness and coverage of information in a rapidly growing web.

- It is especially challenging to maintain freshness and coverage in a centralized search engine. The current approach is to have different re-visit frequencies for different types of pages/websites. There is something inherently wrong with waiting for a crawler to come around and pick up your new content before it can be "found" by people and as the web grows the issues of freshness will get worse.

- Would other alternative search architecture help maintain freshness and coverage? For example, would a distributed search architecture have an advantage here? Would topic-specific dynamic crawling be useful?

3. *Searching.* The challenge is to develop methods for exploiting all evidence, including the web structure, meta-data, user context information, to find high quality information for users.

- How can we capture and then exploit user information and query context to better understand a user's information need? What aspects of a user's context and history can we capture? Can we exploit language models to represent information need more precisely? How can we formally model the interactive retrieval process? How do we deal with variations of queries? How can we identify different types of queries and utilize the fact that specific methods work well for specific query types?

- How do we represent a web document? How do we annotate web pages ("generating new evidence") automatically or semi-automatically?

- The Web is an ideal environment for developing formal retrieval models that support multiple forms of evidence—e.g., multiple types of document representation, use of document structure, etc. The community has talked about this, but never really done it (except possibly for the inference network model). How can statistical language models be combined with other structural evidence? How can domain/linguistic/context knowledge be integrated with statistical methods?

- There is a growing "semantic Web" effort based on widespread use of controlled vocabulary metadata to improve the accuracy and consistency of Web information services. The IR community has a long history of studying controlled vocabulary metadata, especially how to assign it automatically, how to evaluate its quality, how to cope with missing or inconsistent metadata, and how to map among different vocabularies or ontologies. The IR community needs to take semantic Web efforts more seriously and contribute its expertise to controlling the hype and solving the problems surrounding this

issue.

- How can we measure "trust" of information sources and perhaps incorporate this trust factor into retrieval? How can we trust the search engine vendors not to take advantage of their powerful position in providing editorial influence on the material we access from the web?

4. *Data Collection and Evaluation.* The challenge is to develop a shared platform for gathering appropriate data and to design evaluation methods that reflect realistic characteristics of web environment.

- What can, and should, we log from Web searching in order to capture user context, how can we provide this log data as a public utility for researchers and what are the architectural implications for logging such data? What does this mean in terms of privacy issues?

- The collection is changing constantly, we cannot freeze a copy, and the linkage characteristics are changing over time, so how can we evaluate on this "moving target"?

3.4 User Modeling

Much active research in information retrieval is carried out by abstracting the user away from the problem: judgments are captured and held constant, non-binary relevance is ignored as too complex, evaluation is on a single round of query-results without opportunity to adjust the query, and so on. This abstraction has been incredibly successful in enabling research to advance rapidly, creating more effective and efficient systems for retrieving and organizing information.

However, those improvements in retrieval accuracy appear to have dwindled in the past half dozen years. It may be that one reason researchers are unable to advance beyond the current plateau is that the evaluation model forces systems toward user-generic approaches that are "good enough" for everyone, and therefore "never great" for anyone.

Additionally, some information retrieval tasks are ill-defined without taking the user into account: summarization is meaningful only in the context of a particular task and/or class of users. Other tasks, such as cross-lingual document retrieval, appear by classic (not user oriented) measures to be quite effective, but no one knows how such systems would be used by a person.

We claim that greater focus on the user will enable major advances in information retrieval technologies, perhaps dwarfing those made possible by better core algorithms. The focus may involve better modeling of the user's background and type, personalization, the context in which information access occurs, the interactive information seeking process, explanation of what happened, results presentation, or combinations of all of those.

Near-Term Challenges

1. *Shared data-gathering environment.* It is very difficult to gather information about users completely enough that it can be used in experiments. The process is expensive, labor intensive, and requires expertise that is not universal in the information retrieval community. It requires constructing a user's working environment that tightly integrates information retrieval.

   The community should support one or more groups in creating such a data-gathering laboratory, defining what types of information need to be acquired (interaction histories, annotations, explicit preferences, etc), and sharing all results. This process would be cooperative among a large number of sites. As a result, it will not be necessary for every group to gather these types of data in order to do experiments. The cost would be amortized over many experiments.

2. *Testbed of IR interactions.* One or more standard experimental data sets need to be created that are heavily annotated with information about users, context, etc. Such a research collection should allow evaluations that focus on the user in more detail, but in a way that does not require expensive user studies.

Such a dataset would allow for some degree of comparability and comparison between results from different groups.

3. *Evaluation successfully incorporating the user.* The community needs to develop an evaluation methodology that explicitly leverages information about the user and the context of the information retrieval process. Such an evaluation would have the goal of dramatically improving effectiveness for a particular user rather than providing merely satisfactory results for an unknown user.

4. *Privacy.* Support for rich modeling of the user that also preserves privacy.

5. *Extended user models.* Many, if not all of the other areas discussed in this report explicitly demand or implicitly require much more sophisticated models of users than are presently constructed for any IR purposes. There is therefore a significant need for research in how to include characteristics of such elements as users' goals, tasks, and contexts in user models, as well as how to make use of this knowledge in affecting IR system design and response.

6. *Long term user models.* As it becomes possible to gather and store more historical information about a user's interactions, we need to develop models that span tasks and applications. Such a model would be in contrast to the short-term, single-task models that are more common to date.

Language Modeling

Very little has been done that explicitly uses language models to capture the user. Arguably, past work modeling judged documents for relevance feedback is a form of language modeling, but the connection is tenuous.

One hope for language models is to represent the user by a probability distribution of interests (words and phrases), actions (information seeking and use behavior), and annotations (judgments). This information, if appropriately captured, might plausibly be used to improve the accuracy of a

retrieval system by automatic disambiguation or user-focused query expansion.

3.5 Filtering, TDT, and classification

Semi-structured text and media objects, created and managed by users, form an increasingly large proportion of online data. Information access techniques that require schemata at data-entry time are not appropriate for this data. Semantics must be imposed after the fact, in a dynamic task context, and often under the guidance of a nonprofessional user. The imposed semantics may be as simple as "show me/don't show me," as in topic tracking, message routing, and content filtering systems. Or it may be as complex as sorting documents into a hierarchical taxonomy, or producing richly linked hypertext. Systems for topic detection, terminology discovery, named entity extraction, and unsupervised learning find novel semantic regularities across media types and time-varying data, aiding users in imposing semantics.

Some of these technologies have been adopted widely, for automated metadata assignment in knowledge management systems, routing of email and phone requests for help in customer relationship management, categorization of trouble tickets to support data mining for process improvement, and many others. Despite this commercial penetration, only a tiny fraction of text data is ever classified, and most applications require setup and maintenance by computing professionals.

Near-Term Challenges

1. *User Models.* Supervised machine learning has allowed the production of classifiers from manually labeled example documents, reducing the need for rule writing by specialists. Unfortunately, the emphasis in machine learning research on a *tabula rasa* approach means that a burdensome number of documents often must be labeled. Better ways to elicit and incorporate user knowledge in machine learning algorithms, to leverage domain-specific databases, and to gather implicit feedback from user actions, are all needed to reduce the demands for labeling. Promising new approaches to leveraging unlabeled data, including pseudo-feedback, co-

training, active learning, and transduction, need to be made more predictable in their effect.

In addition to algorithmic progress, we need a better understanding of inherent properties of classification. How do, and how can, users and communities of users classify documents to support their tasks? What properties do sets of classes (whether informal shared interests or structured hierarchical taxonomies) have, and how are they connected with latent or link structures inherent in the text data itself.

2. *Semi-structured data.* The fastest growing area in text classification is in classifying real world entities by associated text: customers by their comments, broken devices by partially textual repair reports, students by their essay exams, and so on. Most such applications to date rely on a known connection between text and real world entity. Broadening the range of such applications requires systems that can detect those connections as well ("hardening" the text database), and in particular detecting textual records from multiple sources are associated with the same entity. Techniques such as link analysis and information extraction, long of interest in the intelligence community, need to be improved and extended to a wide range of text mining applications.

3. *Novelty Detection.* Detection of novel data is quickly gaining popularity as a necessary complement to real-world filtering systems. The problem has received relatively little attention, but the exploding growth of available information makes redundancy a very real obstacle to a satisfactory user experience.

Existing novelty detection systems are based primarily on pairwise comparison of potentially novel items to some form of "history," representing classes of items already known to the user. Error rates of such systems can be exactly predicted from the effectiveness of the comparison function, and this error rate rapidly increases with the size and richness of history. This makes novelty detection very challenging in large-scale filtering environments and necessitates development of conceptually new approaches to the problem.

3.6 Summarization

Text Summarization is an active field of research in both the IR and NLP communities. Summarization is important for IR since it is a means to provide access to large repositories of data in an efficient way. It shares some basic techniques with indexing, since both indexing and summarization are concerned with identifying the essence of a document. High quality summarization requires sophisticated NLP techniques in addition, which are normally not studied in IR. In particular, for domains in which the aspects of interest can be pre-specified, summarization looks very much like Information Extraction. Summarization is therefore a good challenge problem to bring together techniques from different areas.

In comparison with IR, the field of summarization suffers from the difficulty of defining a well-specified and manageable task. Since truly reusable resources like the TREC test collections did not exist for summarization, it was hard to measure progress. Importantly, the high amount of variance across human summaries complicates evaluations. Improved, more tightly specified tasks are currently being developed within the DUC program.

Near-Term Challenges

1. Define clearly specified summarization task(s) in an IR setting. Examples: headline-length summarization, topic-based summarization.

2. Move to a new genre, since producing text summaries is almost trivial for newswire and newspaper documents.

3. Move beyond extractive summarization. Extractive summaries are clearly sub-optimal with respect to obtainable compression rate and overall coherence.

4. Integrate the user's prior knowledge into models. Users do not want in summaries material they know already.

5. Specification of a clear task. Suggested tasks have included query biased summaries, news clustering summaries, browsing tools for finding relevant passages, and summarizing updates to the news.

6. Development of an evaluation measure. The BLEU measure developed for machine translation may provide inspiration. It has been successfully applied to headline-style summaries of documents.

Resource Requirements

1. Improved models that capture a user's background knowledge

2. Large data sets with example summaries (preferably outside news domain)–e.g., the creation of a SummBank, similar to the one created at the JHU summer workshop in 2001, but significantly larger. Ideally the SummBank would contain summaries of the same texts created from different perspectives, at different compressions (from 30% down to headline only), and in different styles (fluent, telegraphic, bullet points, etc.).

Language Modeling

Language Modeling has successfully been applied for content selection, compression of sentences and documents, generation of headlines and reverse-engineering the cut-and-paste process applied by human summarizers. Since it is highly unlikely that summarization can do without modeling higher order structure, integrating linguistic intuitions into probabilistic models poses a particular challenge. Also the scalability of increasingly complex models is an important issue for working systems providing on-line summarization.

3.7 Question Answering

A Question Answering (QA) system takes as input a natural language question and a source collection, and produces a targeted, contextualized natural language answer. To build the answer it gathers relevant data, summary statistics, and relations from the sources (which might be structured or unstructured), fuses or summarizes them into a single answer as appropriate, and also gathers information about the epistemic status of the answer (including its reliability, recency, factuality/hypotheticality, etc.). Building answers might also involve a dialog with the questioner to obtain clarification or otherwise refine the question and answer. Specifically, QA:

- includes questions with short, factual answers ("factoids")

- includes questions with longer natural language answers (passages and multi-passage summaries)

- includes answers that involve some data mining or gathering of summary statistics, possibly fused from multiple sources; summary statistics are operations such as maximum, mean, clustering, conjunctions and joins (thus including some aspects of what databases provide)

- excludes answers that require complex reasoning with models that require no data or information from external sources (thus not including mathematical story problems or solving college physics exam questions)

- excludes taking actions (not "make a flight reservation for me" or "what is the translation of this document", but does include "what is the earliest flight from LAX to LGA?")

Near-Term Challenges

1. Improve performance of "factoid" QA to the point that the general public would find it reliable and useful.

2. Create systems that provide richer answers (beyond factoids):
   - Answers that require summary statistics. For example, use the web pages of a million companies to answer the question "Which state in the country had the largest increase in high-tech job openings last month?" Necessary capabilities might include extracting from formatted and tabular data, mining structured data sources, creating structured data from unstructured sources, ability to select the appropriate summary statistical operations, and generating natural language answers from structured data.

   - Longer answers that have structure to them. This includes answers ranging from those with fairly standard structure (biographies, event stories, etc.) through those with some structure (causal accounts "What are the causes of the Korean war?"), pro/con advice ("should I take hormone replacement therapy?") to ones that have very little structure as in general-purpose descriptions of events and resources ("What is this movie about?") and objects ("Describe a telephone"). Solutions might involve using a language model and discourse model that depends on the question.

3. Create systems that leverage richer data sources:
   - Give answers that require integrating multiple passages, and multiple data sources.

   - Give answers that require integrating structured and unstructured data sources, including databases, semi-structured text and numerical content. This might include the ability to extract the desired capital city from a table of countries with capitals. This work could be used to build (semi)structured knowledge bases.

   - When using dynamic sources, track how an answer changes over time. This relates directly to novelty detection.

4. Create systems that provide richer interaction with the human questioner, and also provide increased transparency.

   - Provide epistemic status of answer (including is it a fact/opinion/hypothesis; at which times was it a valid answer; what is the trustworthiness of the answer and the reliability of the source; etc.) The system should provide its confidence in its own analysis and answer.

   - Support interactive QA (dialogue) to clarify and explore the topic (here there may not be only one answer but a series of answers, each leading to the next). Systems should be able to detect when no answer is possible or when the answer is not to be found in the resources at the system's disposal; thus the system should know when to reply "no answer available," "no answer possible," or "question is inconsistent". This relates to other dialogue systems.

   - Take into account user models and session models, and keep the models updated. Don't tell the questioners what they already know. Remember what you have told them and don't repeat. Give them answers in their preferred style. Handle sequences of contextually interrelated questions and responsive answers of varying complexity that need to be understood within a larger model of the user's information needs/requirements.

5. Develop better evaluation methodologies that are scientific and repeatable. Is there an "optimal" answer? What does "optimal" mean?

Language Modeling

Current statistical/language models are proven useful for at least the first part of QA - locating documents/passages with candidate answers. For the second part, the accuracy required in pinpointing exact answers (for both factoids and for longer answers that weave together factoids) demands more than current language models can support. One suggestion is to extend language models to include more structured patterns (like information extraction template patterns). (Most current work use hand-tailored special-purpose pattern matching for dates, amounts, abbreviations, etc.)

We envision a unified statistical approach in which to accurately integrate evidence, find the right answer, and emit it in the best natural language.

3.8 Metasearch and Distributed Retrieval

Metasearch is the process of retrieving and combining information from multiple sources, and it is typically studied in one of two forms: (1) data fusion, the combination of information from multiple sources that index an effectively common data set and (2) collection fusion or distributed retrieval, the combination of information from multiple sources that index effectively disjoint data sets. As more and more retrieval mechanisms become available over common data sets (e.g., the Web) and specialized data sets (e.g., medical and law libraries), the process of identifying likely sources of relevant information, retrieving information from those sources, and effectively combining the information thus gathered will only grow in importance. A future wherein ubiquitous mobile wireless devices exist, capable of forming ad hoc peer-to-peer networks and submitting and fielding requests for information, gives rise to a new host of challenges and potential rewards.

The issues typically addressed by a distributed retrieval system include resource description, resource ranking, resource selection, searching, and merging of results. Many of the techniques developed to address these issues are *ad hoc* in nature, though language modeling techniques have been successfully employed to address resource description, ranking and selection.

Classic techniques for metasearch data fusion typically address the following issues: (1) relevance score normalization to a comparable space and (2) normalized score combination to obtain a final score from which a final ranking may be obtained. Other data fusion techniques that have been developed include (1) modeling the problem as a multi-candidate election and employing rank-aggregation algorithms from Social Choice Theory (e.g., the Borda Count and Condorcet methods) and (2) various supervised learning techniques (e.g., boosting, cranking, naive Bayes, etc.).

Near-Term Challenges

1. Can a standard resource descriptor be devised such that if a resource published its descriptor (e.g., on the Web), it could participate in a generic distributed retrieval system? What data must be present in such a descriptor? A language model of the underlying data set? A semantic description of the content? A model of the required query syntax which permits interoperability? Could it be useful both for resource selection *and* metasearch data fusion?

2. The performance of language modeling techniques is, at present, on par with that of ad hoc techniques. Can a theoretically grounded model of distributed IR be developed which consistently outperforms ad hoc techniques?

3. The performance of distributed IR techniques is approaching that of a "single database," at least within research environments and when evaluated for high precision. Can this be achieved in practice? Furthermore, through the judicious use of resource selection, distributed IR should, in theory, outperform a "single database." Can this be achieved?

4. The performance of search engines varies from query to query. The goal of metasearch is often to outperform the (*a priori* unknown) best underlying search engine on a per query basis, and this can typically be achieved when combining systems of similar performance. However, this goal is often unachieved when combining search engines of widely varying levels of performance. Can a metasearch technique be developed which consistently outperforms the best underlying search engine? Or can a technique be developed which is capable of distinguishing the "good" underlying systems from "bad" on a per query basis?

5. Techniques for data fusion typically assume (and often implicitly require) that the underlying search engines index an effectively common data set; techniques for distributed IR typically assume that the underlying search engines index effectively disjoint data sets. Can techniques be developed which effectively combine the results of underlying search engines that index data sets of varying and unknown overlap? (Search

engines on the Web fall within this category.) Can the metasearch problem be modeled in a unified way such that data fusion and collection fusion are merely two extremes of a single underlying problem?

3.9 Multimedia Retrieval

Devices for creating, storing, and transmitting multimedia information are increasing in prevalence and capacity, and are decreasing in price. With little prospect of such changes slowing in the foreseeable future, it is not hard to predict with some confidence that content-based access to multimedia information (indexing, retrieval, browsing, summarization, etc.), is set to become a significant problem for web and enterprise search as well as for personal retrieval.

The problem space of this topic is large because the types of objects to be retrieved are varied ranging from collections of audio (e.g., speech, music), images (e.g., photographs, clip art, scanned documents), video (e.g., TV, camcorder, or security camera output), as well as less common objects (e.g., vector drawings, recorded pen strokes, and VRML). The methods available for indexing and retrieving such objects vary significantly depending on their type, which has a strong impact on the forms of retrieval possible; however, it is also clear that the forms of retrieval users will wish to conduct will vary for each media type.

Near-Term Challenges

The current challenge in multimedia retrieval centers on indexing: given a non-text media object, the following options are available. Text may be associated with the object (e.g. captions, juxtaposed text); part of the object might be convertable to text (e.g. through speech recognition or OCR); metadata might be assigned manually or media specific features might be extractable.

1. Extracting good indexing features from most forms of multimedia is hard (except within restricted domains). As an alternative, fragments of text or pieces of metadata may be located and used for indexing. For certain media types, text may be easy to come by, for others, however, little or no text will be directly associated with media objects. Automatic annotation may be one way of associating text with such media objects. This may involve learning from examples how text and media features are associated and then using that information for annotation.

2. The context of user activity when a media object is created may provide a good approach to indexing: for example, emails, diary entries or location-based information associated with the time that a photo was taken could be examined for pertinent text. Context at retrieval time will also be important: the location of the searcher (a particular country; a particular part of an office) or the type of device a user is using will inform the type or content of objects to be retrieved.

3. One of the major reasons for the successes of (text) IR has been the application of formal models for indexing and retrieval. Applying such models to the multimedia domain has been challenging partly because the features which are often most useful do not easily lend themselves to indexing. The extraction of appropriate features which can be used for indexing is a challenge. So is the application of formal IR models to existing features. Multimedia (image, video and audio) IR tasks need to be formulated which cannot be solved using text alone but will require advances in finding such features and advances in applying formal IR models to such tasks.

4. We also need to think of different kinds of tasks which involve data mining and retrieval of time sequences of images, video and audio from other domains. Example: A few years ago there was a paper on collecting time sequenced images of storms from radars and then retrieving similar storms from a database of time sequenced images of storms to predict the future track of the storm. This is basically a natural application of video retrieval applied to this task.

5. To deal effectively with multimedia retrieval, one must be able to handle multiple query and document modalities. In video, for example, moving images, speech, music, audio, and text (closed captions) can all contribute to effective

retrieval. Integrating the different modalities in principled ways is a challenge.

Resource Requirements

In order to assess research effectively in multimedia retrieval, task-related standardized databases on which different groups can apply their algorithms are needed. In text retrieval, it has been relatively straightforward to obtain large collections of old newspaper texts because the copyright owners do not see the raw text being of much value, however image, video, and speech libraries do see great value in their collections and consequently are much more cautious in releasing their content. While it is not a research challenge, obtaining large multimedia collections for widespread evaluation exercises is a practical and important step that needs to be addressed. We suggest that task related image and video databases with appropriate relevance judgments be included and made available to groups for research purposes as is done with TREC. Useful video collections could include news video (in multiple languages), collections of personal videos and possibly movie collections. Image collections would include image databases (maybe on specific topics) along with annotated text - the use of library image collections should also be explored.

Language modeling

The application of information retrieval and other statistical machine learning techniques, analogous to language modeling, may be useful in multimedia retrieval. Language modeling has been successful in text related areas like speech, optical character recognition and information retrieval. There is some evidence that some of these models may be useful in automatic annotation, combining image and text retrieval and image segmentation.

3.10 Information Extraction

Information extraction (IE) fills slots in an ontology or database by selecting and normalizing sub-segments of human-readable text. Examples include find names of entities and relationships between them.

Information extraction is at the heart of much of the anticipated progress in many fields. Question answering, novelty detection, cross-lingual retrieval, and summarization all hope to leverage IR to improve their effectiveness. IE is also viewed as a database-filling technique that serves as a first step toward data mining and other decision support systems.

Near-Term Challenges:

1. *Sufficiently high accuracy* of traditional entity extraction that it can be straightforwardly incorporated into systems that consume it. Current accuracy in the low 90s percent may possibly be sufficient for use in isolation, but is not sufficient when these results are combined into n-ary relations, or incorporated into question answering, novelty detection, or data mining—all of which cause errors to compound.

2. *Ability to extract literal meaning from text.* This could also be called automated proposition extraction. The University of Pennsylvania, BBN and New York University are creating a "PropBank", analogous to the "TreeBank" for parsing. We should have as a goal to enable automated "PropBank"ing of new text. This would be a major step toward automated computer "understanding" of language.

3. *Large-scale reference matching.* In many important applications there are many thousands or even millions of entity strings to be de-duplicated and matched. Performing this task efficiently and accurately requires new approaches and techniques. Also, many references (such as "police") are generic and ambiguous. New representations are needed that will facilitate top-down and knowledge-based disambiguation, or useful consumption of representations that preserve the ambiguity.

4. *Cross-lingual information extraction.* This in-cludes the ability to combine evidence from passages in multiple languages when filling a database slot. (Both a British and a Jordanian newspaper mention an event; use them both to more accurately build a complete database record of the event.) Cross-lingual IE also

includes the ability to normalize entities across multiple languages. (The entity "President Bush" is written differently in Chinese than in English, and furthermore, has multiple Chinese transliterations in common use.)

5. *Relation extraction.* What makes a database or knowledge base most interesting is not lists of isolated entities, but the relations among those entities. To make further progress in accurately extracting relations, new models are needed. Relation extraction also includes event extraction.

6. *Confidence.* Information extraction systems should provide accurate estimates of its confidence in its own analysis and answer.

7. *Robust ability to be trained from limited labeled data.* Make efficient use of unlabeled data and human interaction.

Language Modeling

Language modeling has been at the heart of many of the most successful information extraction systems. Language modeling has been applied to extraction tasks, particularly to name extraction, including hidden Markov models, maximum entropy models, and conditional random fields. Such trained systems achieve performance comparable to the best handcrafted, rule-based systems. Additionally, language modeling is being applied to relation extraction, co-reference resolution, and extraction of the literal meaning of text.

There is much evidence that language modeling also lies at the heart of future needed progress. The models will need to be more sophisticated, make more targeted approximations, and have improved parameter estimation procedures.

3.11 Testbeds

Over the previous decade, the IR research community has benefited from a set of annual US government sponsored TREC conferences that provided a level field for evaluating algorithms and systems for IR. The conferences have included exploration of several new areas such as spoken document retrieval, video retrieval, question-answering, etc.

In addition to the evaluation exercises, these conferences created a number of significant data sets that fueled further research in IR such as the use of language models for IR. The TREC events have created a set of document collections (a few million documents) with queries (a few thousand) with corresponding relevance judgments (a few million). These data sets have played a key role to promote progress in the field. However, given the significant increase of online content over the past few years (the current Web is estimated to be about ten billion pages) and of the increasing rate of using search (tens of millions of queries per day), the current TREC data sets are too small (perhaps a thousand-fold too small) to be representative of the "real world".

Hence a community-based effort is needed to create a more realistic (in scale and function) common data set to fuel further research and increase the relevance of the research activities to the commercial and government activities in IR. We outline below some of the elements needed for creating a set of common resources including data sets, possibly annotated, and testbed platforms for IR research.

The first element is a data collection effort to capture real users performing real search tasks encompassing a sufficiently large set of queries and corresponding set of retrieved documents; the acquisition of these data would allow the exploration and development of algorithms for learning retrieval models. To facilitate the data collection effort, access to a state-of-the-art (SOA) search capability is required to convince users to use this testbed for their search tasks. One approach is to use a proxy or a meta-search engine to a SOA search service and provide users with its search results. This approach would rely on instrumenting the proxy to log users' interactions over a session including all results and ancillary documents to create a complete data set of user queries and corresponding results and underlying documents; the details of what needs to be logged would be determined through the IR community's input. The scale of this data collection could easily be on the order of tens of thousands of

users each performing hundreds of queries over a period of several months.

The management of user privacy and document IP rights would require an experienced organization such as the Linguistic Data Consortium (and/or the Internet Archive) to coordinate the data collection process, in addition to the actual system building activity to create the instrumented platform to perform the data collection experiment. In addition, the LDC could play the role of a data distribution center similar to the role it currently plays in similar data collection efforts for current human language technologies research. The scope of the data collection can easily reach millions of queries with corresponding tens of millions of retrieved documents with associated logs of user activity.

A second element for a common resource would be the creation of a snapshot of a significant fraction of the web (larger than say 10%, i.e. from 200 million to 2 billion pages) to be made available for research in conjunction with data logs based on real users doing real tasks. The instrumentation of the interface will be based on the community's input.

A third element of the common resource, would be the creation of a complete testbed framework which would enable researchers to incorporate new modules into the testbed using the specified API set of the testbed and to conduct user-based research and data collection by introducing new functionality beyond the SOA search technology. The creation and maintenance of a testbed for conducting plug and play experimentation would require significant effort but could be leveraged by several research groups to advance the SOA in IR.

A fourth element would be the annotation of certain data sets for specific research projects such as the labeling of "aspect" information in a set of documents (i.e. identifying the various subtopics in a document relative to the query) retrieved by a query. The identification of an efficient approach to the annotation task would be a challenge for the specific project that undertakes the task. The development of annotation standards would increase the sharing of effort across many teams.

The above-shared common resources can be developed using a parallel approach in developing the various elements with an appropriate mechanism for integration at a set of planned timelines to benefit from the shared work. To enhance the chance of creating the shared resources in a timely fashion, the creation and funding of specific projects to satisfy the various elements would be needed. NSF via its ITR program may be a source of funding to help create the Common Testbed for IR; in addition, other agencies such as ARDA or DARPA may contribute to the creation of this testbed. It is anticipated that such an approach is required in the IR field to enable the next generation of research to be relevant to the new challenges and opportunities created by the explosion of the Web.

4. Conclusion

During the last decade, IR has been changed from a field for information specialists to a field for everyone. The transition to electronic publishing and dissemination (the Web) when combined with large-scale IR has truly put the world at our fingertips. However, information specialists and ordinary citizens alike are beginning to drown in information. The next generation of IR tools must provide dramatically better capabilities to learn, anticipate, and assist with the daily information gathering, analysis, and management needs of individuals and groups over time-spans measured in decades. These requirements require that the IR field rethink its basic assumptions and evaluation methodologies, because the approaches and experimental resources that brought the field to its current level of success will not be sufficient to reach the next level.