Automated Analysis of Unstructured Texts

Technology and Implementations

By
Sergei Ananyan
Michael Kiselev

Why natural language texts?

Automated analysis of natural language texts is one of the most important knowledge discovery tasks for any organization. According to Gartner Group, almost 90% of knowledge available at an organization today is dispersed throughout piles of documents buried within unstructured text. Books, magazine articles, research papers, product manuals, memorandums, e-mails, and of course the Web, all contain textual information in the natural language form. Analyzing huge volumes of textual information is often involved in making informed and correct business decisions. Traditional analysis methods based on statistics fail to help processing unstructured texts and the society is in search of new technologies for text analysis. There exist a variety of approaches to the analysis of natural language texts, but most of them do not provide results that could be successfully applied in practice. This article concentrates on recent ideas and practical implementations in this area.

The implementation of new approaches to the natural language text analysis brings much closer the fulfillment of an old time human dream of having an intelligent, relentless, loyal, and inexpensive electronic adviser. Upon outlining the basic principles of the new technology in this paper, we discuss their concrete implementation in the text mining system TextAnalyst. This system is capable of automated analysis of natural language texts from arbitrary application fields. It can distill the meaning of a text and help navigate a textbase, create summaries of documents, cluster documents, and carry out semantic information retrieval on a collection of texts (Figure 1). TextAnalyst is available as a standalone application or as a set of COM-based modules implementing individual analytical functions.

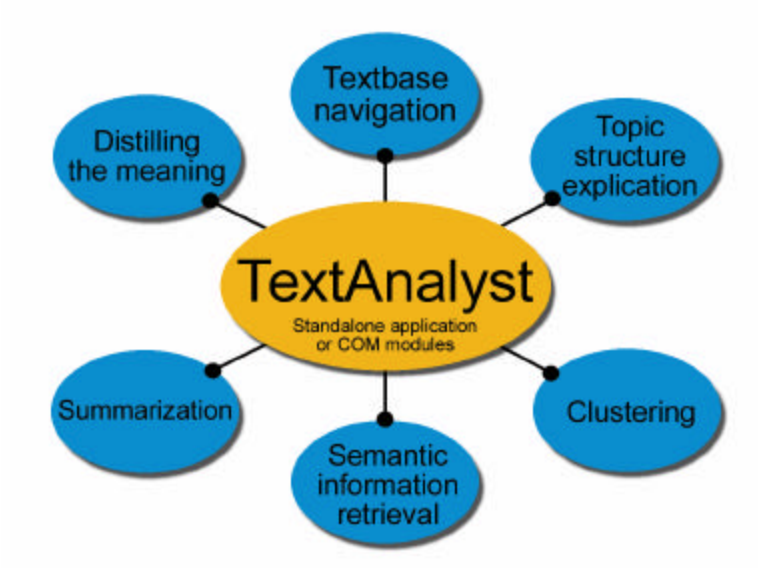

Figure 1. TextAnalyst functionality.

New opportunities: business prospective

Let us consider the most common tasks arising in relation to text analysis. First, one would like to be able to automatically distill the meaning of a text in a concise form and store the results as a list of the most important concepts from the text hyperlinked with the corresponding places in the original text. This procedure would provide a new efficient mechanism for navigation through texts, automated creation of summaries of documents, clustering and classification of texts, comparison of texts, as well as natural language information retrieval. Achieving this functionality could have profound practical implications for our everyday text processing activities.

By and large, we all have to deal with reviewing large volumes of textual information. At the same time, for some professions automated intelligent text analysis capabilities can be critical. An automated text summarization function could be used by government and business analysts, magazine editors, venture capitalists, lawyers, and students, who wish to see accurate summaries before plunging into the full documents. An efficient navigation through a textbase, as well as summarization, clustering and classification of texts, could enhance the effectiveness of working with large textbases including academic documents (for researchers), electronic news flow (for marketers and investment bankers), and e-mail systems (for all users). An automated classification of incoming messages to different subject groups and priorities through the analysis of their contents, as well as their efficient retrieval at a later time could help to heal the trauma of our e-mail experience. The semantic information retrieval capability could save millions of man-hours by increasing the relevance and precision of a database search or Internet surfing. Clustering a collection of documents that represent the press reaction to the latest marketing moves of your company and your competitors could help assess the effectiveness of your marketing campaign. A combination of all of these functions with a

natural language information retrieval capability could facilitate creating a new generation of powerful and intelligent corporate Help Desk and Call Support Center solutions.

The prospects look bright, but the problem is that all the attempts to build practical systems for automated analysis of natural language texts have not produced satisfactory results thus far. The created systems usually work well only in a certain application field and require significant and costly human interference at the stage of tuning the system to a new field. Thus the objective would be to develop a new approach for more versatile and automated analysis of texts from different subjects. Let us first briefly discuss traditional text analysis techniques in order to identify their strong and weak sides.

The history of the subject

There is a long history of attempts by researchers in the fields of Artificial Intelligence (AI) and Artificial Neural Networks (NN) to understand and model the information processing capabilities of the human brain. Research work in this area has been going on since the fifties. A variety of partially successful approaches to processing natural language texts have been developed.

In general, systems based on traditional approaches analyzed a natural language text in a certain way at the level of individual sentences. The objective was to create a semantic representation of a sentence in the form of structured relations between important words comprising this sentence. To solve this task, various predeveloped linguistic molds were tried with the sentence and its components. When a mold matched the sentence well, a corresponding semantic construction was associated with the sentence. This technique provides a good first guidance for understanding the meaning of a text. But as it turns out, the main problem with this approach is that there can be too many different molds that one needs to build for analyzing different types of sentences. In addition, the list of exceptional constructions in this approach quickly grows prohibitively large. In other words, this approach works well only for a limited subset of natural language texts.

One of the traditional branches of Artificial Intelligence (AI), known as the field of natural language computer-human communication, is mainly devoted to automated processing of texts. This branch includes machine translation, semantic search for information, and creation of expert systems. Here purely linguistic methods are implemented for the analysis of the text semantics. The results of the analysis are represented in the form of a *semantic network* displaying a list of the most important words from the text and relations between them. A semantic network is a convenient text representation object often used in cognitive sciences. It should be noted that a set of the created linguistic rules works well only with texts within the subject for which these rules have been developed. Thus the performed analysis is strongly dependent on the background knowledge of the analyzed field. This also implies that a human expert must be involved at the stage of the development of linguistic rules for a subject. Such an approach works well for the creation of expert systems that are utilized only in a single

application field. Yet, in order to successfully analyze texts from arbitrary fields, one needs to employ more general algorithms.

Another approach applicable to processing unstructured texts, artificial Neural Networks (NN), was developed with the hope that a homogeneous artificial processing media made out of connected elements similar to the brain neurons could indeed process information similarly to the human brain. Again, it has been demonstrated that systems based on this approach are capable of successfully solving simple analysis tasks. However in general, a homogeneous processing media is not suited well for the analysis of linguistically structured information. Developing a new type of structured processing media is required for tackling this task.

Summarizing, both AI and NN approaches provide important insights into the problem but demonstrate only limited success in practical applications. In fact, the most promising techniques for the analysis of natural language texts reside in the overlap of the two fields. One very interesting approach is to employ for text analysis the media consisting of parallel processing units, as in the NN approach, while structuring this media according to cognitive models of AI, where the task is split into a number of subtasks connected by information flows represented in terms of cognitive models. In this way, a more rich and complex construction suitable for the further text analysis can be formed. The information is processed automatically as in NN, while at the same time giving birth to semantic structures, which are often encountered in AI.

New approach to text analysis

Basics

Let us outline the principles of a new approach in more detail. In the new hybrid method, the text is considered as a sequence of symbols organized into words and sentences. This sequence is moved through a window of variable length (from two to twenty symbols can be seen simultaneously), shifting it by one symbol at a time. The snapshots of the text fragments visible through the window are recorded in dynamically added neurons.

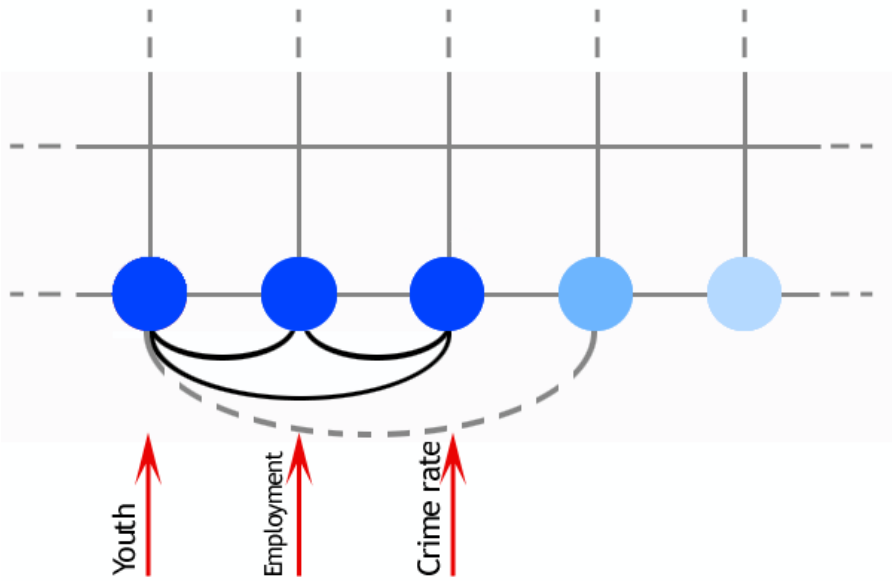

Figure 2. Dynamically growing neural network records new text fragments.

The created hierarchical neural network contains several layers: those fragments that occur in text more than once are stored in neurons that belong to the higher levels of the network. This neural network realizes frequency-based multi-level dictionaries of different text elements (letters, syllables, stems, morphemes, words, and phrases). Words are selected as basic operational elements, while other elements are used as auxiliary information during the analysis.

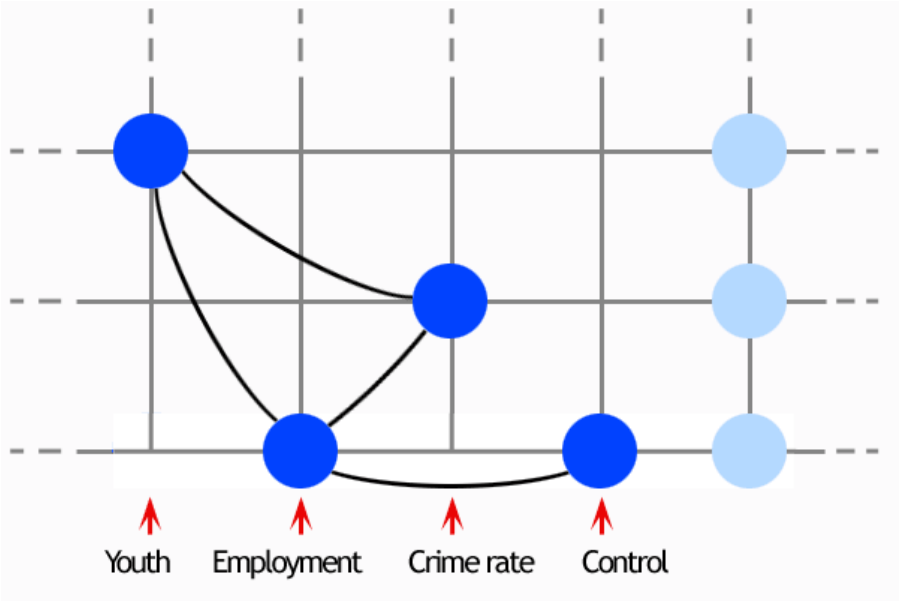

Figure 3. Hierarchical recurrent neural network traces frequencies and relations of terms.

Preprocessing
Ideally, one wishes to get rid of all supplementary and commonplace words, which carry no semantic meaning. Also one would like to identify stems of the words, while separating prefixes, suffixes, and endings (morphemes). This step is called preprocessing. All further work can be carried out with stems only, thus improving the quality of the analysis. For example, the words “mean” and “meaningful” will be identified as having the same stem by this system.

In fact, obtaining an efficient preprocessing mechanism requires fine-tuning the system to a specified language in order to efficiently filter out supplementary words and morphemes native to this language. One could utilize the same hierarchical neural network in order to build a filter for unwanted elements. When processing a large corpus of texts from diverse subjects, supplementary words and morphemes are the fragments appearing most frequently in the text. By working with various fragments of words, the hierarchical neural network allows one to automatically catch both supplementary words and morphemes at the same time. Note that this preprocessing is the only place where language dependency enters in the discussion of the new analysis technique and where some human analyst guidance is desirable. All other components of this technology are language independent and work equally well with texts in any alphabet-based language. Applying a threshold to the neural network developed on such a corpus of texts, one creates a filter that can be used later for separating the stems of semantically important words for further analysis. While performing the analysis with individual stems, the network still holds the information about complete words.

Let us assume that we managed to filter out meaningless elements and process the significant information. The nodes of the developed neural network now hold all important words and word combinations from the text with the frequencies of their occurrence. Simultaneously, the same network assesses frequencies of joint occurrence of different semantic elements within certain structural text units, for example sentences. One obtains a graph-like structure that contains statistical weights of words in the nodes and statistical weights of joint occurrences of these words in the links.

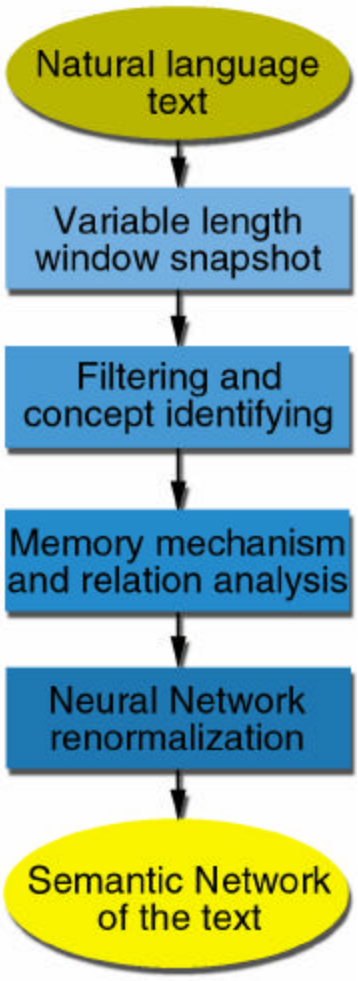

Figure 4. Semantic text analysis workflow.

Renormalization

This graph does not provide an accurate semantic picture of the analyzed text yet. One still needs to adjust individual statistical weights of the words and relations between them to provide a consistent text representation. The weights of those words, which are strongly related to other frequent words in the text should be boosted, and vice versa. This is accomplished by assigning the statistical weights of individual words to the nodes in a one-dimensional Hopfield-like neural network where all neurons are completely interconnected. Simultaneously, the statistical weights of relations between words are assigned to the links between individual nodes in this network. When released, this Hopfield-like network evolves by changing the weights assigned to the nodes and links between them to a stable configuration corresponding to the minimum of an energy-like function characterizing the network. The renormalized weights of words and relations between them are called semantic weights and the resulting reshaped graph-like structure is called a semantic network (which is a list of the most important words and word combinations from the text and relations between them).

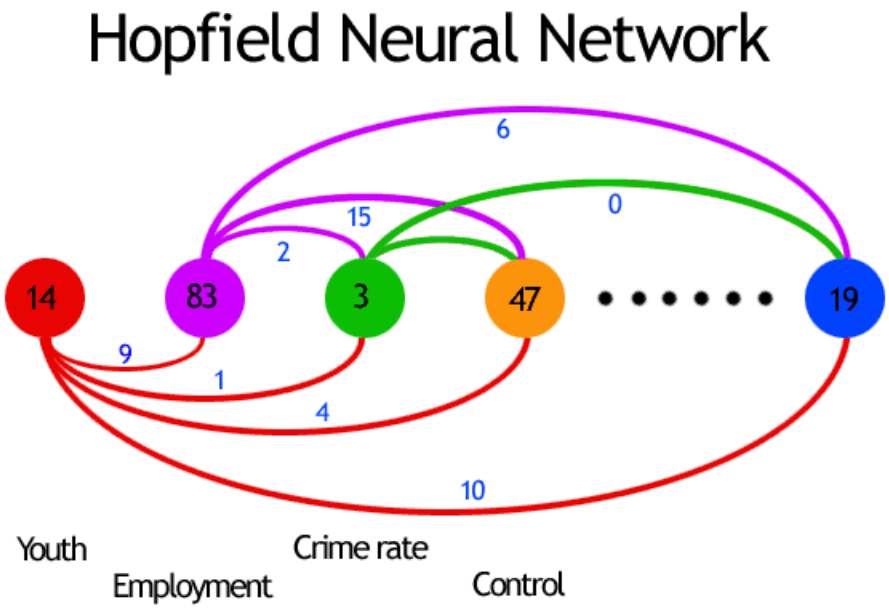

Figure 5. Hopfielfd-like neural network refines semantic accuracy of the analysis.

Since the analysis of a text has been performed with no recourse to any background knowledge of the subject of interest, the meaning of a word in the created semantic network is defined purely by those other words, which are related to it in the network. Correspondingly, the words and word combinations comprising a semantic network have a special name - semantic concepts. The semantic network represents a linguistically accurate and concise picture of the analyzed text. This construction can lie in the foundation of many further analysis techniques implementing user-needed text processing functionality.

TextAnalyst: natural language text analysis software

The new text mining system, TextAnalyst, implements a variety of important analysis functions based on utilizing an automatically created semantic network of the investigated text. This system is built on the results of twenty years of research and development of a new paradigm by a team of mathematical linguists. The key advantage of TextAnalyst against other text analysis and information retrieval systems is that it can distill the semantic network of a text completely autonomously, without prior development of a subject-specific dictionary by a human expert. The user does not have to provide TextAnalyst with any background knowledge of the subject – the system acquires this knowledge automatically.

TextAnalyst empowers the user with the following functionality:

Textbase navigation
In TextAnalyst, concepts stored in the semantic network are hyperlinked to those sentences where they have been encountered, and the sentences are in turn hyperlinked to the places in the original text from where they have been retrieved. Thus the automatically created semantic network provides an efficient navigation through the texts stored in the textbase. Keeping in mind that thousands of texts can be processed simultaneously, the outlined semantic navigation might turn out to be a very handy capability.

Topic structure
The system can identify the most important concepts from the semantic network and transform the network into a tree-like list of nested topics of descending importance by breaking links representing weak relations and substituting certain indirect relations with direct ones. This transformation reveals the hierarchy of themes in the investigated text.

Clustering
This function goes a step further and eliminates those links in the topic structure whose strength falls below a certain threshold value. In this way a joint topic structure of a collection of texts breaks into islands representing certain largely independent themes, which help understand the clusters of information in the investigated textbase. Then individual documents can be assigned to different thematic groups, thus facilitating clustering of the documents in a textbase. Of course, occasionally large documents might have several parts corresponding to different thematic clusters. Such documents can be treated as multi-topic, or they can be split in separate parts.

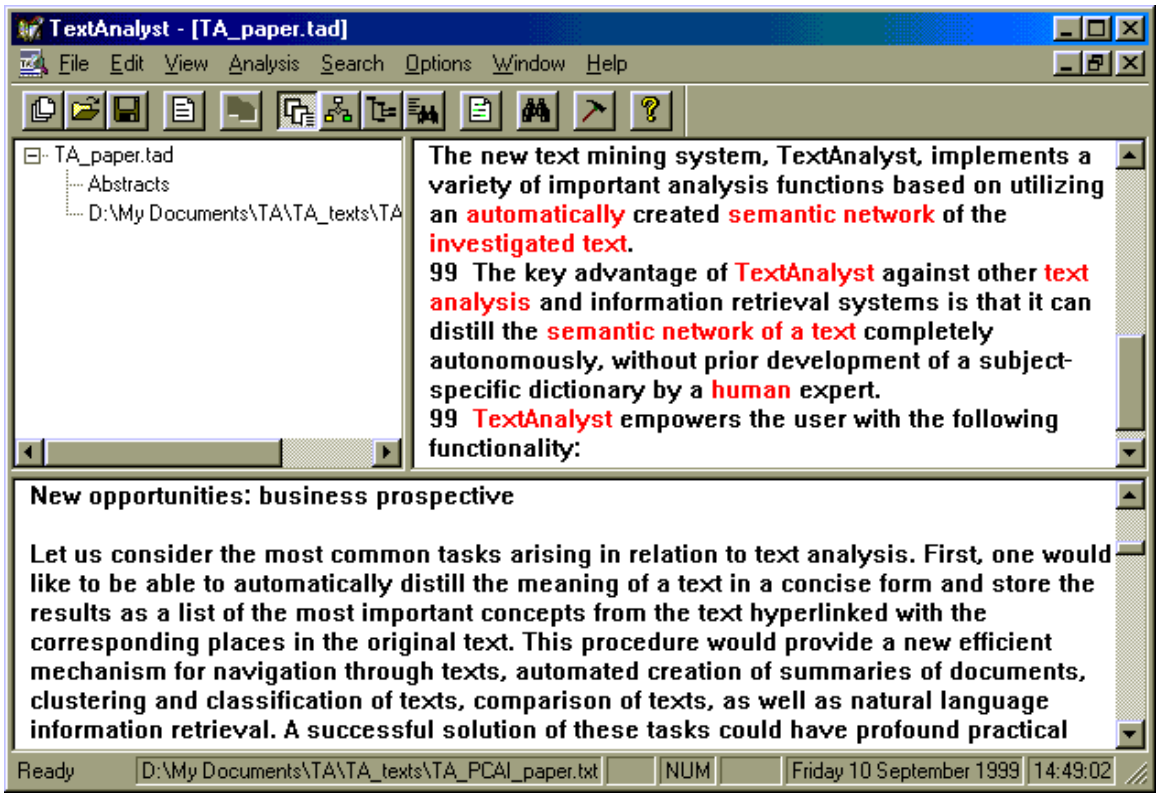

Figure 6. TextAnalyst document summarization.

Summarization

The semantic network can be utilized to score individual sentences in the investigated text. The larger the number of important semantic concepts in a sentence is and the stronger these concepts are related with each other, the higher the semantic weight of the sentence itself is. Then the system collects only those sentences that have a semantic weight higher than a certain adjustable threshold value. This results in summarizing the investigated text. The size of the summary is controlled through changing the sentence selection threshold. An advanced algorithm used for developing an accurate semantic network ensures the high quality and relevance of the created summary.

Natural language information retrieval

The system determines whether an issued natural language query contains words present in the developed semantic network of the investigated text. After that, the sentences containing the identified words are retrieved. Thus one does not have to come up with a predetermined list of key words for a search: the system automatically extracts from a natural language query the best words to utilize. Still more important, the system displays a subtree of concepts that are related to the theme of the query in the context of the analyzed text. These concepts are taken from an immediate neighborhood in the semantic network of the text of the words distilled from the query. This feature allows the user to view an immediate semantic context of the searched theme in the textbase and dive into related subjects to refine the search.

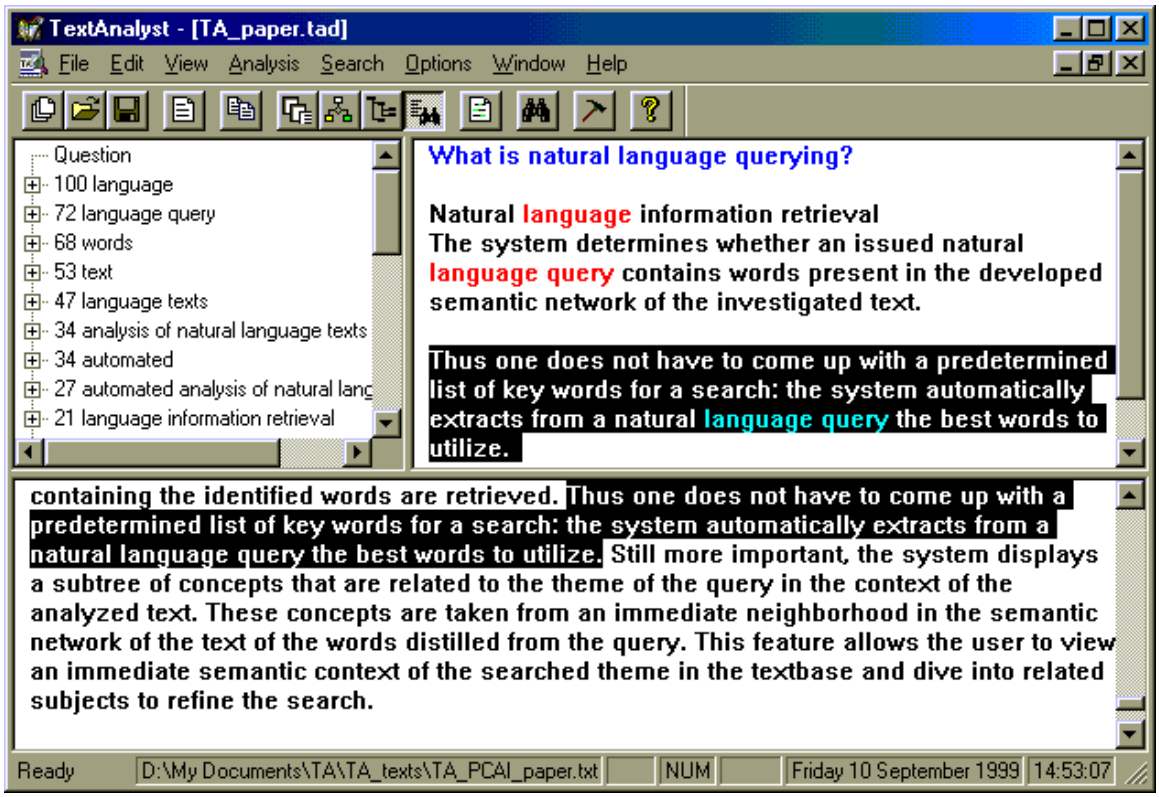

Figure 7. Natural language text retrieval function.

In addition to these important functions one can utilize the described technology of automated creation of an accurate semantic network of the text to provide the user with many other crucial text analysis capabilities. Currently the development team of TextAnalyst is working on implementing automated classification of documents. Measuring the similarity of individual texts is another future feature under consideration.

TextAnalyst is available either as a standalone application for MS Windows or a set of COM components that can be easily integrated in an external decision support system. Further information about the system and an evaluation copy of TextAnalyst are available at **www.megaputer.com**.

Sergei Ananyan is the President of Megaputer Intelligence Inc. He has M.Sc. degree from Moscow State University and Ph.D. degree from College of William and Mary. Sergei joined Megaputer in 1994 and since then carried out numerous data analysis projects, guided the development of data mining tools, and wrote numerous papers in leading academic and trade publications. He can be reached at s.ananyan@megaputer.com.

Michael Kiselev is the Director of Research and Development for Megaputer. He has M. Sc. Degree from Moscow State University and Ph.D. from Moscow State Technical University. Since 1993 Michael headed the majority of Megaputer R&D projects resulting in the creation of some of the world's most advanced data, text, and web data analysis tools. He can be reached at m.kiselev@megaputer.com.