# Neural Network-Based Proper Names Extraction in Fax Images

**Noura Azzabou, Laurence Likforman-Sulem**

GET-Ecole Nationale Supérieure des Télécommunications
Signal and Image Processing Laboratory & CNRS LTCI
46 rue Barrault 75013 Paris, France {azzabou, likforman}@tsi.enst.fr

## Abstract

*In this paper, we are interested in the sender's name extraction in fax cover pages through a machine learning scheme. For this purpose, two analysis methods are implemented to work in parallel. The first one is based on image document analysis (OCR recognition, physical block selection), the other on text analysis (word feature extraction, local grammar rules). Our main contribution consisted in introducing a neural network to find an optimal combination of the two approaches. Tests carried on real fax images show that the neural network improves performance compared to an empirical combination function and to each method used separately.*

## 1. Introduction

Messaging systems convey heterogeneous data such as image documents, e-mails and voice and store them in a unified framework. While e-mail messages are provided with complete information (sender's name, subject and date), document and voice messages, are only indexed by date and phone or fax numbers from where the message was sent. Intelligence can be added to such systems by extracting remaining information such as the identity of the sender and message subject. Such information enables a user to visualize the received documents in a suitable form for access, sorting, and indexing. The task of extracting specific items in business documents (invoices, forms, letters) is generally performed through physical decomposition followed by logical labeling. Models are derived from empty documents, or predefined models which rely on numerical assumptions [1-4].

Besides that, extracting proper names on text strings belongs to the named-entity recognition task reported in MUC conferences [5]. Strings are assumed to be error free and to include punctuation. Proper names include a great variety of names ranging from person, place and organization names. Approaches for extracting proper names typically use manually constructed rules to match searched patterns. More recently, HMM approaches enable easier learning [6].

By dealing with fax images, we face several problems. The first one is their low quality: low resolution and character degradation. The resulting text strings obtained from an OCR system are corrupted and lacking punctuation. Deep parsing of these strings would lead to too much errors as phrases are not precisely delimited. Also, relying only on predefined patterns would miss the corrupted ones. When using rule based patterns, one also assumes that header identifiers and their corresponding field follow each other, which is not always the case in fax transcriptions. The second problem is that fax images are characterized by their variability in header position and presentation. Predefined models and numerical assumptions on the position the specific information searched for can hardly be applied on such documents.

We are interested here in extracting the sender name in fax images. The proposed approach consists in extracting the sender name through a machine learning scheme. Each layout component is classified according to one of the following categories: being the sender name or not. The classification uses a neural network and is based on a feature vector combining features extracted from an image and from a textual analysis. The problems mentioned above are addressed by including simple textual features that can hardly be corrupted during fax transmission and by using only local grammar rules. Adding image features help to localize the searched names between all names included in the image and to compensate for textual analysis.

Thus, section 2 deals with the contribution of image analysis and textual analysis separately for sender's name extraction. Section 3 is devoted to the empirical combination of both analyses. In section 4, we are interested in the introduction of neural networks for sender's name extraction and we focus on the influence of word's context on its classification as a sender's name or not. Performances are evaluated on a real world facsimile database.

## 2. Image and Textual Analysis

Two analyses are conducted in parallel, performed on each fax image and its textual transcription respectively. Image analysis begins with the extraction of layout components at a pseudo-word level, using the RLSA algorithm. Layout components extracted may contain a single word or a group of words and a low level pre-classifier discriminates between handwritten and printed ones.

Image analysis mainly consists in pointing at layout components possibly containing proper names. For this purpose, the underlying structure provided by headers can be recovered by searching for keywords such as (*from, to, name, sender*, aso) in the OCR transcription. Searching only for these words may induce errors as such general words do not always correspond to headers. Spatial constraints between hypothesized headers are added in order to select the best ones. Spatial relations between layout components enable to select components possibly including the sender name. These spatial relations and constraints are defined from very general cues such as proximity and alignment. The selection of potential sender name blocks is described in [7].

Textual analysis mainly consists in pointing at strings possibly being proper names. These string components must satisfy internal and/or external clues. Each word of the OCR transcription is first checked by edit distance matching against two dictionaries: a first name dictionary (1200 words) and a language dictionary (200 000 words). As the transcription contains recognition errors, exact matching is not required. A set of rules applied on each word or 2 words group enables to check for name patterns and initials. Then, several word features are computed according to internal and external textual clues:

- internal clues : is the word written in capital letters, does the word begin with a capital letter, is it an initial, does it belong to one of the dictionaries (first name, general) ?
- external clues : is the word near an identity marker (Mr, Mrs, Dr, …), is the word included in a predefined pattern such as (initial + capitalized word or first name + capitalized word) ?

As a result of these analyses, each word or layout component is associated with a set of binary features from which different feature sets will be extracted and provided to classification. Thus, image analysis relies on OCR transcription where to find headers belonging to a keyword dictionary and general visual cues. As mentioned before, because of corrupted text strings, headers maybe missed and the set of hypothesized layout components has to be restricted. On the other hand, textual analysis focuses on proper name detection regardless their nature since it doesn't make the difference between sender, recipient or even names found in the text body or in some addresses. So, we can say that the two analyses provide us complementary information type and this leads us to deal with a combination of the two approaches to improve sender name detection

## 3. Empirical combination

From the analyses presented previously, we extract a set of 5 binary features f1, f2, …f5 for each word. This feature set is composed of:

- **f1** : indicates whether the layout component associated to the word is considered as a potential sender name block according to the image analysis
- **f2** : indicates whether the word can be considered as a proper name and this is done according to some patterns (cf. external clues)
- **f3** : indicates whether the word is printed or handwritten because we assume that being handwritten argues in favour the possibility of being a proper name because in some cases cover pages are printed and users fill in the fields by hand.
- **f4** : indicates whether the word begins with a capital letter or if it is capitalized.
- **f5** : indicates whether the word was not found in a general dictionary and this has a positive correlation with the possibility of being a proper name

Features f1 and f3 (resp. f2, f4, f5) are issued from image analysis (resp. textual analysis). The combination approach consists in assigning each word a score which reflects the possibility of being the sender name. The score is computed as a linear combination of the features. We used empirical weighting coefficients that reflect the robustness of a feature compared to another. For instance, (**f1**) and (**f2**) are considered as the strongest arguments in favour of the assumption that a word is a sender name or not. Thus, we associated to them respectively 3 and 2 as coefficients. For the other features, all the coefficients are set to one. After score computation, we select the sender's name among the words which have the highest scores, in fact the set of words that correspond to the 3 highest scores (top3 choice).

**Table 1. Empirical combination**

| system | recall (%) | precision (%) |
|---|---|---|
| (I) | 53 | 18 |
| (T) | 53 | 34 |
| (I+T) | 75 | 22 |

At a first time we compared the three following systems:
(I) an image based analysis system where we use only (**f1**) and (**f3**) in score computation.
(T) a textual analysis based system where score computation includes (**f2**) (**f4**) and (**f5**)
(I+T) an hybrid system where score computation includes all features.

Tests were carried out on 150 real faxes images, using recall and precision rates. From the results presented in Table 1 we can say that the combination method improves the sender name detection because additional textual information cope with the variance in fax layouts and the spatial features avoid us to consider all proper names as potential candidates.

## 4. Neural network-based combination

In the previous section we presented the combination of the image analysis and the textual analysis where weighting coefficients are set empirically. To optimize these coefficients and thus improve name extraction, we introduced a machine learning scheme based on neural networks.

### 4.1 Perceptron model

Our objective is to detect the sender's name by discriminating between two categories: sender's name (SN) and simple words (SW). As a first approach, we adopted a perceptron model which is a single neuron with a linear weighted net function. The weighting coefficients can be seen as the parameters of the hyper plane that separates (SW) from (SN) so a good classification relies on an efficient choice of the weighting coefficients. A sequential learning online algorithm is then performed to adjust weighting coefficients to the training data. This is based on a the steepest descend gradient algorithm.
To perform learning, we extracted from 30 faxes words with their feature's vectors. To each word we assigned a label: 1 if it is the sender's name class and 0 if not. We notice that weighting coefficients are completely different from those set empirically. For instance **f2** has a very low coefficient compared to the other features and this can be explained by the fact that it is correlated with **f4** and **f5**. Also, we can see that textual characteristics are more important then those based on image analysis.

$$\text{Score}= 0.18f1+0.06f2+ 0.14f3+0.4f4+0.37f5 \quad (1)$$

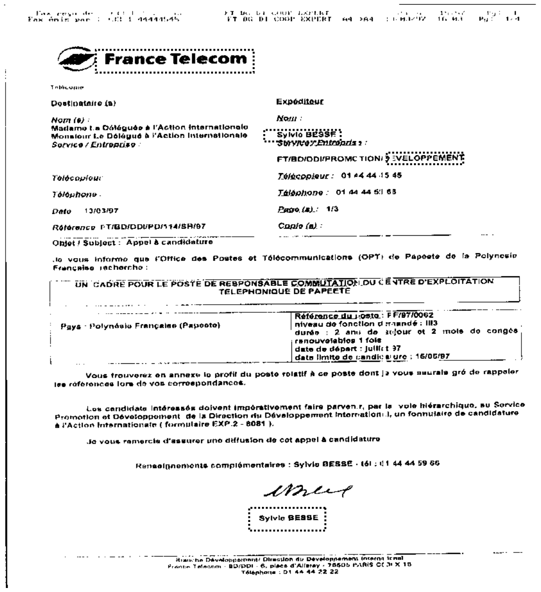
Results in Table 2 show us that a perceptron model outperforms the empirical model in terms of recall and precision. Indeed, better results are achieved with a trained classifier. But in the other hand, a perceptron separates linearly separable data which is not the case in our application. For this reason, we replaced the perceptron by a multi-layer perceptron (MLP).

### 4.2 Multi-layer perceptron model

In this section, we are interested in a more complex architecture of the neural network in order to make non linear separation of the two categories SN and SW. Furthermore, we adopted more basic features which do not use any rule-based patterns. These features are:

- **g1** : indicates if the word is located in a potential sender's name block
- **g2** : indicates if the word is found in a first name dictionary
- **g3** : indicates if the word is an initial
- **g4** : indicates if the word begins with a capital letter
- **g5** : indicates if the word is found in a general dictionary

The MLP we used in our application is composed of, an input layer containing 5 neurons, a hidden layer with 3 neurons and the output layer with 2 neurons which correspond to the probability of belonging to each class. The training step is based on error back-propagation algorithm [8]. Word score is given by the output neuron corresponding to the SN class. Words with a score that goes above a threshold (0.7 in our case) are classified as sender's name.



Fax reçu de : [illegible]    FT BG DI COOP EXPERT
Fax émis par : +33 1 44444545    FT BG DI COOP EXPERT  A4 >A4  13/03/97  16:03  Pg: 1/4

France Telecom

Télécopie

Destinataire (s)

Nom (s) :
Madame La Déléguée à l'Action Internationale
Monsieur Le Délégué à l'Action Internationale
Service / Entreprise :

Télécopieur

Téléphone

Date 13/03/97

Référence FT/BD/DDI/PD/114/SB/97

Expéditeur

Nom :
Sylvie BESSE
Service / Entreprise :
FT/BD/DDI/PROMOTION / DEVELOPPEMENT

Télécopieur : 01 44 44 15 45

Téléphone : 01 44 44 53 68

Page (s) : 1/3

Copie (s) :

Objet / Subject : Appel à candidature

Je vous informe que l'Office des Postes et Télécommunications (OPT) de Papeete de la Polynésie Française recherche :

UN CADRE POUR LE POSTE DE RESPONSABLE COMMUTATION DU CENTRE D'EXPLOITATION TELEPHONIQUE DE PAPEETE

Pays : Polynésie Française (Papeete)

Référence du poste : FF/97/0062
niveau de fonction demandé : III3
durée : 2 ans de séjour et 2 mois de congés renouvelables 1 fois
date de départ : juillet 97
date limite de candidature : 16/05/97

Vous trouverez en annexe le profil du poste relatif à ce poste dont je vous saurais gré de rappeler les références lors de vos correspondances.

Les candidats intéressés doivent impérativement faire parvenir, par la voie hiérarchique, au Service Promotion et Développement de la Direction du Développement International, un formulaire de candidature à l'Action Internationale ( formulaire EXP.2 - 8081 ).

Je vous remercie d'assurer une diffusion de cet appel à candidature

Renseignements complémentaires : Sylvie BESSE - tél : 01 44 44 59 66

Sylvie BESSE

Branche Développement/ Direction du Développement International
France Telecom - BD/DDI - 6, place d'Alleray - 75505 PARIS CEDEX 15
Téléphone : 01 44 44 22 22

**Figure 1.** SN expressions extracted (in dotted rectangles) with a MLP + contextual (M) rule

Results in Table 2, show that the MLP model ameliorates the recall but deteriorates the precision. In fact, recall and precision are two antagonist variables. For example, we can obtain an important recall with a very low precision just by proposing all the words of the fax as sender's names but we obtain a very poor precision. For this reason, we introduced a post processing step, that relies on the image context of the word, to study its impact on the performance of the two proposed models.

## 4.3 Contextual analysis

In the previous sections, we processed each word in isolation but in real cases the sender name is composed of a group of words in the form (**First Name + Last Name**) or (**Initial + Last Name**). This introduces the concept of sender's name group where words are gathered to form an expression. To be gathered as an expression two words w1 and w2 must satisfy the two following conditions:

$|y(w1)-y(w2)| < s_y$ and $|x_{ini}(w2)-x_{end}(w1)| < s_x$

where y is the horizontal position of a word, $x_{ini}$ (resp $x_{end}$) is the vertical position of the first (resp. last) character in the word. Empirical values for $s_x$ and $s_y$ are respectively 100 and 5 pixels. Words that contain some non alphabetical characters are discarded from grouping. Also, expressions whose length is above a threshold are discarded. Once expressions are formed, we assign to them a score according to one of two following rules:

**(M)** Expression's score is equal to the mean of expression word's scores
**(G)** Expression's score is the greatest of expression word's scores

Word scores are those given by the perceptron or the MLP (cf. § 4.1 and 4.2). Only high scored expressions (above 0.7) are considered for being SN expressions. We tested the contextual analysis for the MLP model and the perceptron model. Results in Table 2 show that with the greatest score rule (G), recall is improved at the expense of precision. In fact, some words are presented as members of the sender's name group because they lie in the neighbourhood of a high scored word. So, we recover some sender's name components that have a low score such as initials. On the other hand, the increasing number of candidate words decreases the precision of the algorithm. But retrieving complete strings is qualitatively better than retrieving isolated words.

Using the mean rule (M) increases precision but decreases the recall rate because in some cases the sender's name is located near words that do not have a high score so that the score of the expression containing the right name decreases.

Figure 1 shows sender name expressions extracted using the MLP+(M) rule. The sender name was found as the expression with greatest score (below header *Name*). The sender name also appears in the signature but was retrieved with a lowest score as well as the expression including the sender society name (upper left) and a capitalized word.

**Table 2. Neural network-based combination**

| system | recall (%) | precision (%) |
|---|---|---|
| perceptron | 77 | 25 |
| MLP | 80 | 20 |
| perceptron+ (G) | 84 | 12 |
| perceptron+ (M) | 75 | 21 |
| MLP + (G) | 83 | 16 |
| MLP + (M) | 71 | 25 |

## 5. Conclusion

We have presented a method for sender's name extraction based on the combination of image and textual features. Several feature sets were tested. The best one used textual features not highly dependent on context. As far as the combination rule is concerned, we compared the performance of three systems: an empirical linear combination, a perceptron model and a MLP model. Results obtained confirmed that trained systems perform better. In the other hand, comparison of the perceptron model and the MLP one shows that they are equivalent and the choice of one of the two systems depends on the application and the priority accorded to the recall rate or the precision rate. The contextual analysis which was introduced as a post processing step produces complete sender name expression and improves the recall rate. Finally, recall rate will be improved for documents generated electronically (such as pdf documents), for which error free textual transcriptions are available.

## 6. References

[1] Alam H. *et al.*, FaxAssist : an automatic routing of unconstrained fax to email location , IS&T/SPIE conf. on document recognition and retrieval, San José, 2000, p. 148–156.

[2] Baumann S. *et al.*, Message extraction from printed documents: a complete solution, ICDAR'97, Ulm, 1997, p. 1055–1059.

[3] Lii J., Srihari S. N., Location of name and address on fax cover pages, ICDAR'95, Montréal, 1995, p. 756–759.

[4] F. Cesarini *et al.*, *INFORMys : a flexible invoice-like form reader system,* IEEE PAMI, Vol 20, no 7, 1998, p. 730-745.

[5] Named-entity task definition, 6th Message Understanding Conference, MUC-6 Columbia, 1995.

[6] Bikel D. *et al.*, Nymble a high-performance learning Name-finder, 5th Conf. on Applied Natural Language Processing, Washington, 1997.

[7] Likforman-Sulem L., Name block location in facsimile images using spatial/visual cues, *ICDAR'01,* Seattle; 2000, p. 680-684.

[8] Yu Hen Hu, Jenq-Neng Hwang, Handbook of Neural Network Signal Processing, CRC Press, New York, 2001