

# **SOME CONSIDERATIONS ON TECHNOLOGIES INVOLVING NATURAL LANGUAGE FOR INFORMATION ACCESS AND USAGE**

Renato De Mori  
LIA-CNRS, France  
McGill University, Canada  
renato.demori@lia.univ-avignon.fr

## **INTRODUCTION**

A very large quantity of documents used by Governments and Public Administration is written in natural language and, often, using different languages. Furthermore, interaction with computerized information services takes place with human-computer interfaces. Accessibility would be facilitated if these interfaces are designed to use, among other media, natural language.

Furthermore, research on the use of natural language has received a great attention in the last decade for application in document classification and retrieval, document translation, human-computer interaction, multimedia classification and summarization.

It appears that human-computer communication and data entry is an area of high strategic importance, even if data classification, search and retrieval are even more important.

A concise analysis will be made for these sectors of research after a brief recall of technical aspects. Based on it, some considerations will be made on research policy and funding strategies. These considerations probably apply to a wider horizon.

## **CLASSIFICATION AND RETRIEVAL OF DOCUMENTS IS A DIFFICULT TASK**

Information Retrieval (IR) systems accept a query from a user and respond to a query by proposing documents stored in a data-base that may correspond to the request formulated in the query.

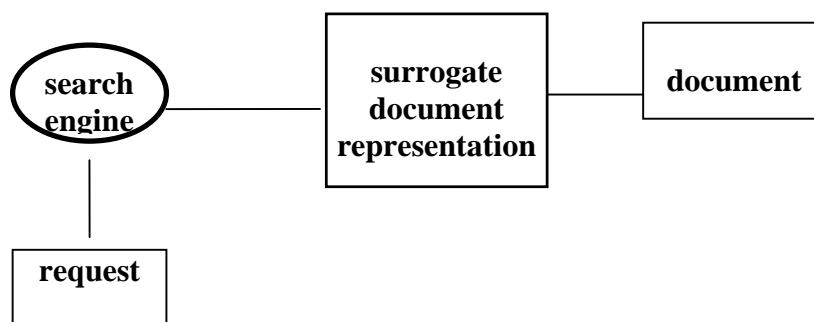


Figure 1 - The Information Retrieval process

This process, represented in figure 1, differs from classical data-base access because the queries are formulated by humans who may not know any formal query language and documents are represented by surrogate structures which are often unknown to the users who formulate the queries. Matching between queries and documents is often imperfect and is performed by a search engine which may return a ranked list of document candidates.

Introductions to problems and algorithms can be found, for example, in (Salton and Mc Gill, 1983).

Classical Information Retrieval (IR) systems represent a query  $Q$  in terms of elements of a keyword set and compare this with the content of an archive in search of some form of match between the query and representations of some documents. As a result of the match, a set of documents is proposed to the user. Matching can be precise (usually based on finite-state automata recognizing a query expression) or imprecise (based on a comparison of a vector of numbers derived from statistics of words in a query with a vector of numbers derived from statistics of document representations).

When documents are stored into the archive, two main processes are performed. The first process consists in obtaining a surrogate representation, while the second process consists in assigning an identifier (ID) to a document in such a way that it can be stored into and retrieved from a data-base. In the case of text, obtaining a surrogate information involves segmenting a text into words, ignoring the non key-words using a *stop-list* of words considered to be irrelevant for retrieval, computing the *stem* for each word and weighting terms in the document based on their stem statistics. The stem of a word is a prefix obtained by an algorithm that selects the shortest prefix of a word that contains enough letters for distinguishing that word in the set of key-words.

Stemming is also performed on the query before attempting to match a query representation with document surrogates. A decision process performs a selection on the candidates of the matching process. Selected documents are ranked and used to enrich the history of queries and satisfactory answers. These histories may be used in the future to enrich the representation of a new query that has affinity with one of the histories. Keywords of the satisfactory answers to similar queries can be used to enrich the representation of the actual request in terms of word statistics. This process, controlled by the user, who expresses satisfaction to an answer, is called *relevance feed-back*.

There are some major problems in IR, namely:

- document retrieval is very often a search for concepts,
- semantic interpretation is an open problem,
- there are many ways to express the same concept,
- words are ambiguous,
- a document that is pertinent for a query may not contain the same word as the query for indicating the same concept.

Among other things, it is important to investigate the cases of:

- *homonymy* (the same word may represent two different unrelated concepts, e.g. the *bark* of a dog vs. the *bark* of a tree)
- *polysemy* (the same word may represent two different, but related senses, e.g. the *review* of a paper or to *review* an activity).

Manually conducted experiments have shown that:

- word meanings are highly correlated with relevance judgments,
- there is a high degree of lexical ambiguity even in a limited domain corpus,
- morphological variants of a word grouped by a *stemmer* improves IR efficiency,

- grouping inflectional variants (plural and tensed verbs) or grouping derivational variants (e.g., -ize -ity) degrades IR performance,
- only *related* morphological variants should be grouped,
- using the results of Part Of Speech analysis has in general a negative impact; it may have a positive impact to identify *meaningful phrases* and to filter candidates (e.g. nouns) for *word-sense disambiguation*,
- some short *phrases* should be considered as a unit.

Problems and observations support the conclusion that it is difficult to cast the experimental findings and conjectures into precise algorithms. Moreover, once some algorithms are conceived and implemented, system based on them have to be evaluated. IR systems are evaluated on the time and memory complexities of the associated algorithms and data structures. Retrieval performance is measured by the following two parameters:

*recall r*: the number of documents in the correct category assigned to that category over the total number of documents in the correct category;

*precision p*: number of documents correctly assigned to a category over the total number of documents (correctly and incorrectly) assigned to the same category.

Controlled experiments, carried under the supervision of the National Institute of Standards and Technology (NIST) of the USA have shown some of the limits of these techniques. Recent progress and results of test campaigns can be found in [1].

If documents are not available in digital form, then they can be acquired using Optical Character Recognition or automatic dictation. Both types of process are error prone.

Once documents are available in digital form, then it is important to address problems of distributed archive organization, digital libraries, discovering where the information is and what are the links between documents at different locations. The problem is represented in Figure 2. Search and data mining methods are relevant for this purpose and the available solutions are far from being perfect. Problems are even much harder if documents are multimedia.

Most methods rely on language and other mathematical models. A key problem is about the value of models for realities which may not follow stable laws which can be formulated in mathematical form.

## NATURAL LANGUAGE INTERFACES

Person-machine Communication (PMC) can be seen as an exchange of information coded in a way suitable for transmission through a physical medium. *Coding* is the process of producing a representation of what has to be communicated. The content to be communicated is structured using words represented by sequences of symbols of an alphabet and belonging to a given lexicon. Phrases are made by concatenating words according to the rules of a grammar and associated in order to be consistent with a given semantics. These various types of constraints are *knowledge sources (KS)* with which a symbolic version of the message to be exchanged is built. The symbolic version undergoes further transformations that make it transmittable through a physical channel.

Dictation and interpretation systems perform a *decoding process* using *KSs* to transform the message carried by a speech signal into different levels of symbolic representation. Decoding can produce word sequences or conceptual hypotheses.

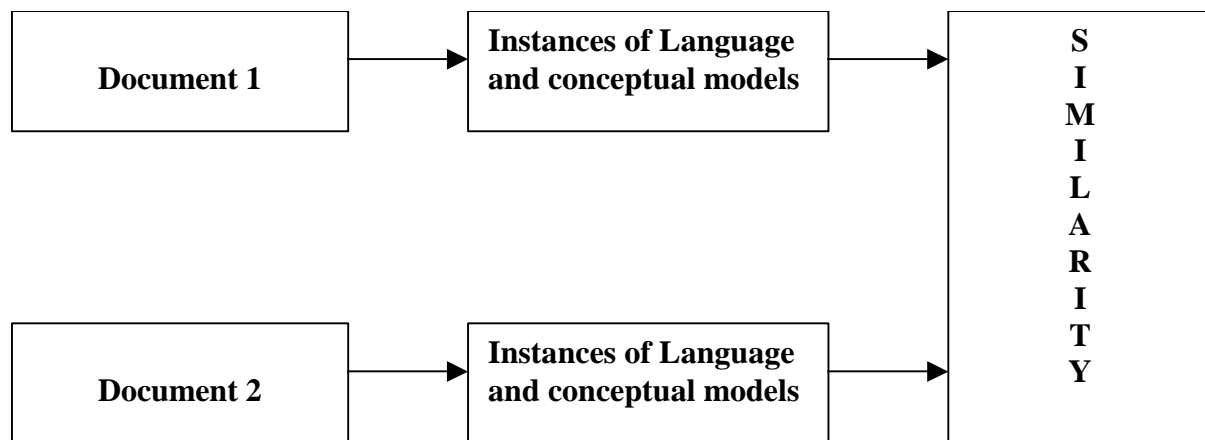


Figure 2 – Detecting relations between documents

Unlike person-to-person communication, PMC is expected to produce instances of computer data-structures in a deterministic way. Deterministic here means that a computer system has to produce the same representation for the same signal, every time this signal is processed. So, with current technology, speech interpretation by machine is not “creative” in the sense that it is performed by predictable reactions to the data. The *KSs* used by machines in the decoding process are only *models* of the ones used by humans for producing their messages.

One of these *KSs* is the *Language Model (LM)*. An LM is a collection of constraints on the sequence of words acceptable in a given language. These constraints can be represented by rules of a *generative grammar G*. *G* can be used to produce sentences of a language  $LG(G)$ . *G* is defined as a 4-tuple :  $G = (\sigma, V_T, V_N, P)$ , where  $V_T$  is a set (an alphabet, in the case of Natural Language, a lexicon) of all the words of  $LG(G)$ ,  $V_N$  is a set of non-terminal symbols representing abstractions of language components, for example, syntactic categories.  $\sigma \in V_N$  indicates the abstract category of all the sentences in  $LG(G)$ . *P* is a set of rewriting generative rules of the type  $\alpha \rightarrow \beta$  where  $\alpha$  is a sequence of symbols that should contain at least one in  $V_N$ ,  $\beta \in (V_T \cup V_N)^*$  is a string of symbols in  $V_T$  or  $V_N$  with which, starting with a rule of the type  $\sigma \rightarrow \beta$  and further rewriting the components of  $\beta$ , it is possible to generate a sentence in  $LG(G)$ . If  $\alpha$  can be only one symbol in  $V_N$ , then *G* and  $LG(G)$  have the property of being *context-free*.

An important difficulty in Natural Language (NL) analysis is that it is almost impossible to conceive a grammar *G* capable of generating all and only the sentences of a natural language. This is due to many factors, probably the most important one being that NL evolves in a way difficult to characterize with formal models. Nevertheless, with grammars it is possible to build very useful, but approximate LMs. Grammars with a large number of detailed rules can accurately model certain NL aspects but be too limited for other aspects. These grammars are said to have limited *coverage*. Other grammars can have a complete coverage but, being too general, they can generate sentences that do not belong to an NL. A good example of these *overgenerating* grammars is one that can generate every pair of words

in a NL vocabulary (*word pair grammar*). Overgeneration can be mitigated by associating probabilities to grammar rules in such a way that undesired sentences will be generated with lower probability than legal sentences in a given NL. Some of these grammars are particularly useful for ASR because they can be represented by *Stochastic Finite State Automata (SFSA)* in which states correspond to symbols in  $V_N$  and arcs are labeled with words in  $V_T$ . Probabilities are associated to arcs. For example, *bigram* probabilities can be associated to word pairs in a stochastic word pair grammar.

Arcs in these SFSA can be replaced by other (possibly stochastic) automata, one for each word representing alternate pronunciations of each word. Arcs of these word automata are labeled with phonemes and pronunciations are obtained with a *Lexical Model (LeM)*. In turn, each model can be replaced by a corresponding *Acoustic Model (AM)* relating each phoneme to distributions of acoustic parameters or features that can be observed when that phoneme is uttered.

In this way, an *Integrated Network (IN)* can be obtained and effectively used to generate word or interpretation hypotheses about a given speech signal. The decoding process has to deal with *ambiguity* due to distortions introduced by the transmission channel, the limits of the knowledge used, and often to intrinsic imprecision of the spoken message. The imprecision is due to the fact that the speaker's intention may not be that of producing exactly an instance of the data structures belonging to the knowledge of the decoder.

To a certain extent, ambiguities can be reduced by exploiting message *redundancy*. In practice, knowledge is used to transform the input signal into more suitable sequences of vectors of parameters and to obtain from them various levels of symbolic representations. The first level of symbolic representation can be a word, a syllable, a phoneme or simply an acoustic descriptor.

Interpretation is usually obtained by a *search* process that considers an IN to be the generator of an observable description

$$X = x_1 x_2 \dots x_n \dots x_N$$

of the signal to be interpreted. The search process attempts to find the best sequence of IN states identifying a path through which  $X$  is generated. Competing sequence candidates are ranked based on *scoring methods*. Scores are used by search strategies for progressively growing partial IN paths. These candidates are often called *theories*. Expansion is constrained by knowledge imposing consistency among components. Redundancy can help in making coherent components more evident.

Modern systems are based on probabilistic scores for candidate hypotheses. The speech waveform is sampled and quantized. A window, displaced by fixed time steps on the time sequence of generated samples, groups them into frames. Each frame is transformed into a vector of coefficients that are more suitable than the samples for further processing. The parameter vector obtained from the  $n$ -th frame is considered as an acoustic observation  $x_n$ . A spoken sentence is thus described by a sequence  $X$  of such vectors.

A simple, popular probabilistic model for scoring hypotheses has a decoder knowledge that considers the sequence of acoustic observations

$$X = x_1 x_2 \dots x_n \dots x_N$$

as the output of an information channel shown in Figure 3 that receives at the input a sequence of symbols representing the intention of the speaker. If these symbols are words, then they are usually represented by the sequence

$$W = W_1 \dots W_k \dots W_K.$$

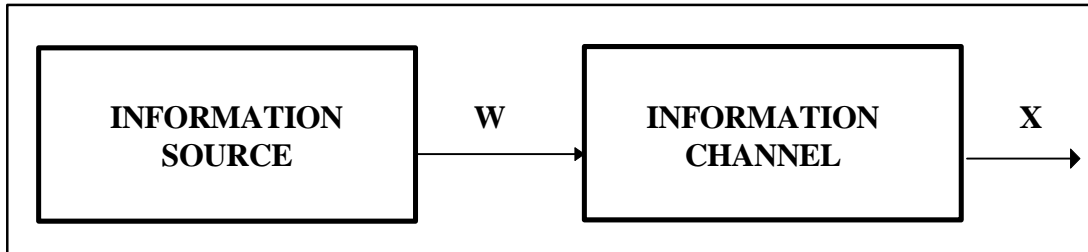


Figure 3: A simple decoder model

$X$  is a coded version of  $W$ . The objective of recognition is to reconstruct  $W$  based on the observation of  $X$ . This is done by using knowledge about the coding process. If the same  $X$  can be generated by different  $W$ , or knowledge is incomplete or imperfect, then reconstruction may not be successful.

In the case of dictation, ambiguity and imprecision make it necessary to consider recognition as a search process that generates word hypotheses by selecting candidates for which  $\Pr(W/X)$  is maximum. If the source model provides  $\Pr(W)$  and the channel model provides  $\Pr(X/W)$ , then the following quantity  $\Pr(X,W) = \Pr(X/W) \Pr(W)$  can be computed. Notice that, as  $\Pr(X)$  is the same for all the considered candidates  $W$ , the sequence  $W'$  for which  $\Pr(X,W)$  is maximum is also the sequence for which  $\Pr(W/X)$  is maximum. The decision rule used for recognition ensures a minimum sentence error risk.

$\Pr(X/W)$  is the probability of observing  $X$  when  $W$  is pronounced. In practice, this probability cannot be computed directly from data. It has to be computed using an Acoustic Model (AM).  $\Pr(W)$  is the probability of a sequence of words and is computed using a Language Model (LM).

Current machine dictation systems tend not to employ semantic knowledge in transcribing an acoustic signal into a sequence of words. Instead, given a description  $X$  of the signal, such systems output the word sequence  $W'$  such that  $W'$  maximizes  $\Pr(X,W)$  with respect to all possible word sequences  $W$ . If the objective is understanding, then the system has to find the conceptual representation  $C'$  that maximizes  $\Pr(C/X)$  over all the possible conceptual representations  $C$ . This can be expressed as:

$$\begin{aligned} C' &= \arg \max_C \Pr(C / X) = \arg \max_C \sum_W \Pr(CW / X) \cong \arg \max_{CW} \Pr(CW / X) = \\ &= \arg \max_{CW} \Pr(X / CW) \Pr(CW) \cong \arg \max_{CW} \Pr(X / W) \Pr(CW) \end{aligned}$$

$\Pr(CW)$  can be expressed as  $\Pr(C/W)\Pr(W)$  where  $\Pr(W)$  is computed by the LM and  $\Pr(C/W)$  is computed by a semantic model.

The choice of KSs and the way they are used in a system determines the *system architecture*. System architecture design should be based on a number of performance indices; the most important of them are now briefly reviewed.

A first requirement that has already been discussed is *coverage*. The system has to be able to recognize virtually all the sentences that can be pronounced. Another requirement is *precision*. KSs and methods for their use should produce the lowest recognition or understanding error rates.

A third requirement is acceptable *computational complexity*, both in terms of *time* and *space*. This has an impact on the central memory requirements for the hardware system. Having responses close to real-time is a necessary condition. This implies methods based on algorithms with linear time-complexity or with polynomial time complexity only if the input size is very small.

Knowledge can be manually compiled or obtained by automatic *learning* from a corpus of data. Coverage and precision of manually derived knowledge are often limited. The best results so far have been obtained using component models having a simple, manually decided structure. Statistical parameters of these models are estimated by automatic training. Complex knowledge structures are obtained by composition of basic models.

Figure 4 shows the scheme of a spoken dialogue system in which word hypothesis generation is represented by a single component.

These types of schemes are relevant for communication between users and systems with structures such as call centers. Access can be made through the telephone system or with multimodal interfaces. More details can be found in [2].

The main limitation of these technologies is that they are fragile because user models, language and environment models are highly imprecise. Thus, only applications in limited domains are possible and with cooperative users.

## **OTHER RELATED APPLICATIONS**

There are many other related applications of natural language technologies which won't be covered for the sake of brevity. They include, machine translation, speech to speech translation, document summarization, multimedia smart rooms.

## **POLICY CONSIDERATIONS DERIVING FROM TECHNOLOGY ANALYSIS**

It appears that basic and applied research are still needed in order to improve the technologies for handling documents and for natural person machine communication.

Unfortunately, improving basic technologies is a necessary, but not a sufficient condition. Other technologies have to be considered also, for example those related to telecommunication aspects, such as quality of service, and security.

Even when good technology will be available, they should be applied and integrated into valid prototypes which should be tested with well established procedures.

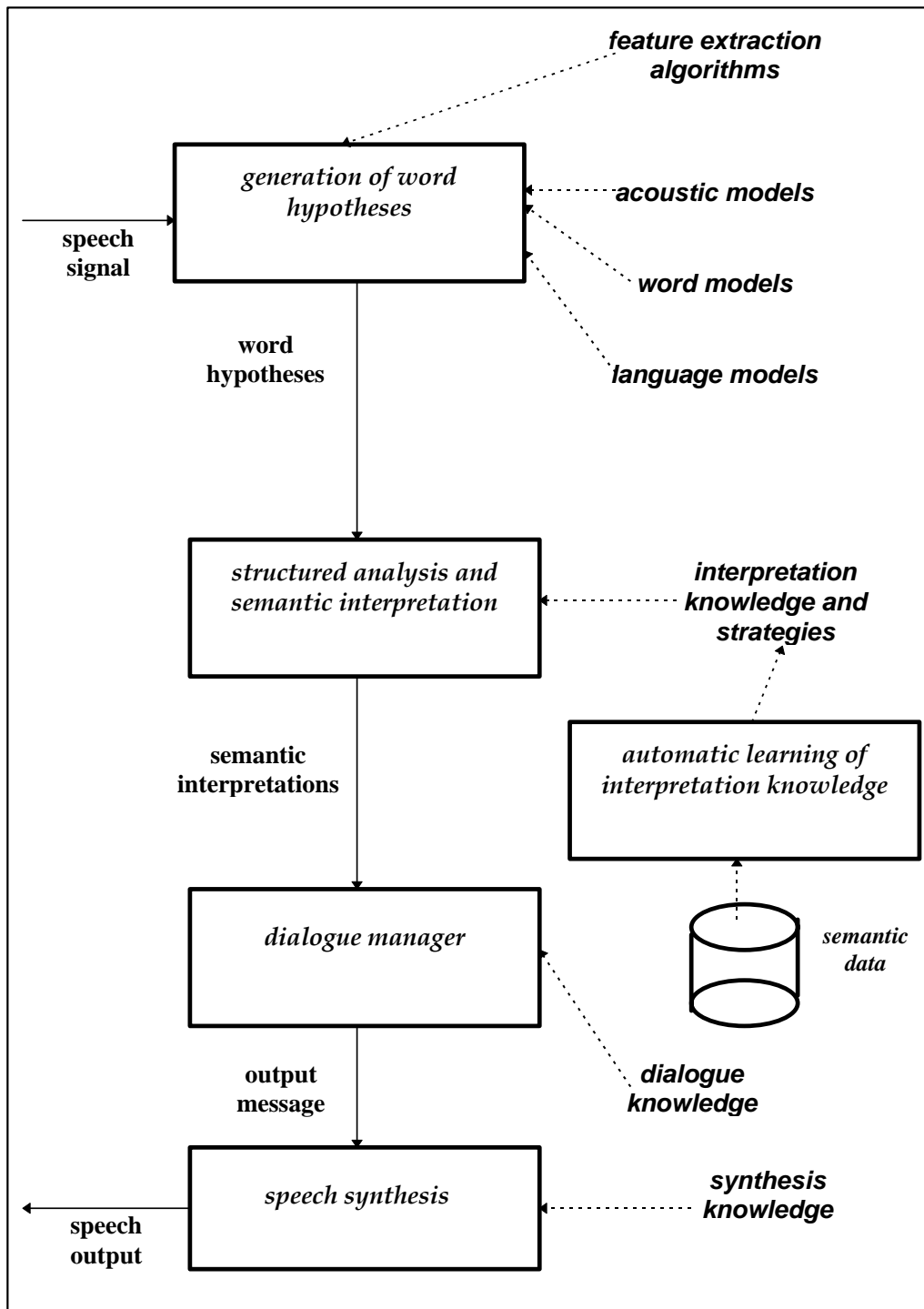


Figure 4 : Dialogue scheme

Well tested prototypes have to be transformed into good products. Successful business models have to be conceived.

It is risky to have successful products which can be executed only on a single hardware/software platform. Great attention has to be played to computer architectures, operating systems, software environments, telecommunication systems.

For the specific technologies considered before, these aspects, among others, are worth to be considered:

- *How close to real-world application are research results?*  
For speech technology, only limited applications are viable now.
- *What are the major problems that remains to be solved?*  
For speech technology, the main problem in robustness of processing and models.
- *What type of progress has been observed in the last years and what is the trend?*  
For speech technology, the progress is tangible with linear performance improvements over time. In absence of major discoveries, it will continue to be linear. If Word Error Rate (WER) is used as performance indicator, a ten times reduction has been consistently observed over many tasks in a ten years period and we are still more than an order of magnitude below human performance.

An interesting analysis of the progress as a function of time for language technologies is presented in [3]. The main conclusion is that there have been important improvements, but they are harder to quantify with respect to disk storage and processor power for which the Moore's law of linear growth as function of time apply with good approximation.

- *What is a likely application scenario for the future?*

For speech technology, the range of applications will be wider and performance will improve, but systems and models will have to be adapted to specific domains. Two types of industries appear to dominate the field. Those who sell just speech technologies (e.g. Nuance, Scan Soft) and those who embed speech and Natural language technologies into information and telephone systems (e.g. IBM and Microsoft).

Developments are along three lines, namely:

- telephone services (corporate call centers, voice browsers)
- devices (cellular phones and palm computers)
- laptops (Natural language integration into word processing software, operating systems and browsers)

Search is going to be predominant over data entry. Video is going to be more requested than speech and text, but their integration into multimedia systems seem to be the most likely scenario.

Better search tools will make appear the need for better and larger repositories.

- *What are the reasons why past business plans appeared to be incorrect?*  
For speech technology, the fact that the cost of system tuning, which is high, has been largely underestimated.

Supply appears to be now greater than demand. But demand can grow if killer applications are developed and more attention is paid to factors such as:

- packaging
- pricing
- platform
- infrastructure integration
- application development facilities
- delivery
- language coverage

A killer application could derive from better search methods.... but then... how much additional time are we going to spend with computerized systems?

- *What is a reasonable model for conducting research?*  
For speech technology, incremental system development should be carried out, considering failures of practical applications for developing new theories, methods and systems to be tried out with large corpora of real-life data.

It is important to distinguish what limits progress:

- Physics (probably the case of coding)
- Investment, especially in *competent* human resources (seems to be the case for document search and natural language interfaces).

It is important to balance research investments among two tendencies:

- Rationalism (making models based on theories)
- Empiricisms (inferring statistical models from the analysis of a large corpus of data).

## FUNDING RESEARCH PROJECTS

European groups are quite active in Information Retrieval research. Table I shows the names of some European Institutions active in the Text REtrieval Conferences (TREC), reporting results of competitions handled by the National Institute of Standards and Technology (NIST) of the USA.

Table I – European Institutions active in the TREC competitions

Alicante University	City University, London	CLIPS-IMAG
Tampere University of Technology	TNO TPD, The Netherlands	CWI, The Netherlands
Cambridge University	Fondazione Ugo Bordonni	Université d'Angers
Dublin City University	University of Amsterdam (2 groups)	University of Avignon
University of Bremen	Imperial College of Science, Tech. & Medicine	University of Glasgow
University of Hertfordshire	Institut EURECOM	IRIT/SIG
ITC-irst	University of Limerick	Kasetsart University
University of Neuchatel	LIMSI	University of Pisa
University of Sheffield	Moscow Medical Academy	University of Twente
University of York		

Most of the European groups participate to US competitions because this offers an access to rich sets of data. Unfortunately, these data do not adequately cover some important European languages and the characteristics of European users. Similar considerations can be done for data entry and human-computer interaction. There is a lot of research activity carried out in Europe, comparable in quality and quantity of results to the research carried out in North America.

There is a wide consensus on the fact that the organization of Information Repositories, search engines and data mining are going to play a very important role in the near future, even more than natural language interfaces which will also be of great importance.

If we consider speech technologies as an example, we observe that, in spite of the great investment and the importance of results obtained in Europe, the biggest and most active Industries are in the US.

The focus in these areas has not been always placed on the right things in Europe. The structure of the European funding system often forces good groups to seek integration into big and often artificial structures. Very often good groups do not find their place into these structures while average or mediocre researchers do. Also very often some good groups are funded on too many contracts, while other good groups are not funded at all.

It appears to be very important that laboratories and individuals with research results documented by prestigious publications or useful patents get support to continue producing useful results. Furthermore, it is also very important that these results are adequately exploited and are rapidly involved in a technology transfer process.

A number questions can be asked today about funding of research in these areas. Some of them, which probably apply to other areas too, are :

- To what extent did past projects succeed?
- What expectation, if any, was raised with unjustified optimism?
- What type of evaluation was carried out?
- Were the evaluators internationally recognized experts active in the field?
- Were good researchers left out because of the heavy bureaucracy and lobbying?
- Were mediocre researchers funded?
- It is well known that computer technology does not have yet a suitable set of standards as, for example civil engineering. What is the effect of unofficial *de facto* standards imposed by companies which dominate the market? Will free software be more suitable than commercial software?

Conceiving an interesting set of questions of this type is certainly a useful but not easy task. Answering them is even much more difficult. Some necessary conditions should be satisfied in order to make such an effort a valuable one.

First of all, the set of questions should be reviewed and refined by different panels of experts. Once a reasonable set of questions is identified, the best possible quantitative and pertinent data should be collected, Finally, panels of experts should comment and try to draw conclusions from them.

Doing all of this is far beyond the purpose of this paper. Nevertheless, some qualitative considerations will be presented, inspired by a personal experience in research, technology transfer and evaluation committees in Europe and North America. This intends to be a contribution to a debate and does not pretend to present certitudes.

Trying to answer these questions inspires a good set of properties a good project could have :

- Really original
  - Evaluation through high standard publications and patents

- Practically feasible
  - Evaluation with public or private corpora and procedures
- Made by people of caliber
  - Paying attention to publication (patent) record and citations

Actual limits to progress in this area appear to be lack of investment in human resources and coordination of experiments and efforts.

Some of the reasons which inspire caution in predicting great progress and which probably apply to sectors other than NL are:

- Difficulty of conceiving a valid theory for all the NL events generated by culture evolution for which an effective computational model has not been found yet.
- Automatic procedures have problems of coverage and precision as well as manual procedures which are error prone.
- Unrealistic or fuzzy visions leading to very different and often contradictory interpretations.

Often great progress is made by well focused small or medium sized projects carried by a small set of groups , even a single one.

The following aspects should deserve more attention in the future:

*systems*

- Problem solving and integration. Fill the gap between research results and their integration into valuable prototypes.
  - This capability can only be predicted based on past experience
- Consider all the components of information systems, ranging from architectures to applications. Pay attention to platform independence.
  - This is a debatable issue, but strong dependence on one non-European vendor does not seem to be a good situation
- Progress will be incremental. Major break-through are unlikely.
  - This should suggest a progressive increase in expected performance, rather than vague statements about global, unlikely results

*policy*

- Perform well controlled and state of the art system evaluations using real data and corpora.
- Consider that publications on top archival journals are a good index of the quality of the work.
- Bureaucracy should be simplified

- Find real experts for evaluation and monitoring .They should have a strong publication record or a strong record of achievements like patents, proven strong contribution in successful industrial realizations. Their selection should be made by panels of experts rotating in this task and not by officers.
- Avoid concentration of funding to small groups ignoring valuable resources in Europe which do not have the critical mass to support unjustified bureaucracy . Do not let a small group of lobbyists dominate evaluation and policy making.
- Pay more attention to real market and job creation potential. Continuously monitor the state of the world in research, industrial opportunities and market.

For what concerns the Networks Of Research Excellence, the following points are worth considering.

- Topic should be of strategic importance
- All the very good researchers should be in (those with extensive publication record or with an impressive records of achievements in things like international competitions)
- All those with poor record should be out.
- Those who show potential but did not achieve very high standards yet should be associate members.

## CONCLUSIONS

An analysis has been made on technical problems in the area of natural language interfaces and document retrieval. Such an analysis has inspired some considerations on research policy and funding strategies. Probably, these considerations apply to areas other than the one considered here.

## References

[1] <http://trec.nist.gov>

[2] R. De Mori. *Spoken Dialogues with Computers*. Academic Press, 1998.

[3] K.W. Church. *Speech and Language Processing: Where have we been and where are we going*. Proceedings EUROSPEECH03, pp. 1-4. Geneva, Switzerland, Sept 2003.