Morphological Tagging Approach in Document Analysis of Invoices

Y. Belaïd and A. Belaïd

LORIA -University Nancy 2
54506 Vandoeuvre-Lès-Nancy
France
{ybelaid,abelaid}@loria.fr

Abstract

In this paper a morphological tagging approach for document image invoice analysis is described. Tokens close by their morphology and confirmed in their location within different similar contexts make apparent some parts of speech representative of the structure elements. This bottom up approach avoids the use of an priori knowledge provided that there are redundant and frequent contexts in the text. The approach is applied on the invoice body text roughly recognized by OCR and automatically segmented. The method makes possible the detection of the invoice articles and their different fields. The regularity of the article composition and its redundancy in the invoice is a good help for its structure. The recognition rate of 276 invoices and 1704 articles, is over than 91.02% for articles and 92.56% for fields.

1. Introduction

Nowadays, the use of forms as common documents for administrative communications and exchanges can lead to their rapid accumulation in the administrations and offices. This is why several document engineering companies launched out in the automatic treatment of forms. However, the lack of generality in the description of form structure led only to very simplistic solutions and all the challenge remains entire for the more complex documents [1, 2, 3, 4, 5].

In this paper, we will describe a generic approach for invoice document processing and structure. We limited the structure to invoice bodies already extracted. The invoice body text is given in ASCII text roughly structured and recognized by OCR. It is question to delimit the different articles composing the invoice body and to extract within each article its different fields such as “designation”, “code”, “unitary price”, “amount”, etc.

Error! Reference source not found. shows an invoice example where the body is framed.

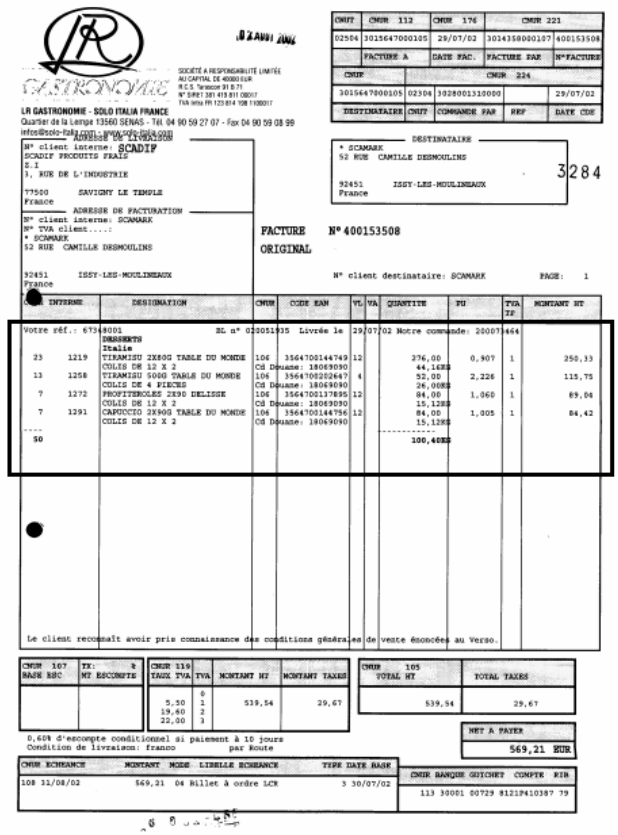

Figure 1. Invoice example

2. Tagging Approach

This approach is based on the concept of part-of-speech tagging (PoS). PoS assumes that the text is written in a natural language with a real syntactic structure. It tries to assign the morphological word class annotation to each word in the text by considering its lexical meanings and syntactic context. The grouping of this tags leads to the extraction of nominal syntagms revealing the real language structure. In the literature, several methodologies

Proceedings of the 17th International Conference on Pattern Recognition (ICPR’04)
1051-4651/04 $ 20.00 IEEE

have been developed for the task: Hidden Markov Models [6], transformation based learning [7], memory based learning [8], maximum entropy[9], etc. The main research investigated is concentrated either on the resolution of the word ambiguity belonging to different classes, or on the imagination of the sense of unknown words.

In the case of invoices composed of successive articles structured in consecutive fields, the article text is written in a non natural language not allowing the direct use of a syntactic approach. Hence, we have adapted the PoS tagging to the particular structure of invoice articles. We conserved the primary tagging by only using the morphology, and the secondary (contextual) tagging is obtained by the use of regular expressions reinforced by some redundancy and regularity factors observed on their occurrence in the text. This approach should overcome the current problems encountered in invoice recognition such as: noise perturbation (stamp, handwriting notes, etc.), OCR errors, structure heterogeneity and structure variation depending on the invoice provider. Furthermore, the approach have to be enough generic to adapt on each new invoice performed.

The system is composed of three main steps :

- Primary tagging and tabular structure extraction;
- Contextual tagging and article delimitation;
- Model generation and syntactic correction.

3. Primary tagging and tabular structure extraction

For each invoice body, an ASCII file is produced containing all the words recognized by the OCR. Each word is accompanied by its upper left corner co-ordinates (within the image of the invoice).

Thanks to the primary tagging, a label stemmed from a specific table (see Table 1) is assigned to each token (word). Then tokens are structured in lines and columns. The column structure is obtained by line projection methods applied only on lines containing a real number (this usually corresponds to a price). This choice allows us to eliminate lines not related to articles (reminder of the order number or the number of the delivery order ...) (see Figure 2). At each cell and each column is assigned the most frequent label (see Figure 3).

Some corrections are performed in the cells and fusions are operated between columns in order to regroup some numerical tokens (integer or real) cut due to the presence of spaces separating the thousands in the amounts.

Table 1. Main primary tags

Label	Meaning	Particularities
AB	Alphabetic	
AN	Alphanumeric	
AV	Token « avoir »	« avoir »
BL	Token « bon de livraison »	« bl », « b.l », « bl : », « bon », « b.e », « be », « livraison »
CM	Token « commande »	« commande », « commandes », « cde », « cmd »
CO	Code	Contains « - » / « / »
GC	Gencode	Begins by « 356 » in France
NE	Integer	
NO	Token « numéro »	« no », « no. », « num », « num. », « n° »
NP	Percentage	Ends by « % »
NR	Real	
NU	Numeric	
RF	Token « référence »	« référence », « référence », « référence », « réf », « ref », « réf. », « ref. »
TT	Token « total »	« total », « tot »
UT	Unity	After NU

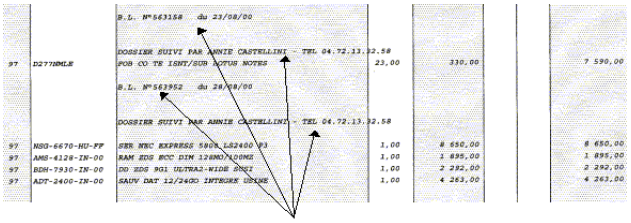

Figure 2: Lines discarded to favour the column extraction

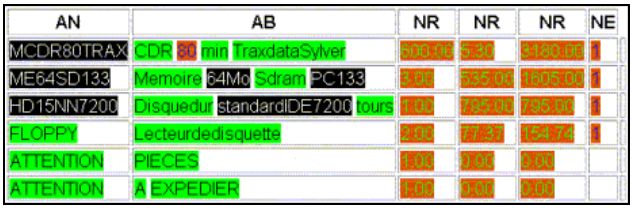

Figure 3: Primary tagging and tabular structure extraction

4. Contextual tagging and article delimitation

The article zone is split into blocks composed of successive lines. Two blocks are separated by at less one white line. Then, each line is processed individually by looking for some features such as: "quantity", "unitary price" and "amount". This is done by looking for numerical terms (integer and real) regularly repeated at the same position in similar column cells. One found, such lines are called main lines.

Once located, the main lines position in the block will contribute to delimit the articles. We have defined several cases:

- one article per line,
- one article per block,
- one article on many lines and the amount is located on the first line,
- one article on many lines and the amount is located on the last line,
- one article and the amount is located either on the first or the last line,
- others cases.

These cases are found by studying the position regularity of the main lines within the blocks. Then, the article category is determined for all the articles according to the most frequent cases. Figure 4 shows a segmentation example (the black strokes separate the different articles extracted).

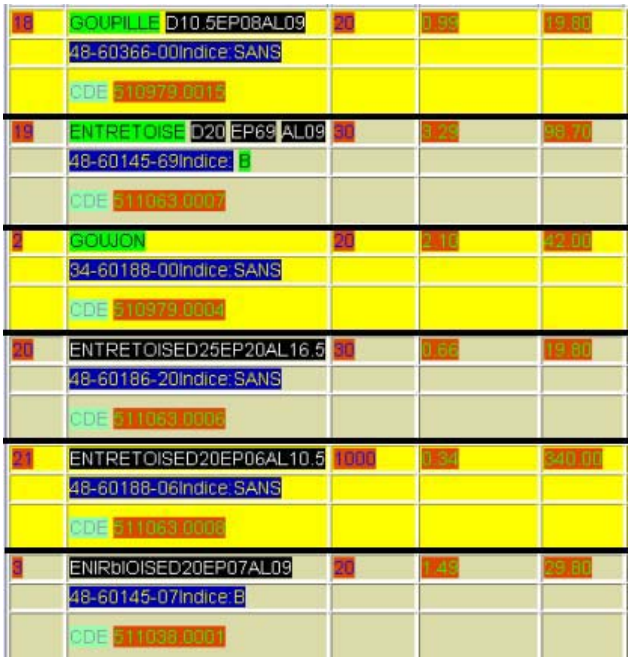

Figure 4 : Article segmentation

Then, we look for the structure of each article considering the contextual labeling operated in the main line. The "Designation" location is made in the cell containing the greatest number of words and containing a majority of alphabetic or alphanumeric words. This location is extended to the close cells verifying these criteria. The international « GENCOD » sometimes cut out in several cells is detected thanks to its particular form: it is composed of 13 digits and generally starts with the number 356 in France. The article "code" is required in the cells located before the labeled cell "Designation". Either there is a cell having the label "CO" or one retains that containing only integers or alphanumeric. The "Quantity", the "Unitary price" and the "Amount" are required in the cells containing a real or an integer. Their precise localization is validated by calculation (unit amount * price = amount). Lastly, it remains to possibly find the price unit before the discount. They are searched in a column on the left unit price and are also checked by calculation. In this labeling step, the regularity is taken into account by the presence of the columns and their labels.

5. Model generation and syntactic correction

A general model is determined from the models of each article by retaining that which is most frequent (see Figure 5): the limits between the various columns are visualized by vertical black strokes and the headings appear in the top of each column). This model indicates the headings of each column and their appearance order. The model is used to carry out corrections in the articles which do not respect this syntactic description. For example, if an OCR error led to a word labeled as alphanumeric in the column quantity, then this word is corrected knowing that the quantity multiplied by the unit price must give the amount.

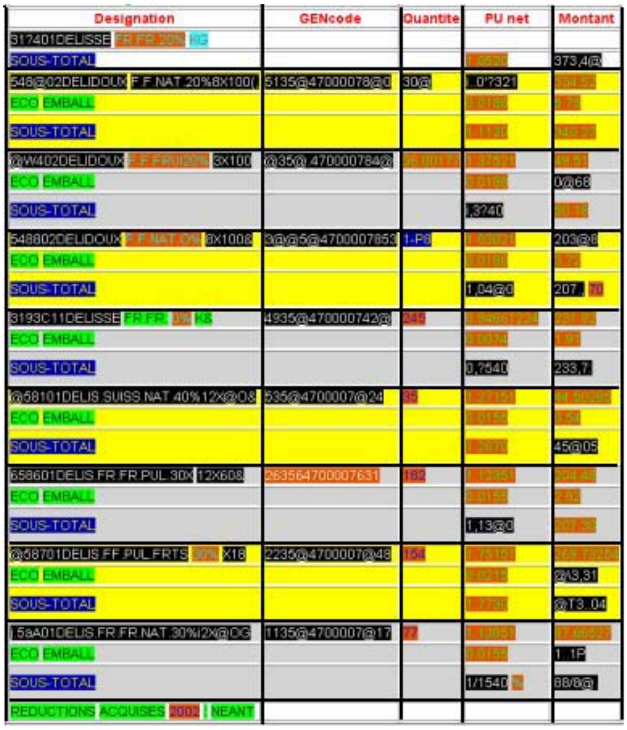

Figure 5 : Invoice model

6. Results and Discussion

Experiments were made on 276 invoices, corresponding to 1704 articles. The percentage of the articles recognized is equal to 91.02%, that of the

headings is equal to 92.56%. Table 2 gives the detail of recognized information.

Table 2. Experimental results

Headings	Recognized	Present	Percentage
Code	**180**	**185**	**97.30%**
Designation	**262**	**274**	**95.62%**
Quantity	**250**	**273**	**91.57%**
UP - discount	**67**	**75**	**89.33%**
Discount	**31**	**39**	**79.49%**
Unitary Price	**250**	**274**	**91.24%**
Gencode	**52**	**59**	**88.13%**
Amount	**253**	**274**	**92.34%**

Segmentation or construction errors of the model are related to:

- OCR: 1) absence of integer or real in the main line of an article; 2) under-segmentation,
- presence of unwanted lines: they are sometimes associated to the article which precedes or which follows (Figure 6: the first article was detected on 2 lines instead of only one; the number of the delivery order was regarded as the continuation of the designation),
- a too weak redundancy (not enough of articles) to be able to compensate the OCR quality,
- the absence of a structure in the column shape (see **Error! Reference source not found.**),
- a bad localization of the article area (absence of the first or last lines or part of the last column).

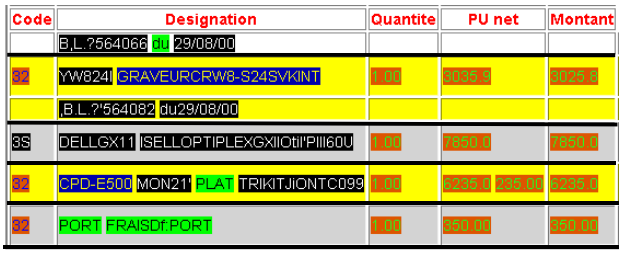

Figure 6. Segmentation errors

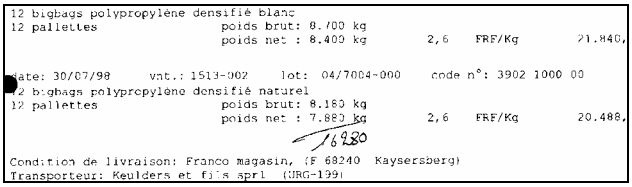

Figure 7. Invoice without column

7. References

[1] A. Belaïd, Y. Belaïd, N. Valverde, and S. Kébairi, "Adaptive Technology for Mail-Order Form Segmentation", In *International Conference on Document Analysis and Recognition*, Seattle (USA), 2001, pp 689-693.

[2] H. Sako, M. Seki, N. Furukawa, H. Ikeda, A. Imaizumi, "Form Reading based on Form-type Identification and Form-data Recognition", In *International Conference on Document Analysis and Recognition*, Edinburgh (Scotland) , 2003, pp 926-930.

[3] H. E. Nielson, W. A. Barrett, "Consensus-Based Table Form Recognition ",In *International Conference on Document Analysis and Recognition*, Edinburgh (Scotland) , 2003, pp 906-910.

[4] A. Simon, J. Pret, and A. Johnson, " A fast algorithm for bottom-up document layout analysis", *PAMI*, Vol. 19, No. 3, March 1997, pp. 273-277.

[5] Y. Belaïd, J. L. Panchèvre, and A. Belaïd, "Form Analysis by Neural Classification of Cells", In *International Workshop on Document Analysis Systems*, Nagano (Japan), 1998, pp 69-78.

[6] T. Brants, " TnT - A Statistical Part-of-speech Tagger", In *Language Technology Joint Conference Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics,* Morgan Kaufman Publishers, Seattle, Washington , 2000.

[7] E. Brill, "A Simple Rule-based Part-of-speech Tagger", In *Third ACL Conference on Applied Natural Language Processing*, Trento, Italy , 1992, pp.152-155.

[8] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis, "MBT: A Memory-based Part-of-speech Tagger Generator", In *Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, 1996, pp.14-27.

[9] A. Ratnaparkhi, "A Maximum Entropy Part-of-speech Tagger", In *Conference on Empirical Methods in Natural Language Processing,* Philadelphia, 1996.

Acknowledgment :

We are grateful to ITESOFT company for the database and for stimulating discussions about the project.