State-of-the-Art Tools for More Efficient Information Discovery and Analysis

Dr. Mark T. Maybury
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
Tel: (781) 271-7230 Fax: (781) 271-2780
http://itc.mitre.org

ABSTRACT
Information discovery and analysis can be enabled by a wide-range of technologies. Typically analysts have to perform data searches against very large heterogeneous repositories, extract information from result sets, summarize and interpret the results, and form conclusions based on the results. In intelligent information access, tools can facilitate these activities for the analyst throughout the process, decreasing task time and increasing comprehensiveness and accuracy of search if tools are appropriately chosen and applied. The purpose of this paper is to provide an overview of five intelligent information access technologies: information retrieval, summarization, information extraction, text clustering, and question answering. We aim to provide a brief characterization of state of the art in each area, point to some example tools, and indicate the potential benefit of each of these areas to end users.

Keywords
Information access, information retrieval, document summarization, information extraction, text clustering, and question answering

INFORMATION RETRIEVAL

The state of the art in information retrieval is ranked lists of documents obtained from keyword queries. Current approaches emphasize speed, scalability, domain independence and robustness, all of which are critical for access to large collections of documents. Information retrieval methods such as document indexing and query processing have helped drive valuable solutions for rapid access to large scale collections (e.g., the web). Today, systems can return documents from many natural languages relevant to a particular subject with around 80% precision but low recall (or vice versa). Recall is the ability to retrieve *all* of the relevant results whereas precision is the ability to retrieve *only* relevant results. So, for example, in a high precision but low recall system, if a user poses a query and gets back documents, 4 out of every 5 documents will be relevant. Yet while most of those documents returned will be relevant, the user will also miss many relevant documents.

Since 1992, NIST has organized the annual Text Retrieval Conference (TREC, trec.nist.gov) to benchmark retrieval system performance for hundreds of international participants from industry, academia, and government. An important scientific advance in the past few years is that advances in automated query expansion and relevancy feedback have achieved near human retrieval performance. Figure 1 illustrates the classical precision versus recall tradeoff for a high performing IR system in the TREC-9 evaluation.

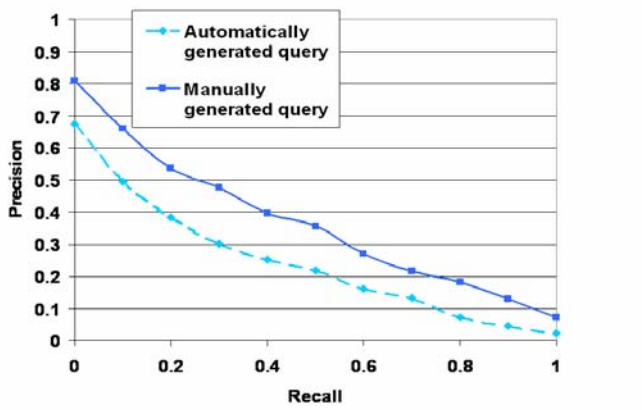

Figure 1. Typical TREC-9 Results for a System

Search engines typically exploit statistics (e.g., term-frequency/inverse document-frequency (TF/IDF), word co-occurrence, document or web structure, and format to enhance retrieval. Traditional retrieval algorithms represent documents as vectors of term weights (word features) from a set of terms (dictionary). Queries are similarly analyzed and matched to the most similar document vector(s). Because irrelevant and redundant features degrade algorithm speed and accuracy, developers work to reduce dimensionality. Other algorithms use extra-document information to enhance retrieval. For example, Google uses link analysis, e.g., is a page an authority (number of other pages that point to it) or a hub (number of pages it points to) to help determine its page rank. Related, Teoma assesses subject specific link popularity to enhance retrieval. Other tricks include using format to help modify relevancy (e.g., words that appear in bold, larger font, higher in the document get more weight.).

It is useful to contrast search engines, directories, metacrawlers, and content providers. Metacrawlers (e.g., metacrawler.com, dogpile.com) aggregate results from multiple subordinate search engines. Services such as Yahoo or Lycos provide taxonomic browsing of the information space. Related to we will discuss further below, some tools create or use taxonomies to support information access. For example, the search engine Northernlight provides taxonomic results navigation, that is it groups results in custom search folders™ by subject (e.g., baseball, camping, expert systems, desserts), type (e.g., press releases, product reviews, resumes, recipes), source (e.g. web pages, magazines, encyclopedias, DB), and language (e.g., English, German, French). Vivisimo similarly takes returns from a keyword search and presents these as clustered results in hierarchically expandable/collapsible folders of results.

Reaching beyond text, systems are emerging that provide content based retrieval of speech, imagery, and video (Maybury 1997). Also, it is notable that the semantic web promises to enhance retrieval through more accurate content-based markup of materials.

SUMMARIZATION

Summarization is a technology process that distills the most important information from a source (or sources), and produces an abridged version of the information as either an abstract or an extract. By *abstract* we mean a summary at least some of whose material is not present in the input (e.g., subject categories, paraphrase of content, etc.). In contrast, an *extract* is a summary consisting entirely of material copied from the input. We can also distinguish three kinds of summaries. *Indicative* summaries characterize the "aboutness" of a source for use in assessing the relevancy of content for selecting documents for more in-depth reading or processing. In contrast, *informative* summaries attempt to cover all the salient information in the source at some level of detail. Finally, *evaluative* summaries are critical and aim to evaluate the subject matter of the source, expressing the abstractor's views on the quality of the work of the author. Summaries can be generic or can be tailored to particular purposes or users.

Summarization software exploits a broad range of methods to include cue phrases (e.g., "in summary", "in conclusion", "important") (Luhn 1958), location/format heuristics (e.g., title, first sentence), frequency analysis (words, phrases), statistical combinations of features (Kupiec et al. 1995), language processing (e.g., syntax and semantic analysis), and discourse/rhetorical analysis (e.g., Marcu 1997).

With newspaper text, analyst can summarize documents to 20% of their source size without information loss, saving themselves 50% of task time (Mani and Maybury, 1999).

There have been a few summarization evaluations such as the TIPSTER SUMMAC Text Summarization Evaluation Conference (Mani et al. 1998), the Japanese Text Summarization Challenge (Fukusima and Okumura 2001), and the Document Understanding Conference summarization evaluation (http://duc.nist.gov). SUMMAC was the first large-scale, developer-independent evaluation of text summarizers and considered three tasks: an ad hoc task in which indicative summaries tailored to a particular topic, a categorization task in which a generic summary to be used to categorize a document, and finally a question-answering task in which a topic-related summary for a document was evaluated in terms of its "informativeness" in terms of containing answers to topic related questions. Automatic summaries proved to be effective for relevancy assessment tasks. A key result was that summaries at rather low compression rates (17% for adhoc, 10% for categorization) supported relevancy assessments as good as full text while cutting judgement time as much as by half (50% for adhoc, 40% for categorization).

Figure 2 illustrates the nature of the performance of automated summarization systems. The graph displays is how many questions a human could answer given a summary at varying levels of compression of the original text (e.g., 10%, 20%, etc compression). A number of industry (e.., GE, SRA, Textwise), academic (e.g., ISI, NMSU, University of Pennsylvania), and industry/academia teams (CGI/CMU, Cornell/SabIR) competed in this SUMMAC evaluation. Model human-generated summaries (Modsumm) that had all the answers to the test questions (hence had perfect recall) are shown as a yellow squares in the graph. While systems utilized a range of summarization methods such as those referenced above, as expected, the highest answer recall is associated with the least reduction of the source. The ratio of accuracy to compression, called the informativeness ratio, is around 1.5.

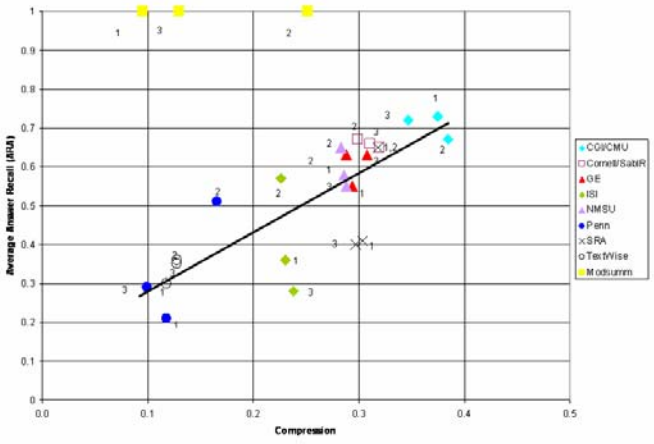

Figure 2. SUMMAC Q&A Evaluation

In addition to the success of answering questions using a summary, another important measure is how much time

users could save using summaries as opposed to full original sources. In the SUMMAC evaluation, using summaries that were just over 20% of the full source text would roughly halve a user's decision time.

Another finding from the SUMMAC evaluations was that content-based automatic scoring (vocabulary overlap) correlates very well with human scoring (passage/answer recall). At each compression level, systems outperformed baseline approaches (e.g., taking the lead sentence of a document or using term frequency to extract summary sentences) in terms of in content overlap with human summaries. Also, human subjective grading of coherence and informativeness showed that human abstracts were better than human extracts which were better than both automated systems and baselines.

Finally, we emphasize that the evaluation show in Figure 2 was on English texts. A similar evaluation of multiple systems was performed on Japanese texts (Fukusima and Okumura 2001) with consistent results, suggesting that that observed effects could be language independent.

INFORMATION EXTRACTION

While summarization can help users sift through large volumes of text to get at key text segments, information extraction (IE) promises more direct access to relevant content. Information extraction is the automated identification of specific semantic elements within a text such as the entities (e.g., people, organizations, locations), properties or attributes such as characteristics of entities, relations (among entities), and/or events. Current systems are able to extract named entities in news with over 90% accuracy and relations among entities with 70-80% accuracy.

For example Figure 3 illustrates a document that an analyst has retrieved on a UN resolution on Iran in which IE software has annotated and color coded entities such as people (Ali Kohrram, Mohammend ElBaradei), locations (Iran, Islamic Republic), organizations (IAEA, United Nations, Mehr News Agency, UN Human Rights and Disarmament Commission), and dates (Sunday, September). While the system has done an excellent job, notice that it is not perfect. For example, it annotates IAEA as a person (green).

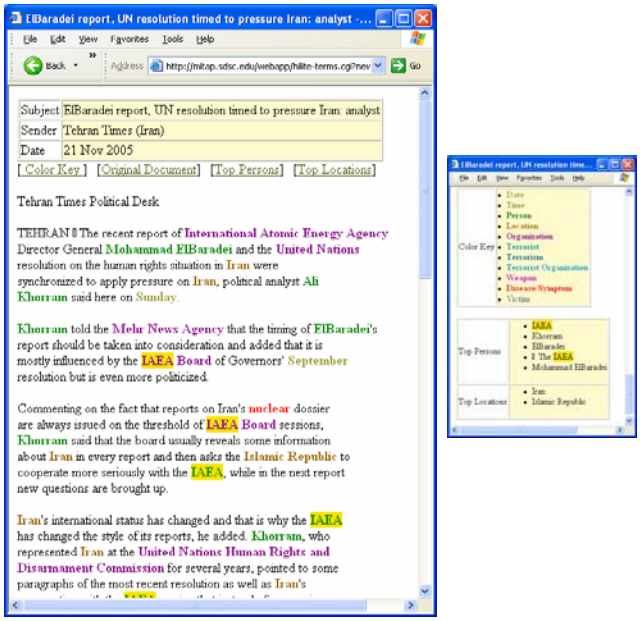

Figure 3. Entity Extraction

The DARPA led but multiagency TIPSTER program started in 1991 and fueled investment in IE. Figure 4 illustrates the performance of the best IE systems for different tasks and languages over the years. IE systems are evaluated both in terms of the detection of the phrase that names an entity as well as the classification of the entity correctly (that is, e.g., distinguishing a person from an organization). The two primary metrics used for text in NLP evaluations are precision and recall for each entity class, where:

- Precision = #CorrectReturned / #TotalReturned
- Recall = #CorrectReturned / #CorrectPossible

A harmonic mean of precision and recall is used as a "balanced" single measure and is called an F-measure.

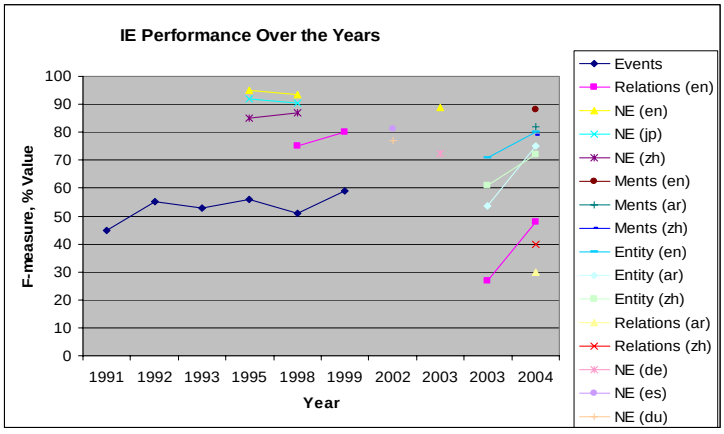

Figure 4. Information Extraction Performance

Figure 4 illustrates that the best entity extraction in English is about an 85-95% F-score. In contrast, relation extraction

among entities (e.g., X is-located-at Y, A is-the-father-of B) drops to about 80%. Finally, event extraction is less than 60% performance, and is improving slowly. Finally when moving from a completely English corpora to a foreign language, extraction performance drop to between ¼ and 4/5ths. Notice in Figure 4 the range of languages being evaluated (e.g., English, Japanese, Arabic, Chinese, German, Spanish, Dutch).

Finally, recent evaluations on bioinformatics have discovered that IE performs better with newswire than on biology texts (90% vs. 80% f-score), but also that newswire is easier for human annotators too.

TEXT CLUSTERING

Text clustering is the process of detecting topics within a document collection, creating a taxonomy of these topics, assigning documents to the topics, and then labeling these topic clusters so they can more easily be used by various tools. There are a number of commercial text clustering tools available, some on line. Some of these tools perform categorization or clustering of post retrieval results sets (e.g., northernlight.com, vivismo.com). Others are intended to create taxonomic classifications to support subsequent browsing or retrieval (e.g., www.semio.com). Still others may provide visualization and interactive exploration with search results such as link node diagrams (e.g., www.kartoo.com).

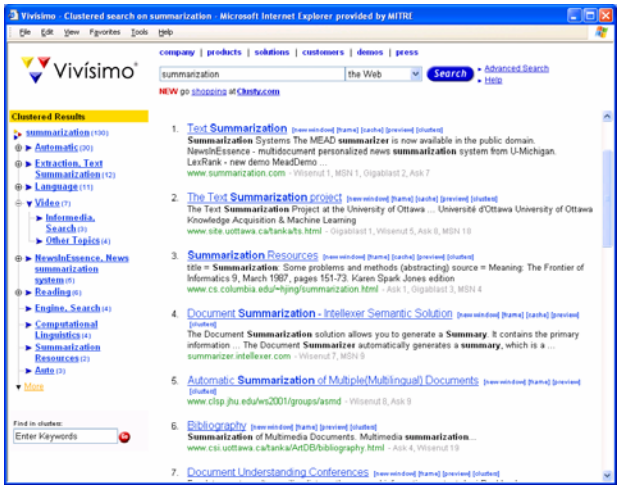

Figure 5. Post Retrieval Document Clustering

Figure 5, for example, illustrates the results from vivismo.com for a search on "summarization" in which the tool has not only listed the most relevant web pages in rank order, but to the left notice that it has automatically clustered and labelled groups of relevant web pages such as those dealing with different aspects of the query such as "automatic" summarization,, "video" summarization or "news" summarization. Note we have further expanded the "video" cluster to find out there is a subcluster of documents (web pages) about CMU's "Infomedia" video summarization system. In this way, the user can explore the constellation of document clusters to get an overview of a collection before diving into specific documents.

Figure 6 illustrates a semi-automatic development process for developing clustering systems that uses existing taxonomies or gazetteers and specialized dictionaries to extract phrases and cluster concepts to enable the creation of taxonomies to which clusters of concepts (and their associated documents) can be attached to enable browsing. For example, using a geospatial gazetteer a user could browse a collection of documents organized by the locations therein or using taxonomies generated by the process show in Figure 6.

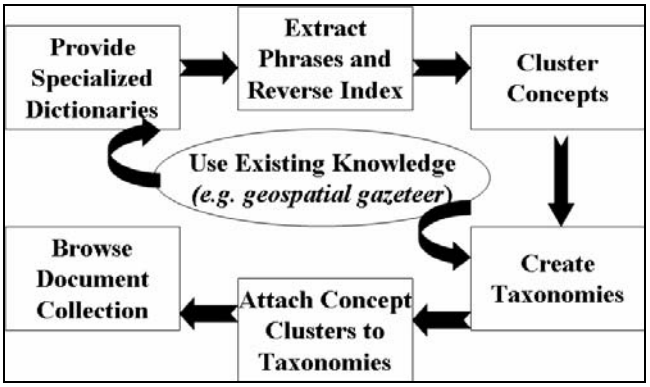

Figure 6. Text Clustering

Cluster evaluation can be intrinsic, for example, measurements of intra and inter cluster similarity. Extrinsic evaluations are based on a particular task such as comparing to a manual classification, such as the degree of precision and recall in terms of correctly creating the same clusters and/or measuring how specialized in a hierarchy of clusters documents are.

TOPIC DETECTION AND TRACKING

Just as a user might want to create a set of clusters from a given collection of texts, they also might want to identify and track topics or stories within and across documents. Topic Detection and Tracking (TDT) research was pursued under the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program (Wayne 2000). TDT aimed to thread together topically related material from newswire and broadcast news in both English and Mandarin Chinese. Five TDT tasks in the Program were:

1. *Story Segmentation* - Detect changes between topically cohesive sections
2. *Topic Tracking* - Keep track of stories similar to a set of example stories
3. *Topic Detection* - Build clusters of stories that dis-

cuss the same topic

4. First Story Detection - Detect if a story is the first story of a new, unknown topic

5. Link Detection - Detect whether or not two stories are topically linked

Figure 7 illustrates 2004 TDT Performance for the topic tracking task measured using detection error tradeoff (DET) curves which display the tradeoff between miss and false alarm probability. This allows easy comparison of different algorithms or under different tradeoff conditions. An ideal system would have a very low probability of missing relevant information and a very low probability of detecting a document as relevant when it was not (false alarm), the bottom left hand corner of the graph in Figure 7. Figure 7 contrasts performance of five systems. Random performance is show as the uppermost straight line in the upper right quadrant. What this curve says is that the very best system from CMU will miss 40% of the relevant topics with essentially no false alarms or at the other extreme will not miss any of the relevant topics if you deal with about a one percent false alarm rate. Because of variance on both the number of on-topics stories and topic difficulty, an average of performance across topics is used to improve the reliability of the performance measures.

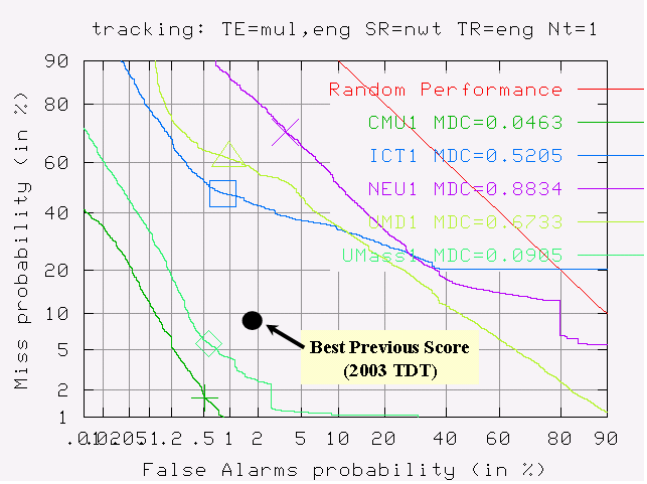

Figure 7. Topic Tracking Performance on Newswire and Multilingual Texts (Fiscus and Wheatley 2004)

NIST calculates a *normalized cost*, ranging from 0 (best) to 1 or more from the miss and false alarm rates for a task and their predetermined costs to reflect the overall strength of an algorithm. Normalized costs are shown in the upper right corner of Figure 7 and reveal that the systems performed in the following rank order CMU (.0463) > UMass (.0985) > ICT (.5205) > UMD (.6733) > NEU (.8834). The best score from the previous year was between the UMass and ICT systems. As in many other government funded evaluations, TDT corpora (English and Chinese news documents) are made available via the Linguistic Data Consortium (LDC) which enables more rapid startup and amortizes evaluation costs across many projects. Finally, Figure 8 illustrates the normalized performance of the best TDT systems over five years on the task of topic detection, link detection, and topic tracking.

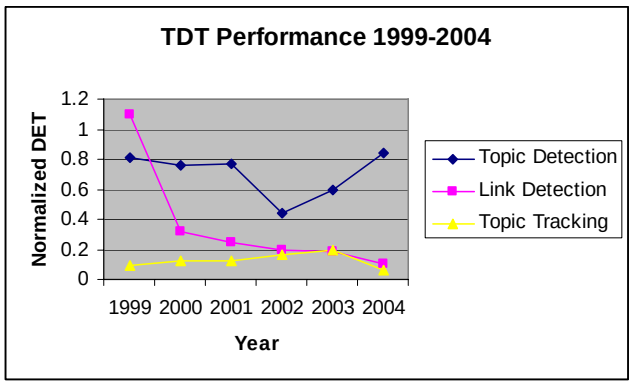

Figure 8. Topic Tracking Performance over time

QUESTION ANSWERING

Question answering uses several of the previously discussed technologies. As illustrated in Figure 9, in question answering, questions are analyzed by the system, documents are retrieved using some representation of the question, answers are extracted from these documents, and a ranked set of possible answers is provided to the user.

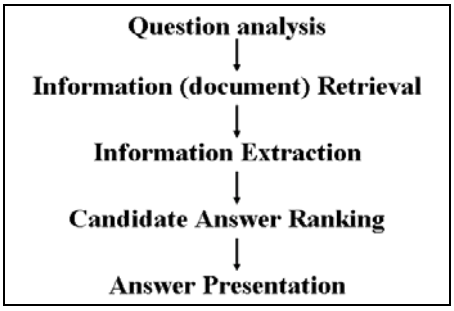

Figure 9. Question Answering

Using the best performing question answering system, an analyst can retrieve answers to simple factual questions from relevant documents at 75% accuracy (Maybury. 2004). Figure 10 illustrates the performance of systems in the TREC 2003 question answering track. Several evaluations tasks included answering questions about definitions, lists, and factoids (factoids and definitions came from actual questions found in MSNSearch and AOL logs). A human "assessor" judges question answer pair correctness which the systems must find from 3 gigabytes of newswire text (1 million articles). Figure 10 illustrates the average scores of the top 15 systems using the mean reciprocal rank of correct answers among the top 5 answers. Figure 10 illustrates that the top system provided 70% of the correct "answers" in the top 5 50-byte passages it returned as answers. On line question answering systems include Ask-

Jeeves (www.ask.com) and Language Computer Corp (www.languagecomputer.com).

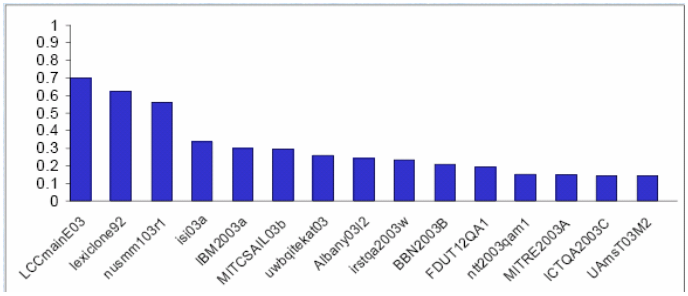

Figure 10. TREC 2003 Question Answering Performance

MACHINE TRANSLATION

Translation is the process of converting a source text or texts in one language into another text in a target language. Translation can either be machine translation, that is translation by computer code, or it could be computer assisted translation in which the machine helps the human translator. Translation memory uses a store of previously translated source texts and their equivalent target texts in a database and retrieves related segments during the translation of new texts.

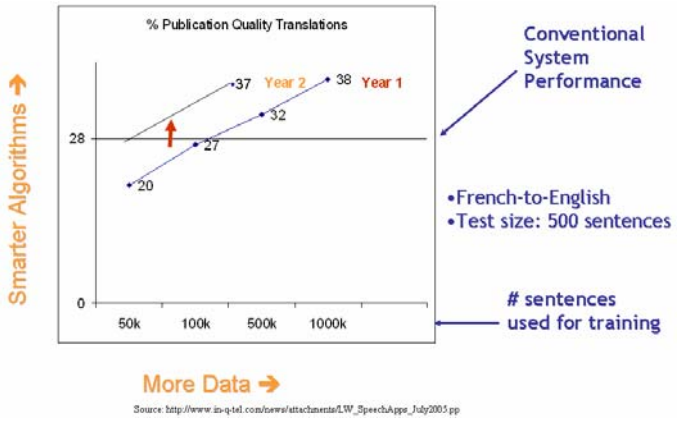

Figure 11. MT Performance

Over several decades of active development in the R&D community, increasingly high quality translations have arisen from successful methods such as rule-based methods (lexical, grammatical, semantic), Statistical machine translation, and example based statistical machine translation are successful paradigms . It is possible now to access gist quality translations from the web (e.g., Systran). As Figure 11 illustrates, applying machine learning to parallel corpora and translation memories enables rapid development and customization (e.g., Language Weaver, which has over 6000 languages, 150 or so important ones, and an increasing but still limited global coverage). Figure 11 illustrates that more data you provide to machine translation (e.g., 50 to 100k) the resultant higher quality.

SUMMARY

Effectively exploited, intelligent information access systems promise many benefits. These include:

- More *strategic* management of intellectual resources – unlocking the full enterprise potential.
- More *efficient* knowledge discovery -- enabling more rapid knowledge discovery with less work.
- More *effective* knowledge application -- tailoring information access to individual needs.

This paper provides an overview of five intelligent information access technologies: information retrieval, summarization, information extraction, text clustering, and question answering. We reported performance measures on each of these key tasks and found the following key findings:

- automated systems exist that can return documents relevant to a particular subject with around 80% precision but low recall
- automated query and relevance feedback is near human performance
- systems can presently identify entities at over 90% accuracy and relations among entities at 70-80% accuracy.
- Automatic real-time translation of source documents to gist quality targets in many languages
- systems can summarize documents to 20% of their source size without information loss, saving users 50% of task time
- systems can also respond to a simple factual question by returning answers from relevant documents at 75% accuracy.

We conclude noting that deployed systems should be useful, usable, user customizable, and open.

REFERENCES

1. Advanced Question Answering for Intelligence (AQUAINT) www.ic-arda.org/InfoExploit/aquaint.
2. Fiscus, J. and Wheatley, B. 2004. Overview of the TDT 2004 Evaluation and Results. TDT Workshop. Dec 2-3, 2004.
3. Fukusima, T. and Okumura, M. 2001. "Text Summarization Challenge: Text summarization evaluation in Japan." Workshop on Automatic Summarization. Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2001). New Brunswick, New Jersey: Association for Computational Linguistics.
4. Google Text Mining References:

http://directory.google.com/Top/Reference/Knowledge_Management/Knowledge_Discovery/Text_Mining/

5. Kupiec, J., Pedersen, J. Chen, F. 1995. A Trainable Document Summarizer, In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995, 68-73.
6. Light, M. and Maybury, M. 2002. Personalized Multimedia Information Access: Ask Questions, Get Personalized Answers. *Communications of the ACM* 45(5): 54-59. (www.acm.org/cacm/0502/0502toc.html).
7. Luhn, H. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal*, pages 159-165, 1958.
8. Mani, I. and Maybury, M., editors, 1999. *Advances in Automatic Text Summarization*. MIT Press
9. Marcu, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 96-103.
10. Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org:80/Press/Books/Maybury-2/)
11. Maybury, M. T. editor. 2004. *New Directions in Question Answering*. AAAI/MIT Pres.
12. Mani, I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T. Chrzanowski, M. and Sundheim, B. 1998. The TIPSTER SUMMAC Text Summarization Evaluation: Final Report. MITRE Technical Report 98W-138. http://www-nlpir.nist.gov/related_projects/tipster_summac/
13. Search engine watch (http://searchenginewatch.com)
14. TIPSTER program. 1991-1996. http://www.itl.nist.gov/iaui/894.02/related_projects/tipster
15. Wayne. C. 2000. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation LREC. 2000.