LAperLA: an integrated graphical-linguistic System for old printed Latin Texts

Andrea Bozzi
Area della Ricerca di Pisa
ILC-CNR via G. Moruzzi 1, 56124 I- Pisa
phone +39.050.3152867 fax +39.050.3152839
andrea.bozzi@ilc.cnr.it

Abstract

LAperLA (*Lettore Automatico per Libri Antichi*) is a prototype for the automatic recognition of Latin texts in old printed books. The strengths of the system are the neural architecture and the post-processing linguistic tool that is represented by an index of Latin forms (more than 500,000) and by a query management system which uses the information of the index to check and correct the interpreted words. The images have been taken from the text of "Contradicentium Medicorum" by Girolamo Cardano in the edition printed on 1663; the main textual material consists of a set of 40 image-files (11 for the training and 29 for testing) with a resolution of 118 DPI. We would like to point out that the interpretation results produced on images chosen as benchmarks by LAperLA have been compared with Fine Reader 4.0 by Abby and Omnipage Pro 10 by Caere. FineReader reaches correctness percentage of 61.19%; Omnipage gets to 54.41%, while LAperLA recognises the 80.95% of words which increases with the aid of the specific linguistic module (93,22%). A very easy to use system interface has been developed not only for the training of the net, but also to select the parts of the image-files to be interpreted.

1. OCR systems: correctness comparative analysis

The most sophisticated OCR systems include a preliminary phase of training on image-files and an effective recognition step. In order to assess our experimental system, named LAperLA (*Lettore Automatico per Libri Antichi*, containing Latin texts), it has been necessary to compare its performances with well known modern commercial softwares.
Fine Reader 4.0 by Abby and Omnipage Pro 10 by Caere have been chosen as commercial systems for the mentioned comparative analysis. The images have been taken from the text of "*Contradicentium Medicorum*" by Girolamo Cardano in the edition printed on 1663; the main textual material consists of a set of 40 image-files (11 for the training and 29 for testing) with a resolution of 118 DPI. The available textual images are characterised by the typographical specificity, by the editorial conventions and by different factors of degradation and noise[1].
Limits and advantages, strengths and breakdowns has been individuated whit reference to every tested software.

1.1 Fine Reader

First of all Fine Reader 4.0 incorporates an image pre-processing system and, during the training process, it allows the user to correct errors of the blobs in which each character has been segmented. Examples of such errors are the union of two or more characters into only one blob, the division of a unique character into several blobs, etc. The training for the character set has not relevant memory constraints; in fact, it is possible to provide many samples for each character. Although the training interface respects requests of functionality and usability at all, it has not been possible to train accented letters because they are not provided by the default alphabet and would have to be inserted manually resorting to supplementary symbols. Real Achilles' heel of this software is however the impossibility to import every kind of dictionary as a spelling checker tool for the recognised text.
As regards to the recognition process, a problem is represented by the image interpretation of words containing hyphen.

1.2 Omnipage Pro.

The second tested software regards Omnipage Pro 10; it is equipped with an image pre-processing system as well, but it allows training just of characters that user considers as high specific. In fact every training file contains a table of only 256 cells; it means that the number of available training samples is potentially less then the previous system. Another important restriction is the impossibility to correct segmentation into blobs containing characters. So the training interface is more limiting and inefficient.
As far as the linguistic module, Omnipage facilitates the introduction of a dictionary of no more then 65,536 characters. With the hypothesis of 6-character average length per word, it is possible to introduce a dictionary of no more than 10,922 words, which is absolutely insufficient for Latin texts interpretation and post-processing.

1.3 LAperLA.

The experimental system LAperLA has been built so as to allow both union and division into blobs that have been badly located by the software during the segmentation process. Moreover, training of single

[1] Possible factors of degradation: ageing of document (paper, ink), deterioration caused by use, shading of underneath page, low resolution of digitalisation, aberrations of optical system.

characters (accented letters too) has no memory constraints. 100 samples for each character have been arbitrarily provided during the training phase described below. Another flexibility element consists of the possibility to define alphabetic table according to text font that has to be recognised [2].
The strengths of our system are the neural architecture described in the next paragraph and the post-processing linguistic tool that is represented by an index of Latin forms (more than 500,000) and by a query management system which uses the information of the index to check and correct the interpreted words.
The experimental work set out so far lacks of some integrations, first of all the graphic module for image enhancement and pre-processing [3].

1.4 Comparative evaluation.

The three systems testing interfaces are almost similar, but we would like to point out that they differ about the final recognition results, which are recorded on the appendix table (figure 1). FineReader reaches correctness percentage of 61.19%, Omnipage gets to 54.41%, while LAperLA recognises the 80.95% of words.
It has to be noticed that percentages are measured with respect to correctness regarding words recognition (the test has been performed on 29 image-files for a total of 2,110 words), that punctuation and numbers (no trained) recognition errors are not considered as well as recognition errors on broken words out of the image frame (for example, the words which are written on the image top left corner or on the image down right corner). Before producing the final analysis, interpretation evaluation tables have been built for each of the three tasted systems according to a rigorous classification method of the occurred errors[4]. On the basis of mentioned tables, individual correctness percentages have been calculated.

2. The LAperLA automatic character recognition

The process begins with the the activation of a computational system for the analysis of the bitmap image (two levels of quantization), which is able to perform the recognition of the writing lines within the page and the word-zones within each single line. Method for extracting characters is based on region growing technique. Segmentation process computes brightness level which is common to connected components belonging to each single character.
To extract information about the graphical typology of the character set, it is necessary to transform the low bitmap image information into a higher level representation. Features extraction aims properly to provide a more synthetic characters description able to catch really significant elements (the features) for characters classification. Therefore the binary matrix is changed into a vectorial synthetic representation which will be used by the succeeding classification phase. One features typology focuses on character bitmap subdivided in zones and computes image density of each zone. The area of a digital binary image zone S is the number of pixels having value 1 on S. Although these criteria of features extraction have to do just with spatial characteristics of images, they product information having high discriminating power.
Another features typology consists of projections. Projection provide a good indication of the character presence, of its position and its extension. Therefore the image density along a set of half-lines (each other positioned with different angles) provides information about character extension along chosen directions.
Then, in the case of LAperLA system, the input for the neural network consists of a features vector extracted previously from the pattern that has to be recognised. The hidden layer of the network is composed by k LVQ (Learning Vector Quantization) neural networks, one for each cluster formed analysing input patterns. Therefore it deals with an architecture composed by some parallel networks able to subdivide the whole problem of recognition into smaller problems. The LVQ network divides patterns space into m classes and a reference vector called codebook is defined for each of them; this codebook tends to become representative of its class during training phase. When training phase finishes, the network is able to determine the belonging class of whatever vector just finding the codebook more similar to it (matching function). Every sub-network LVQ is so dedicated to recognition of a single class, which matches a certain pattern that has to be recognised. Moreover, each subnetwork is described by more than one representative neurone (more codebooks for each class), because it happens often that patterns belonging to the same class have got really different characteristics.

3. The LAperLA Linguistic Module

An integrated linguistic module has been considered essential for our experimental system so as to provide a support for recognition assessment. The created module judges corrects all the words it finds on its thesaurus, while it activates a sequence of controls for the words not found on the thesaurus.
The proposed idea is based on the possibility to individuate the groups of characters that the linguistic module confuse more easily and that are not well processed during segmentation phase. On the basis of this knowledge, it is possible to create two tables: one

[2] For the recognition of "*Contradicentium Medicorum*" was useful to introduce a set of specific logotypes (for instance ae, ct, ffe, si, ssi, ecc.).
[3] Blind restoring techniques to remove degradation have being developed by the Information Processing Institute of CNR (IEI-CNR Pisa).
[4] Some examples: character classification not performed at all, union of two or more words into a unique string, division of one word into several segments, confusion between capital and small letters.

containing variants of characters recognised by linguistic module with low confidence degree (figure 2) and the other one containing probable segmentation errors (figure 3). This information will be useful to consult the dictionary with suitable queries on the basis of the scheme shown in figure 4. The query on dictionary for searching words having some variants is performed generating a tree of the given word alternatives. The alternatives are built substituting unsure characters within a word with characters listed on the variants table (figure 5).

Every tree branch corresponds to a possible word having variants. Words provided by the analysis of each branch are then "filtered" by means of the dictionary to find the existing ones. The filter has simply to do with the existence or non-existence of the word within the dictionary; more sophisticated methods could be developed to decrease rank of words signed by the module.

The statistic table on appendix (figure 6) assesses linguistic module performance on the basis of following two criteria:

(a) corrections of errors are considered as valid ones if the module suggests a unique alternative which is exactly the right one;

(b) corrections of errors are considered as valid ones if the module suggests the exact solution within a set of five alternatives at the latest.

Moreover it is considered a third possibility, which consists of the linguistic module activation on output produced by a pre-processing graphic module able to, for instance, refine distance measuring between words so as to avoid union of two segments belonging to different words or division of two or more words tied on a unique string.

Considering the first point (a) as evaluation meter for the module, linguistic control support reaches a correctness percentage of 87.92%, which means an improvement of seven points in the percentage rate (from 80,95% to 87,92%). If point (b) and described above graphic module intervention are considered, the percentage increase up to 92.32% and 93.22% respectively.

Therefore obtained results validate experimental hypothesis, that is to say the necessity of enrich domain knowledge acquired by the automatic characters recognition system with knowledge elements provided according to linguistic context.

Appendix

Images	Correctness (FineReader)	Correctness (Omnipage)	Correctness (La per La without Linguistic Module)
Test1	56,25%	54,17%	85,42%
Test2	60,26%	57,69%	88,46%
Test3	65,49%	43,36%	84,07%
Test4	73,22%	67,86%	85,72%
Test5	68,29%	70,73%	80,49%
Test6	70,69%	58,62%	77,59%
Test7	59,68%	59,68%	87,10%
Test8	54,17%	45,83%	85,42%
Test9	69,44%	54,47%	88,39%
Test10	67,09%	63,29%	72,15%
Test11	70,97%	53,23%	75,81%
Test12	64,06%	43,75%	78,13%
Test13	67,82%	57,47%	75,86%
Test14	62,03%	46,84%	86,08%
Test15	70,15%	58,21%	77,61%
Test16	38,67%	54,67%	70,67%
Test17	65,15%	43,94%	84,85%
Test18	58,67%	64,33%	76,00%
Test19	60,00%	36,67%	75,00%
Test20	57,90%	52,63%	71,93%
Test21	64,45%	42,22%	81,11%
Test22	60,24%	38,56%	72,29%
Test23	67,86%	41,07%	85,72%
Test24	54,29%	48,57%	82,86%
Test25	60,42%	64,58%	88,54%
Test26	57,30%	58,43%	80,90%
Test27	53,85%	76,92%	78,02%
Test28	47,44%	57,69%	87,18%
Test29	50,00%	65,72%	80,00%
Totale	**61.19%**	**54.41%**	**80.95%**

Figure 1: comparative evaluation of the final recognition results

Unsure character	c	e	t	r	f	s	b	h	i	l	u	n	p	q
Variants	e	c	r	t	s	f	h	b	ls	is	n	u	q	p

Figure 2: table of variants

Incorrect Segmentation	nl ln ul lu	il li	ll
Possible Union	m	u	n h

Figure 3: table of the probable segmentation errors

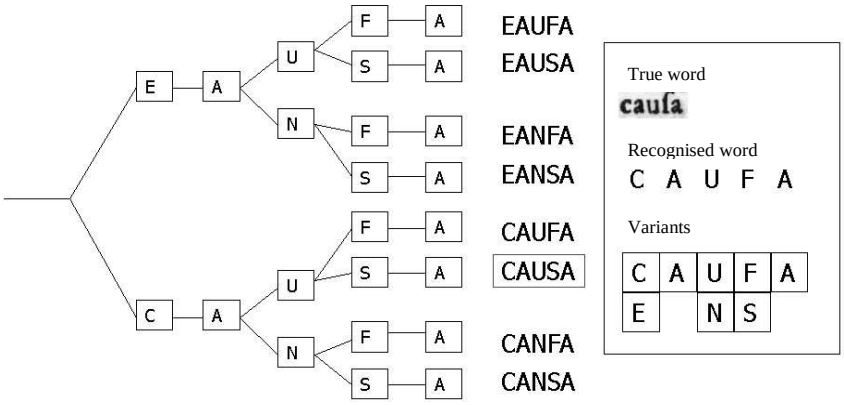

Figure 4: the scheme used to consult the dictionary using variants

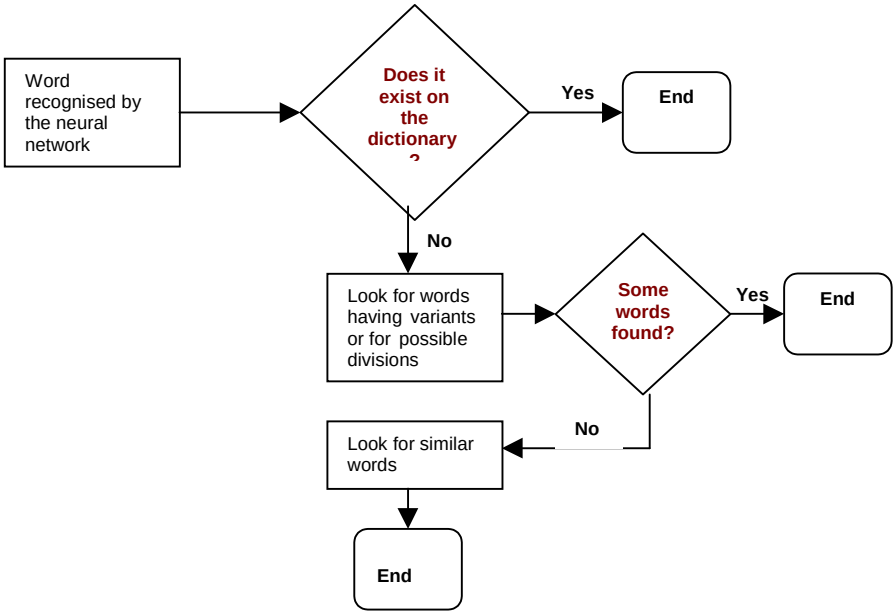

Figure 5: the flow concerning the check of each recognised word on the dictionary

Images	Words number per image	Correctness without Linguistic module (LAperLA)	Correctness with Linguistic module **(a)**	Correctness with Linguistic module **(b)**	Correctness with Linguistic module **(graphic intervent)**
Test1	48	85,42%	91,67%	97,92%	97,92%
Test2	78	88,46%	85,90%	100%	100%
Test3	113	84,07%	95,58%	95,57%	95,57%
Test4	56	85,72%	96,43%	96,43%	96,43%
Test5	41	80,49%	95,12%	95,12%	97,56%
Test6	58	77,59%	91,38%	93,10%	93,10%
Test7	62	87,10%	91,94%	95,16%	98,39%
Test8	48	85,42%	89,58%	93,75%	95,83%
Test9	112	88,39%	92,86%	95,54%	95,54%
Test10	79	72,15%	82,28%	87,34%	87,34%
Test11	62	75,81%	79,03%	83,87%	83,87%
Test12	64	78,13%	89,06%	93,75%	93,75%
Test13	87	75,86%	79,31%	85,06%	86,26%
Test14	79	86,08%	91,14%	93,67%	93,67%
Test15	67	77,61%	82,09%	85,08%	85,08%
Test16	75	70,67%	76,00%	85,33%	88,00%
Test17	66	84,85%	90,91%	95,46%	95,46%
Test18	75	76,00%	78,67%	82,67%	85,33%
Test19	60	75,00%	85,00%	91,67%	91,67%
Test20	57	71,93%	87,72%	91,23%	91,23%
Test21	90	81,11%	86,67%	93,33%	94,45%
Test22	83	72,29%	81,93%	89,16%	89,16%
Test23	56	85,72%	91,07%	94,64%	94,64%
Test24	70	82,86%	88,57%	91,43%	92,86%
Test25	96	88,54%	93,75%	94,79%	95,83%
Test26	89	80,90%	88,77%	93,26%	94,38%
Test27	91	78,02%	90,11%	94,51%	95,61%
Test28	78	87,18%	91,03%	96,15%	96,15%
Test29	70	80,00%	88,57%	92,86%	92,86%
Totale	**2110**	**80.95%**	**87.92%**	**92.32%**	**93.22%**

Figure 6: comparative table of the recognition results when using the LAperLA linguistic module

References

Bozzi, A. (2000), Computer-aided recovery and analysis of damaged text documents, Bologna: CLUEB.

Buscema, M. (1999), Reti neurali artificiali e sistemi sociali complessi, Milano: Franco Angeli.

Cammarata, S. (1997), Reti neuronali. Dal perceptron alle reti caotiche e neuro-fuzzy, Milano: Etas Libri, Milano.

Chen, S. & Haralick, R.M. (1995), Recursive erosion, dilation, opening and closing transforms, IEEE Transaction on Image Processing, 4 (3), 335-345.

Dengel, A. & Hoch, R. & Hones, F. & Jager, T. & Malburg, M. & Weigel, A. (1997), Techniques for improving OCR results, in H. Bunke & M.S.P. Wang (eds.), Handbook of Characters Recognition and Document Image Analysis (pp. 227-258). Singapore: World Scientific Publishing Co. Pte. Ltd.

HA, T.M. & BUNKE, H. (1997), Image processing methods for document image analysis, in H. Bunke & M.S.P. Wang (eds.), Handbook of Characters Recognition and Document Image Analysis (pp. 1-47). Singapore: World Scientific Publishing Co. Pte. Ltd.

Kohonen, T. (1997), Self -Organizing Maps, 2nd ed., Heidelberg: Springer-Verlag.

Aa.Vv. (1993), Optical Character Recognition in the historical Discipline. Proceedings of an International Worgroup organized by: Netherlands Historical Data Archive, Nijmegen Institute for Cognition and Information. St.Katharinen: Scripta Mercature Verlag,.