Science Arena Publications
Specialty Journal of Engineering and Applied Science
ISSN: 2520-5943
Available online at www.sciarena.com
2019, Vol, 4 (3): 10-18

Surveying Business Documents and Their Processing Methods

Amirhushang Tajfar[1]*, Reza Momeni[2]

[1]Faculty Member of Information Technology Department, Payame Noor University, Iran.
[2] Master's student at Payam-e-Noor University, Iran.

***Corresponding Author**

Abstract: *Scientists have always been in a position to move closer to human intelligence in various areas of decision making and computer recognition. One of their areas of interest is natural language processing, which involves the following sub-sections, including speech processing, recognition of printed texts that are sub-types of image processing, analysis of text meaning, and so on. The importance of progress in this area is determined by the extent of available information and documents and the need for their analysis of translation and resolution of defects to understand their meaning. In this article, we discuss the concepts of commercial documents and methods of natural language processing as well as electronic processing of commercial documents. Although it does not do the justice as it deserves and the educated ones should add to its richness by criticizing it.*

Keywords: *natural language processing. OCR OMR.ICR. DIA*

INTRODUCTION

As information technology advances and the need for information increases, the volume of documents containing information gets more and more. Contrary to the use of electronic documents, the volume of paper documents has never been reduced since the volume of publications, newspapers, magazines, reports books, etc., is steadily increasing, and most people prefer paper documents for archiving and reading. But storing and retrieving all of the paper documents that are increasing in size is hard and cumbersome. Instead, e-documents have many advantages, such as increasing use of computers and the Internet, compact and unprotected storage, efficient retrieval, rapid transfer and explicit structure. Therefore, extensive research has been carried out on the conversion of paper documents to electronic documents. The algorithms and techniques applied to the document images to provide a computer-readable description of pixel data, are called document analysis. The purpose of document analysis is to recognize text and graphic components in the image of documents and extract desired information, as human wants. Document structure analysis is the most important step in the process of document analysis and retrieval, including separation of text and graphics, and other components of a document, and then extracting separate text blocks such as titles, headers, subtitles, and image descriptions. The algorithms in the image of a document looking for text, graphics, and other components of image are divided into three major categories: top-down, bottom-up, and mixed. The top-down methods divide the image continuously into smaller components to find desired components. In contrast to top-down methods, bottom-up methods start with pixels or connected components

and, by combining them, make the components of the image. Mixed methods use a combination of the two methods described.

Concept of commercial documents

The document is out of business within the framework of civil law and it is a new system of its own. From the paper mode which is used as a proxy, this document, which is not independent in the civil law context, acquires an abstract code of practice and benefits, and separates it from the simple description of ordinary documents in civil law. Commercial documents are mainly all documents exchanged between traders in daily transactions, their varieties vary widely, depending on the status and type of business. However, in trade law term, trade documents are trademarks that are named by the Commercial Code and placed under certain conditions (Feizi and Faroozi, 2014).

Importance of converting paper documents into electronic documents

As information technology advances and the need for information increases, the volume of documents has included more and more information. Contrary to the use of electronic documents, volume of paper documents has never been reduced, since volume of publications, newspapers, magazines, reports, books, etc., is steadily increasing, and most people prefer paperwork for archiving and reading. But the storage and retrieval of all of the paper documents, which are increasing in size, is hard and cumbersome. Instead, e-documents have many advantages, such as: increasing use of computers and the Internet, compact and unprotected storage, efficient retrieval, rapid transfer and explicit structure. Therefore, extensive research has been done on converting paper documents into electronic documents.

- **Why does the need for paperwork continue?**

In fact, the advantages of none of these two types of document cannot be ignored. The everyday and extensive use of paper documents and the growing use of electronic documents, which, in addition to the features mentioned, have other benefits, such as the ease of long-term maintenance and copying, as evidenced above. What needs to be done is a two way communication between paper and electronic documents so that human beings enjoy the use of both of these documents in their way of life in the world of information and take steps as quickly as possible. "A scan of a printed document is returned to a document descriptor language, which, if needed, can be reprinted exactly". In today's world, documents are being produced more and more on computers. In the meantime, the share of paper documents will not be lost, and because the interface is very easy to use, at least in the next decades, they will stay with us. The only difference from the past is that they will eventually be integrated into our computerized world.

Natural Language Processing

In natural language processing, the computer is trying to extract meaning of natural language resources. Due to the complexity of natural language processing and the ease with which it faces, various specialized tools can be used to deal with any of the problems. The syntactic tag reader is one of the tools that can help natural language processing process. The task of the labeler is to specify grammatical role of words in sentences given as inputs to the computer. This basic tool can be used in machine interpreters, falsifiers, text summaries and more. Labeling is largely divided into two types of law and statistics. Due to the problems of labeling, they are based on the major laws of implementation of Persian language, using statistical methods. Natural language processing is in fact a baseline in both the field of artificial intelligence and linguistics. The ultimate goal of this discipline is the ability to understand human beings from the sources of natural human language in computers. In terms of semantic analysis of the text, there are several things that can be said about speech recognition, word falsification, machine translation, typographical recognition, text retrieval, text summarization, and other applications. Of course, it should add to this list of the ability to consult and answer questions with access to extensive information resources such as web pages and encyclopedias. According to the above definitions, we are looking for a system that can both understand natural language resources and its ability to produce. The first ideas of natural language processing coincided with the birth of artificial intelligence and in the same Turing test. Natural language processing is done on a variety of levels, which can

be divided into the following: phonology and vocalization: recognizing sounds and voice recognition and speech recognition.

Morphology: It focuses on the structure of words and the orientation of vocabulary.

Syntax: It addresses the relationship between words and grammar in groups and sentences.

Semantics: It deals with semantic relation of words.

Pragmatics: The use of language to convey an article to an audience or audiences, in practice, or in natural text and speech.

Discussion: deals with the general communication of a language in one or more specific sentences.

Early attempts at this research were unsuccessful, and afterwards scientists came to the conclusion that they needed another field of science to help with it. Linguistics theory was something that could help to succeed in these efforts, but it did not exist until then. Publication of the book *syntactic structure* by an American theoretical linguist, Chomsky, helped to solve many of these complexities, but there are still many problems facing natural language processing researchers. Problems, many of which have inherent ambiguity, complicate work for researchers. For example, the ambiguity of meaning of sentences in the natural language, due to multiple meaning of some words, tone of speech and natural language grammatical structure, add to this complexity. Another problem is the lack of precision of natural language grammar, so that neither of them is so precise that is always possible to recognize the role of sentence components using their position. Natural language processing requires several methods, algorithms and tools in different fields and processes to achieve its goals. One of the tools that helps natural language processing is normalizer, vocabulary and sentences detector, root finder, parser, semantic tagging of the word, corpora, syntactic tagging of the word and so on.

- **Syntax analysis**

In order to receive and interpret natural-language sentences, they must be syntactically analyzed. Even for semantic analysis and machine translation, there is a need for syntactic analysis, and in machine translation, the syntactic analysis and analysis section is considered to be the main pre-processing unit of operation. There are also different approaches to syntactic errors. But what is clear is that due to the ambiguity and possibility of having multiple parsing for a phrase or sentence in the natural language, it is impossible to use usual parsing methods used in the programming languages of the computer. To illustrate the issue, though, significant improvements have been made to English language tagging, but it's still a far cry from the sidewalk. In the Persian language, research is backward-looking and there is plenty of work for researchers. In the Persian language domain, a few points can be helpful in explaining the issue to the reader. First of all, human mind has many amazing features, including without the ability to do things that make it hard to implement algorithms for the computer. We use tagging algorithms to diagnose role of words that are capable of obtaining multiple labels, or in other words, to resolve ambiguity of determining their role. There is a huge amount of information in non-structured texts on the web, whose analysis of these texts has found great importance, and maybe even more important than the extracted structured data, because of the net volume of valuable information. The analysis of emotions deals with automatic detection and classification of beliefs mentioned in the written text in the natural language (Qasemzadeh and Izadi, 2014). Think Tank Research is the focal point of research from several research areas such as machine learning, natural language processing, information retrieval and text mining. The use of natural language processing methods for optimal use of information resources is one of the requirements of technology era. Due to the increasing spread of information on the web and relationship between belief mining and e-commerce, greater human-computer interaction and direct use of workplace has gained popularity in recent years and has become a widespread research field. Therefore, the search for ideas from Persian language news on the Web has been considered in this research.

What is OCR?

Suppose that we have text on paper and we want to put it into the computer. The first way to get in mind is to type the text into a typo to type with the computer. But can we just enter the same text into the computer so that it does not need to be typed? Of course, Scanner device can enter a picture of that text into a computer, until a part of our problem has been resolved here. But a computer that has neither a rational nor a "language" cannot distinguish between letters and words. For example, if we ask the computer to tell us that in the scanned text, the word "Ali" has come several times, without being ashamed, he says, I cannot diagnose! In fact, this "digital image" should be converted into a "processing image". This is the main issue of OCR.

OCR is the term used to describe full text in English dictionaries in two ways:

Optical Character Recognition

Optical Character Reader

- **types of OCR:**

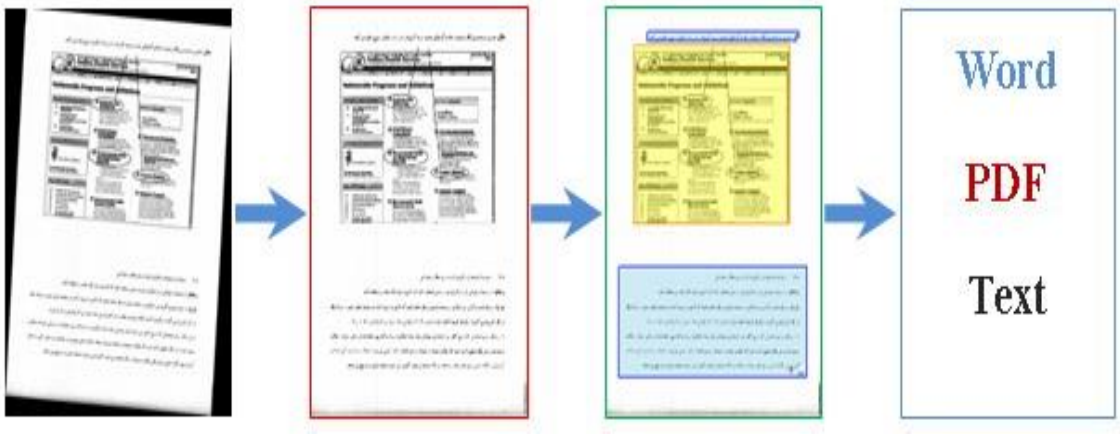
In other languages, especially those written in Latin letters, OCR has been used for many years. But in Iran, it has been two or three years now that we thought of using OCR in Persian.

There are two types of OCR as follows either typing or handwriting. That is, we need to enter a previously typed text (such as books and newspapers several years ago, or even text that is not a file of its type, and only have it printed), or the handwritten text. Handwritten texts also exist in both "discrete" and "continuous" terms: the "manual manuscript" text is the same as what we write on paper every time we miss it, either a letter or a piece of poetry ... But "discrete handwritten text" is the same written writings that are written separately and discretely, such as the name and surname written in the test forms for registration, in each letter in a cadre. OCR design of Persian split is almost at the final stages of work, but OCR continuously seems to be working for many years[1].

The word "OCR" was first used only for recognition of digits and letters. Optical extension in this expression was placed against a magnetic ink symbol to distinguish this method from earlier method of recognizing characters with magnetic intruder, MICR. Over time and with significant progress in this field, there were methods for recognizing handwritten and printed texts that brought the domain to words and phrases. Despite the lack of precision in OCR with these cases, the name was also used for these methods. They now use OCR more to recognize printed documents such as books, magazines, and letterheads.

A character reader, such as a typist, reads the text of a document and converts it into a format suitable for saving on a computer. Usually a scanner will provide the document image for OCR. The script system recognizes objects in document's image, which are digits, letters, symbols, and words, and stores corresponding string in appropriate format. An image file is large and text search is not possible. This is while the file output system is very small and searchable. While text file information can be easily edited and used elsewhere, the information in the document image cannot be edited. Character reader systems like many other smart systems have a lot of complexity. Image processing and pattern recognition are the two main bases of these systems. The complexity of these systems is different for different languages. For example, writing OCRs for Latin languages is easier because their letters are written separately, from languages such as Persian and Arabic, which stick together the letters of a word. This, in addition to low population of Farsi users, has caused some Persian script systems to be available for Persian language.

[1] iranwsis.org



An OCR system consists of several parts. One must first analyze the input image and correct it if it has a rotational text. After changing rotation, location of text blocks, shape, and table must be specified in the document image. After determining position of different blocks (regionalization or document structure analysis), the text blocks must be identified; that is, the text strings are found, and then position of words is specified, and in the next step, position of letters in word will be specified; eventually, each letter should be known and combined to formulate their equivalent word. This was the whole process of automatic text recognition, which was expressed very simply. It should be noted that due to complexities of Persian language with Latin languages, OCR Persian production is far more difficult than English ones. Therefore, the products available in this area are very limited in Persian.

- **Persian Language and its problems**

We need to address some of the problems of Persian line or indeed features of this line. Firstly, we write in letters and cover letters continuously in Persian, and it is very difficult to distinguish characters from words written by computer (which is to be typed in the next steps). Imagine that the simple word "است " is written to different states: one places a tine for a "س ", one does not place, one pulls it and one does not, and ... Now if these different forms of "س " are linked to "ت", It's hard to recognize the letters for us, let alone the computer (Hosseinpour, 2010).

- **Similarities of letters**

Another problem with our line is that Persian letters are very similar. Consider, for example, that the difference "R" with "Z" with "D" or "B" with "T" is only one point, and because the point of the component is very small, if a line or even a small stack falls on the paper, it is difficult to distinguish between letters and creates a serious problem in recognizing letters by computer. These are new problems in Persian line.

There is also a problem with Persian numbers: Zero is a small point that can make a computer mislead; numbers 4, 3, 2, and 1 are very similar, and their only difference is a small dent. Typed texts are continuous, and there is much to do with recognition of manual manuscripts by computer, since in discrete manuscripts, although the letters are similar, they are written at least separately. In continuous text, we do not have difficulty of writing a letter or a broken letter (Ebrahimi and Behin, 2010). For the reasons stated, OCR in our current situation in "discrete manuscripts" is in our country, or, according to officials of Paya, there are currently software that converts manual manuscript into disjointed letters, but error factor of this software is high and not industrialized.

- **Recognition of Letters and Patterns**

So far, the picture of a page in which, the letters are spatially separated (each letter in a box) is written to the computer by scan. The next step is to recognize letters, that is, their location is recognized from other lines (such as lines of box in which it is written), and if the text is continuously typed, the letters are separated and the corners of the image are removed. For example, if the student wrote "S" in a way that was out of the box, to tell the computer that it was reckless, he should read the same letter in the box.

In the next step, called "Template Recognition", it is understood by a number of conditions that, for example, it is "A" or not, and the computer recognizes that it is "C" or "B". For this diagnosis, it is necessary to match the letter "A" with the prototype elves that were previously given to the computer. The sample alphabet has already been developed from a large educational collection and its common features have been extracted. But since the diversity of facial expressions is a very large manuscript writing, a statistical model is extracted in which the similarity of previous extracted properties with the sample of input into computer is examined. Here, "Pattern Recognition" is done with statistical methods, which is a common practice in OCR systems. If you think that the work is over, you are wrong, because we are just coming up with letters. For example, if someone writes the same letter "S" in sequence, the computer must recognize that this is just "S", or "Y" is sticking to it, for example.[2]

Modeling or linguistic processing

The next stage is called "linguistic modeling" or "linguistic processing". Letters that stick together, which make the word, must be meaningful or well-known. At this point, what words are in the language? What are the combinations of words allowed? And, of course, in more advanced stages, there is grammatical modeling (grammar) and semantic modeling that recognizes that the sentence is verbose and semantic is true or unobtainable. But in the discrete OCR, which is most used to register, the likeness of a word in the name, family name, city, and ... is sufficient.

We need databases to identify permitted combinations of a word or meaning of a word. In these banks, for example, all small and big names of Iranians have already been collected and when it comes to matching a word with it, it is determined by computer's handwriting. Therefore, the role of this database is very important, because if the name is not registered, the word that includes that name is automatically deleted from OCR program or message that "this word is wrong" if it is possible for example, the name "Hisham" exists among Iranian names, but has not already been registered in the database.

- **Destiny of Handwriting OCR**

In the case of manually handwritten OCR, the process is as described, but what makes the task more difficult is segmentation and separation of letters together and their diagnosis. If this trend continues, it is hoped that one day the Persian handwritten OCR will be widely used. Of course, continued manual OCR, even in English, is still not widely used. "The English OCR is available in the Windows operating system, which is sold with the office, but do not assume that the English office letters written with the handwriting are all typed with OCR," says Rosazi, an engineer. This must be done to achieve at least one 10-year trend. The Persian OCR is a step back, so it will take more time. "

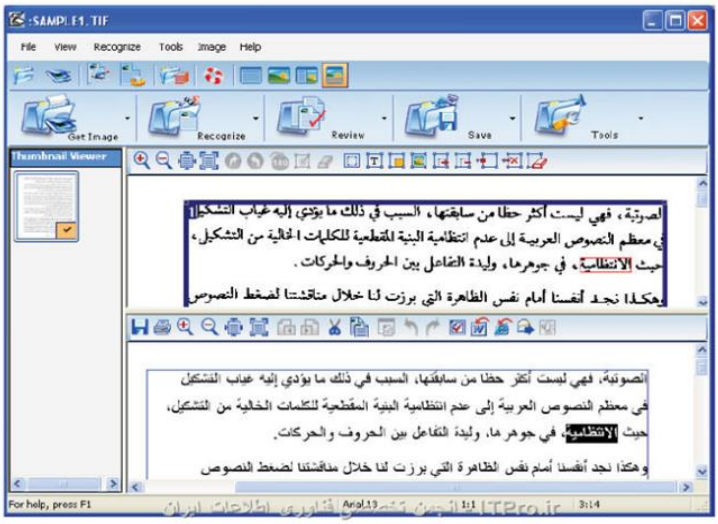
Engineer Sediq, director of Paya Corporation, said: "The discrete Persian OCR of the same kind was a dream a few years ago, but we saw that it has been realized and will progress even more. Therefore, although the design of Persian OCR will take years, it will surely come to the fore. The project is a research project that is being pursued at universities and is still not available to the general public on a large-scale industrial application. But now there are university and laboratory samples inside the country and is the subject of doctor's dissertation of some students. "

So, in the next 10 to 20 years, according to Paya's officials, the Persian handwritten OCR will also enter the market.

In response to the question of whether "the design of Persian consecutive OCR will be realized one day," Dr. Fili replies: "Yes, but gradually." However, according to a contract with "Iranian" companies, up to less than a month, the Persian OCR will be provided with a discrete and continuous manual. "The discontinued OCR project is in its final stages, but it has problems in identifying types of scanners and fonts that we are fixing," says the director of New Age company. The software is currently able to accurately detect Persian type fonts, but the serious problem is that it results in inappropriate scanners." He also commented on the importance of plan in national dimension: "Given that OCR problem has been virtually solved for many languages in the

[2] Iranian Information Technology Society

world, such as English, if final outcome is reached in our country, the increase in level of Persian information in digital world today (including on the Internet) will be very important.



- **Documents Image Analysis**

Document Image Analysis, also called DIA, refers to algorithms and techniques that are applied to document images to provide a computer-readable description of pixel data. The purpose of Document Image Analysis is to identify graphic components and the text in image of document and extract the information it wants as it seeks. A well-known product of image analysis is the optical character recognition software (OCR). Optical recognition of letters allows the user to edit or search the contents of the documents.

Document Image Analysis can be divided into two categories: First, text processing deals with text component of image. Some applications include: Detecting and debugging, finding columns, paragraphs, text lines, words, and finally identifying text (and possibly some text features like size, font, etc.) using OCR. The second category is graphic processing that deals with symbols and non-textual lines, which include linear diagrams, direct lines between text segments and company logos. The images constitute the third most important component of document, but, apart from detecting their location on page, more analysis is required of machine vision visualization and image processing. After using these graphical and text analysis techniques, several megabytes of original data have been collected to provide a brief semantic description of document (Abadim and Ghadiani, 2011).

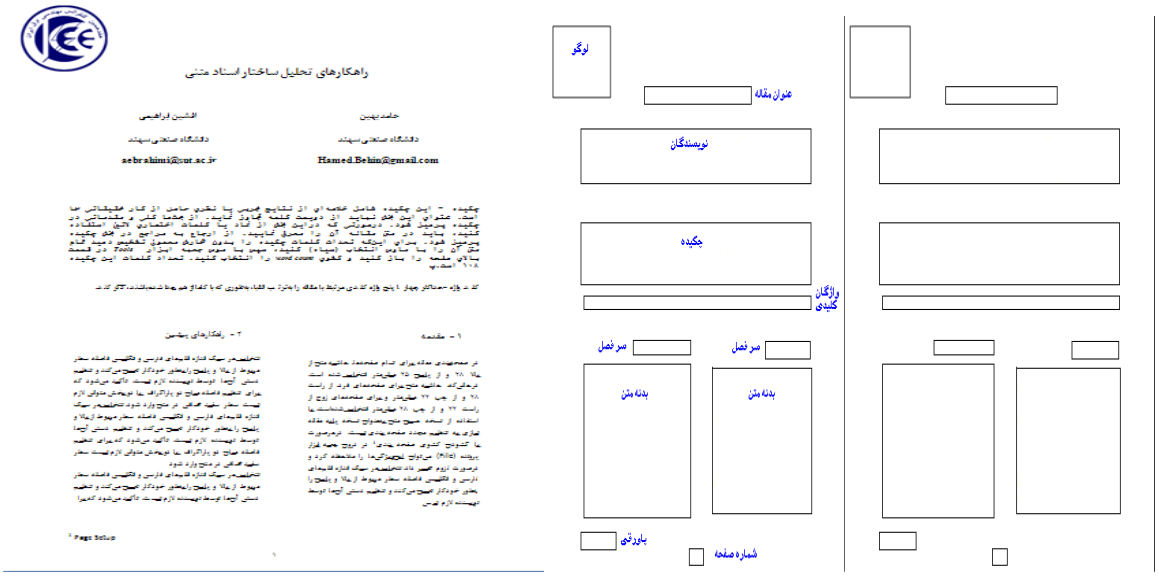
- **Documents Structure Analysis**

Usually after recognition, the image is returned to the zero angle and then structure analysis is performed. Document Analysis Structure is divided into two categories:

Physical structure analysis

Logical structure analysis

Physical structure analysis is a physical segmentation that divides the document's components into groups. Depending on formatting of document, segmentation is performed to separate words, lines of text and structural blocks, such as separate paragraphs or lists of titles.



The main image, together with the results of its physical and logical structure, shows physical structure of blocks obtained from image segmentation. The formatting knowledge of a particular journal helps to label blocks in a logical structure (Ebrahimi and Behin, 2010).

What is OMR?

OMR, which stands for Optical Mark Reader / Recognition, means optic marking is used, and is mainly used in multi-choice tests. In this way, certain forms are designed and tester brands fill (usually oval or rectangle) with soft fonts. To read these forms automatically, one can use branded devices or use brand marking software. OMR Scanner devices are a special scanner that can detect position of brands with the help of light shading on various rows of exam sheet. These devices have threshold values for finding items that have multiple filled in options or no filled in options. OMR software actually simulates the work of devices; thus, they receive scanned image and, by processing it, find the position of different brands. Due to ease of processing information, this technology is still used in many tests of the day and in Iran, there are many companies working in this field.

What is ICR?

ICR stands for Intelligent Character Recognition. This term, which is most often embedded in business sector, is a branch of text recognition that specifically addresses recognition of handwritten letters and figures. In this way, user's requested information is entered on paper forms and in separate boxes (for example, to write name of M H M D letters in four separate square houses), ICR technology is used for automated data acquisition of this information.

In this method, after scanning the forms, position of different fields (such as name, family, ID number, etc.) is determined by image processing, and each field is analyzed to identify letters and digits in it. After finding position of each character, characters are detected using artificial intelligence techniques (such as SVM, neural network, or structural methods).

Conclusion

Many documents, such as newspapers, magazines and brochures, have a very complex structure due to layout of shapes, captions and subheadings of shapes, complex backgrounds, artistic text formatting, font size variations, and line spacing. To recognize the content of a document, humans use a set of evidence, such as general content of document, a series of contracts, as well as information about language, along with a complex argumentation process. Automatic analysis of an arbitrary document with a complex work structure is far more complicated and beyond the ability of today's document structure analysis systems. On the other hand, text and sub-words are graphical components that, according to a convention that exists in each language, have a different meaning for us. It distinguishes from other graphic components. Therefore, complete separation of text and graphics without understanding their true meaning is not possible, although it may be possible with techniques to approach the goal: texts have strong edges, short and continuous, in different directions. Texts have edges in different directions, and the contrast of text area is greater than the contrast of a non-textual area relative to the background, because the text should be readable by the human eye. A text area usually consists of a series of connected components that are arranged horizontally and vertically according to the type of language. To identify a text in a paragraph, regular lines and almost constant lines can be used. For automated reading systems, quick, flexible, and precise segmentation techniques are vital. Segmentation accuracy is a key factor in improving the efficiency of automatic reading systems.

Reference

1. Abadim M., Ghadiani M. (2011). "Providing an Approach Based on Natural Language Processing Processes for Faithfulness in Persian News", Tarbiat Modares University, Faculty of Engineering, Industrial Engineering.
2. Ebrahimi A., Behin H. (2010). "Structure Analysis of Text for Extraction", Sahand University of Technology, Faculty of Electrical Engineering.
3. Feizi M. R., Faroozi F. (2014). "Presentation of a Method for Labeling Components of Word for Persian Language" Faculty of Electrical and Computer Engineering, Software Engineering Department.
4. Hosseinpour Gh. (2010)."Analysis of Textual Image Layouts Based on Classification of Areas in a Hierarchical Decision Making Structure", Quarterly Journal of Electrical Engineering, First Year, No. 3
5. Qasemzadeh M., Izadi S. (2014). "Using Natural Language Processing Techniques to Match Questions in Persian Q & A Systems", Yazd University, Electrical and Computer Faculty, Computer Engineering Department.