**Third Int. Workshop on GNs, Sofia, 1 Oct. 2002, 7-13**

# An Intelligent System For Machine Reading Of Trade Documents

**Valentina Veselinova Radeva**
**Centre for Biomedical Engineering - Bulgarian Academy of Sciences**
**Acad. G. Bonchev str., Bl. 105, Sofia-1113, Bulgaria**

**Abstract:** The paper describes the process of machine reading of trade documents and extracting structured data. A generalized net is used to model an existing implementation the system.

**Keywords:** Abstract systems, Abstract Systems with Properties (ASP), Character recognition, Document reading, Generalized Nets (GN)

## 1 Introduction

Research in machine reading in the past decades was primarilly focused on character recognition. Still the character recognition is just the first step of the process of reading. While the complete task of machine reading is very complex imposing some kind of restriction e.g. on the type of the input documents and the data structures which need to be read, makes it feasible. In this paper an intelligent system capable of reading a certain class of trade documents is described. The system had been developer and is being constantly refined in response to customer requirement at several deployment sites.

## 2 Document reading and extracting structured data

1) Character and word recognition

On reception of a new graphical document for recognition and extraction of structured data, the first task is to recognize the components of which it is composed and to extract its characters. A detailed description of the related subtasks is given in [4] in terms of Generalized Nets (GN; see [1]). The process of character recognition is not an object of detailed study in this paper. The availability of a mechanism for character recognition of the document is assumed, which is able to output for each character its numeric representation (ASCII or UNICODE), the document page number where the character has been found, the position and extent of the bounding rectangle of the character relative to the upper left corner of the page (called coordinates for short later in the text), as well as a recognition confidence measure. Such methods and mechanisms are described in terms of GN in [4] for printed text and in [2] and [3] for handwritten text.

The transition $z_1$ represents the document character recognition process. The token $\alpha_{img}$, representing the input document in graphical format $I$ with properties $P(I)$ as follows:

Input image $I$ with properties $P(I)$:

– resolution;

– file format;

– color depth.

For the next stages of automated recogition and data extraction are needed also the configuration of the expert system for document recognition and structured data extraction,

represented by token $\kappa$ in input place $k_1$ or $k_3$, the configuration of the expert system for extracted data validation, represented by token $\epsilon$ in input place $e_1$ or $e_3$ and the data base for extracted data validation, represented by token $\delta$ in place $d_1$ or $d_3$.

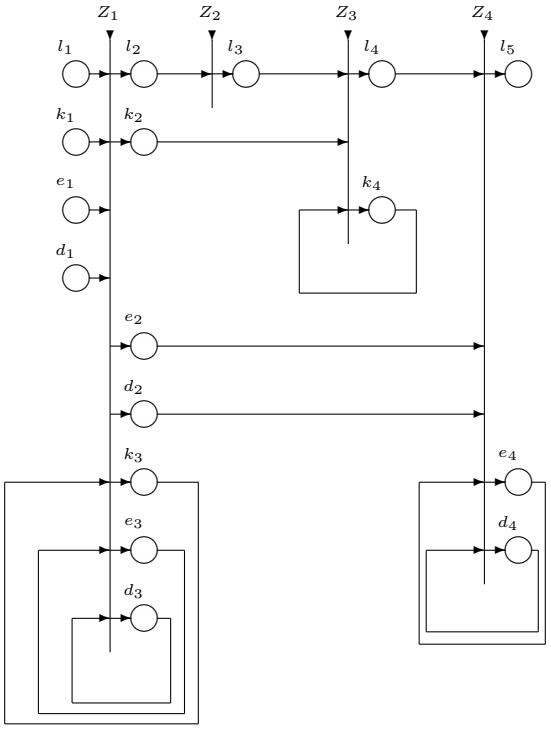


Fig 1. A GN model of intelligent system for document reading and extracting structured data

The transition is activated if a document is waiting for processing in an input place and when the above requirements are fulfilled. A token with characteristic "recognized characters $L$ with properties $P(L)$" enters in place $l_2$.

Recognized character $L$ with properties $P(L)$:

– character code (ASCII or UNICODE);

– position in the document and size of the bounding rectangle - coordinates.

The formal representation of the transition $z_1$ is:

$$z_1 = \langle \{l_1, k_1, e_1, d_1, k_3, e_3, d_3\},$$
$$\{l_2, k_2, e_2, d_2, k_3, e_3, d_3\},$$
$$R_{z_1},$$
$$(l_1 \wedge (d_1 \vee d_3) \wedge (k_1 \vee k_3) \wedge (e_1 \vee e_3))\rangle$$

where:

$$R_{z_1} = \begin{array}{c|ccccccc} & l_2 & k_2 & e_2 & d_2 & k_3 & e_3 & d_3 \\ \hline l_1 & w_{l_1,l_2} & w_{l_1,k_2} & w_{l_1,e_2} & w_{l_1,d_2} & w_{l_1,k_3} & w_{l_1,e_3} & w_{l_1,d_3} \\ k_1 & w_{k_1,l_2} & w_{k_1,k_2} & w_{k_1,e_2} & w_{k_1,d_2} & w_{k_1,k_3} & w_{k_1,e_3} & w_{k_1,d_3} \\ e_1 & w_{e_1,l_2} & w_{e_1,k_2} & w_{e_1,e_2} & w_{e_1,d_2} & w_{e_1,k_3} & w_{e_1,e_3} & w_{e_1,d_3} \\ d_1 & w_{d_1,l_2} & w_{d_1,k_2} & w_{d_1,e_2} & w_{d_1,d_2} & w_{d_1,k_3} & w_{d_1,e_3} & w_{d_1,d_3} \\ k_3 & w_{k_3,l_2} & w_{k_3,k_2} & w_{k_3,e_2} & w_{k_3,d_2} & w_{k_3,k_3} & w_{k_3,e_3} & w_{k_3,d_3} \\ e_3 & w_{e_3,l_2} & w_{e_3,k_2} & w_{e_3,e_2} & w_{e_3,d_2} & w_{e_3,e_3} & w_{e_3,e_3} & w_{e_3,d_3} \\ d_3 & w_{d_3,l_2} & w_{d_3,k_2} & w_{d_3,e_2} & w_{d_3,d_2} & w_{d_3,d_3} & w_{d_3,d_3} & w_{d_3,d_3} \end{array}.$$

The predicates of the index matrix $R_{z_1}$ are:
$w_{l_1,l_2} = true$;
$w_{l_1,k_2} = w_{l_1,e_2} = w_{l_1,d_2} = w_{l_1,k_3} = w_{l_1,e_3} = w_{l_1,d_3} = false$;
$w_{k_1,l_2} = false$;
$w_{k_1,k_2}$ = an updated configuration of the document recognition and structured data extraction exists;
$w_{k_1,e_2} = w_{k_1,d_2} = false$;
$w_{k_1,k_3} = w_{k_1,k_2}$;
$w_{k_1,e_3} = w_{k_1,d_3} = false$;
$w_{e_1,l_2} = w_{e_1,k_2} = false$;
$w_{e_1,e_2}$ = an updated configuration of the data validation expert system exists;
$w_{e_1,d_2} = w_{e_1,k_3} = false$;
$w_{e_1,e_3} = w_{e_1,e_2}$;
$w_{e_1,d_3} = false$;
$w_{d_1,l_2} = w_{d_1,k_2} = w_{d_1,e_2} = false$;
$w_{d_1,d_2}$ = updated validation database exists;
$w_{d_1,k_3} = w_{d_1,e_3} = false$;
$w_{d_1,d_3} = w_{d_1,d_2}$;
$w_{k_3,l_2} = false$;
$w_{k_3,k_2}$ = actual confuguration of the document recognition and structured data extraction expert system exists;
$w_{k_3,e_2} = w_{k_3,d_2} = false$;
$w_{k_3,k_3} = w_{k_3,k_2}$;
$w_{k_3,e_3} = w_{k_3,d_3} = false$;
$w_{e_3,l_2} = w_{e_3,k_2} = false$;
$w_{e_3,e_2}$ = actual configuration of the data validation expert system exists;
$w_{e_3,d_2} = w_{e_3,k_3} = false$;
$w_{e_3,e_3} = w_{e_3,e_2}$;
$w_{e_3,d_3} = false$;
$w_{d_3,l_2} = w_{d_3,k_2} = w_{d_3,e_2} = false$;
$w_{d_3,d_2}$ = actual database for validation exists;
$w_{d_3,k_3} = w_{d_3,e_3} = false$;
$w_{d_3,d_3} = w_{d_3,d_2}$. When the characters included in the document are recognized, they are joined into strings. Specific properties are assigned to each string $W$.

Joined strings $W$ with properties $P(W)$:
– a character sequence ($ASCII/UNICODE$);
– min coordinates;
– max coordinates;

– recognition confidence measure.

The joining of characters is performed by geometrical proximity criteria. The min coordinates are calculated by taking the maximum value of the upper bounds as an upper bound and the minimum value of the lower bounds as a lower bound. The max coordinates are calculated by taking the minimum value of the upper bounds as an upper bound and the maximum value of the lower bounds as a lower bound. Characters with unproportionally large or small bounds are discarded during this process. The recongition confidence measure of the string is calculated by averaging the confidence measures of its characters.

After the characters are joined into strings, a search for known words is performed on them. This is done using a predefined dictionary. This search detects words which match some of the strings exactly or as substrings; in the latter case, the string is split into several parts, one of which is the dictionary word. The matching criterion is based on the Levenstein string distance metric. A input string is replaced by a dictionary word if their Levenstein distance is below a predefined threshold. The properties of the string are modified appropriately - the confidence measure is recalculated, taking into account the replacement of a substring with a dictionary word. When a string is split, the bounding rectangles are adjusted accordingly.

The formal representation of the transition $z_2$ is:

$$z_2 = \langle \{l_2\}, \{l_3\}, R_{z_2}, (l_2) \rangle$$

where:

$$R_{z_2} = \begin{array}{c|c} & l_3 \\ \hline l_2 & w_{l_2,l_3} \end{array}.$$

The predicates of the index matrix $R_{z_2}$ are:
$w_{l_2,l_3} = true$.

Using the recognized words as input, the expert system recognizes the type of the document and extracts the structured data. For each document $D^{'}$ the following properties are determined $P(D^{'})$.

Recognized document $D^{'}$ with properties $P(D^{'})$:

– document type;

– data type;

– data values;

– position - coordinates of data in document;

– data recognition confidence measure.

2) Document type recognition and reading

The process has two phases: document type recognition and extraction of structured data based on the data layout knowledge associated with the document type.

For the document type recognition a system for calculating the correspondence between the input document and the predefined types stored in the knowledge base has been developed.

The knowledge base contains document type descriptions, rules for telling them apart and for extracting data from them. The descriptions are coded in a specially designed language stored in XML files. This language can describe special data fields on title, middle or last pages of the document, as well as any kind of tables with any kind or number of columns. It is also possible to describe fields which are encountered only once per document, or many

times on the title page, last page or any page, as well as one or many tables with different number of columns. A record in a table can span one or several lines of text.

The document data consists of strings, composed according to the predefined extraction rules, specific to the document type. Their properties are the strings themselves, their data types, defined in the document type description, their positions and the recognition confidence measure.

Tables are processed using a robust line separation algorithm, which is not affected by some kinds of document image distortions, e.g. skew. The records in the tables are automatically detected and data values are extracted. A variable number of lines per record within the same table is also supported. The table processing algorithm is able to detect the end of each table, which makes possible the processing of documents with variable number of records from a single document type description.

After the most appropriate document type is selected, the system creates an instance of the document type description, customized to the current input document and its parameters. It is used by an expert system for structured data extraction using the methods described above.

The formal representation of the transition $z_3$ is:

$$z_3 = \langle \{l_3, k_2, k_4\}, \{l_4, k_4\}, R_{z_3}, (l_3 \wedge (k_2 \vee k_4)) \rangle$$

$$R_{z_3} = \begin{array}{c|cc} & l_4 & k_4 \\ \hline l_3 & w_{l_3,l_4} & w_{l_3,k_4} \\ k_2 & w_{k_2,l_4} & w_{k_2,k_4} \\ k_4 & w_{k_4,l_4} & w_{k_4,k_4} \end{array} .$$

The predicates of the index matrix $R_{z_3}$ are:
$w_{l_3,l_4} = true.$ $w_{l_3,k_4} = false.$ $w_{k_2,l_4} = false.$ $w_{k_2,k_4}$ = an updated configuration of the document recognition and structured data extraction expert system exists;
$w_{k_4,l_4} = false.$ $w_{k_4,k_4}$ = an actual configuration of the document recognition and structured data extraction expert system exists.

After the structured data is extracted, it is validated using a customer-supplied database by a second expert system. For each validated document $D''$ the following properties $P(D'')$:

Verified document $D''$ with properties $P(D'')$:

– information about the verification method which leads to the closest match between the document data and the customer database;

– document type;

– data type;

– data value;

– position - data coordinates;

– recognition confidence measure.

After the data coordinates and types are determined, a second character recognition is performed with a restricted character set (e.g. digits only for numeric data types). Usually this second recognition leads to significantly better results than the first one, which was performed without any knowledge about the data. Additionally, characters are coerced into the appropriate type with equivalent replacements. The recognition confidence measure is lowered when such replacements are made.

The validation works only with user-specified fields of specific interest, which are represented in a database. The process is performed by a second expert system. The knowledge base for this expert system consists of the correspondence between the input document fields and the customer database record fields. The description is coded in a custom language, stored in XML files. The methods used are part of the expert system. They are used to find the best match from the database for the recognized document record. The fields used in the comparison are assigned priorities to resolve possible ambiguities. When a perfect match is found, the confidence measure for the record is set to the maximum possible value. If the match is not perfect, but within the preset Levenstein distance, the input record is replaced by the user database record and the confidence measure is set to the distance between them. Finally, if the system cannot choose a candidate meeting the minimum confidence criterion, the input record is not modified, but its recognition confidence measure is lowered. Thus, three classes of records are output - valid, probable candidates and unknown. In the user interface portion of the system, those classes are color-coded to attract the user attention to problematic data.

The formal representation of the transition $z_4$ is:

$$z_4 = \langle \{l_4, e_2, d_2, e_4, d_4\}, \{l_5, e_4, d_4\}, R_{z_4}, (l_4 \wedge (e_2 \vee e_4) \wedge (d_2 \vee d_4)) \rangle$$

where:

$$R_{z_4} = \begin{array}{c|ccc} & l_5 & l_4 & d_4 \\ \hline l_4 & w_{l_4,l_5} & w_{l_4,l_4} & w_{l_4,d_4} \\ e_2 & w_{e_2,l_5} & w_{e_2,l_4} & w_{e_2,d_4} \\ d_2 & w_{d_2,l_5} & w_{d_2,l_4} & w_{d_2,d_4} \\ e_4 & w_{e_4,l_5} & w_{e_4,l_4} & w_{e_4,d_4} \\ d_4 & w_{d_4,l_5} & w_{d_4,l_4} & w_{d_4,d_4} \end{array}.$$

The predicates of the index matrix $R_{z_4}$ are:
$w_{l_4,l_5} = true$;
$w_{l_4,e_4} = w_{l_4,d_4} = false$;
$w_{e_2,l_5} = false$;
$w_{e_2,e_4}$ = an updated configuration of the data validation expert system exists;
$w_{e_2,d_4} = false$;
$w_{d_2,l_5} = w_{d_2,e_4} = false$;
$w_{d_2,d_4}$ = an update database for validation exists;
$w_{e_4,l_5} = false$;
$w_{e_4,e_4}$ = an actual configuration of the data validation expert system exists;
$w_{e_4,d_4} = false$;
$w_{d_4,l_5} = w_{d_4,e_4} = false$;
$w_{d_4,d_4}$ = an actual database for validation exists.

3) The system as ASP In terms of ASP the process, represented with this GN, can be described as follows:

Input object $X$:
•Input image $I$ with properties $P(I)$:
– resolution;
– file format;
– color depth.
Output object $Y$:
• Verified document $D''$ with properties $P(D'')$:

– information about the verification method which led to the closest match between the document data and the customer database;

– document type;

– data type;

– data value;

– position - data coordinates;

– recognition confidence measure.

Global system state object $C$:

•Database $DB$ with properties $P(DB)$:

– database availability;

– database tables availability;

– database table columns availability;

– database size;

– database indexing;

•Document type recognition and structured data extraction expert system configuration $KnB$ with properties $P(KnB)$:

– knowledge base availability;

– knowledge base validity;

– knowledge coherence;

– knowledge base size.

•Data validation expert system configuration $ESC$ with properties $P(ESC)$:

– information validity;

– knowledge coherence;

– configuration parameters.

# 3 Conclusion

Each of these subsystems can also be represented with a GN model and/or as a separate abstract system with properties by describing its input, output and internal parameters and their properties. This could be an object of future research.

# References


[1] K. Atanassov, Generalized Nets. Singapore, New Jersey, London, World Scientific, 1991.

[2] G. Gluhchev, K. Atanassov, S. Hadjitodorov. Handwriting analysis via generalized nets. Proceedings of the international Scientific Conference on Energy and Information Systems and Technologies. Vol. III, Bitola, June 7-8, 2001, 758-763.

[3] A. Shannon, G. Gluhchev, K. Atanassov, S. Hadjitodorov. Generalized net representing process of handwriting identification. Cybernetics and Information Technologies, Vol. 1, 2001, Sofia, 71-80.

[4] G. Gluhchev, K. Atanassov, S. Hadjitodorov. Automatic document processing. Proceedings of the Second International Workshop on Intuitionistic Fuzzy Sets and Generalized nets, Warszawa, 25-26 July 2002 (in press).