A Hybrid Strategy for Knowledge Extraction from Biomedical Documents

Margherita Berardi Michelangelo Ceci Donato Malerba
Dipartimento di Informatica - University of Bari
Via Orabona 4, 70125 Bari, Italy
{berardi, ceci, malerba}@di.uniba.it

Abstract

Biomedical documents raise a number of challenges for document image understanding research. In fact, this kind of documents demands to go beyond a geometric-based mapping of layout structures into logical structures and to add semantics to the complex process of document image understanding. On the other hand, processing of the huge amount of available biomedical documents requires the use of the automatic functionalities that are peculiar of machine learning techniques. In this paper we propose a hybrid strategy that enriches the layout-based approach to document image understanding with textual-based features. The proposed solution has been implemented in a new version of the DIAR system WISDOM++, that encapsulates a SVM algorithm for automatic text classification. We report an application of the improved version of the system to a specific dataset of biomedical documents reporting studies on mitochondrial diseases associated to mtDNA mutations.

1. Introduction

In biomedicine, the decoding of the human genome has increased the number of online publications leading to information overload. Every 11 years, the number of researchers doubles [10] and Medline, the main resource of research literature, has been growing with more than 10,000 abstracts per week since 2002. Therefore, it becomes more and more difficult for researchers in biomedicine to keep up with research progresses without the help of automatic tools. Indeed, biologists often maintain some internal and proprietary databases designed for their specific topics of research. However, these databases require to be periodically updated in order to be aligned with results published in recent researches of the field. In addition, biomedical documents are often in digital (i.e. image or pdf) or paper format that cannot be directly stored in databases without the support of specific tools that transform the original unstructured document in a structured format.

If we take into account that documents are organized according to a well-defined structure and that biologists need to retrieve information from "specific" parts of documents [13], we expect that machine learning techniques for Document Image Analysis and Recognition (DIAR) systems might be profitably applied in this domain. More precisely, biomedical documents are organized according to a regular section structure (composed by Abstract, Introduction, Methods, Results and Discussion), and often biologists already know which part of the documents may contain a certain kind of information. From this perspective, the main challenge is to enrich the classical layout-based strategy [2] for document image understanding [14] with textual-based features which allow to add semantics to the complex process of document image understanding. As clearly explained in [4], research trend in document image understanding turns towards the need of document management systems which should be able to employ hybrid strategies for knowledge capture in order to handle different dimensions of information (e.g. textual, layout, format, tabular, etc.). In particular, it seems that document management systems will not give answers to real-world needs until they continue to tailor their solutions for the individual dimensions in which the whole process of document understanding can be articulated.

In this paper, we present a solution to the mentioned issues. The proposed solution has been implemented in a new version of the DIAR system WISDOM++, that encapsulates a support vector machine (SVM) [15] approach for automatic text classification. In this way it is possible to capture both the layout and the textual dimension. We also report an application of the new functionality of the system in the specific field of biomedical documents where, this hybrid strategy is strongly demanded. In particular, we present results on a corpus of biomedical papers selected to contribute to the annotation in the HmtDB resource (http://www.hmdb.uniba.it/) of variability data associated to clinical phenotypes.

The paper is organized as follows. In the next section the upgrade of WISDOM++ is described. In Section 3 the auto-

matic text classifier implemented in WISDOM++ to support content-based document image understanding is presented. Experimental results are shown in Section 4 and some conclusions and future works are presented in Section 5.

2. Supporting hybrid document image understanding in WISDOM++

WISDOM++ is a document analysis system that can transform textual black and white paper documents into XML format [2]. This process involves several steps. First, the image is converted in black and white and is segmented into basic layout components (non-overlapping rectangular blocks enclosing content portions) by means of an efficient variant of the Run Length Smoothing Algorithm. These layout components are classified according to the type of their content: text, horizontal line, vertical line, picture and graphics. This classification is performed by means of a decision tree automatically built from a set of training examples of the five classes. Then, layout analysis is performed in order to detect structures among blocks. The result is a tree-like structure which is a more abstract representation of the document layout. This representation associates the content of a document with a hierarchy of layout components, such as blocks, lines, and paragraphs. Considering the extracted layout, the document image classification is performed. This aims at identifying the membership class (or type) of a document (e.g. business letter, newspaper article, and so on) by means of some first-order rules which can be automatically learned from a set of training examples [9]. In document image understanding, layout components are associated with logical components. This association can theoretically affect layout components at any level in the layout hierarchy. However, in WISDOM++ only the most abstract components of the layout hierarchy are associated with components of the logical hierarchy. Moreover, only layout information is used in document image understanding. Two assumptions are made: documents belonging to the same class have a set of relevant and invariant layout characteristics; logical components can be identified by using layout information only. Document image understanding also uses first-order rules [9]. Once logical and layout structure have been mapped, OCR can be applied only to those textual components of interest for the application domain, and its content can be stored for future retrieval purposes. The result of the document analysis is an XML document that makes the document image retrievable.

This approach differs from approaches to document image understanding proposed by other authors [8] which make also use of textual information, font information and universal attributes given by the OCR. In our approach, only some layout components of interest are subject to OCR and hence document image understanding precedes text reading. Nevertheless, to find responses to real-world demands, as those arisen in the biomedical sector, a more involved approach to document image understanding should be adopted. In particular, it is possible that layout components cannot be completely "understood" by considering only geometrical features. In fact, by using layout-based document image understanding approaches, they are labeled as generic logical components (e.g "paragraphs", "section titles", etc.). This motivates to go beyond this approach and to consider the textual content of layout components in order to enrich the logical structure with "semantic components" such as "introduction section", "results section", "methods section" etc. (see Figure 1).

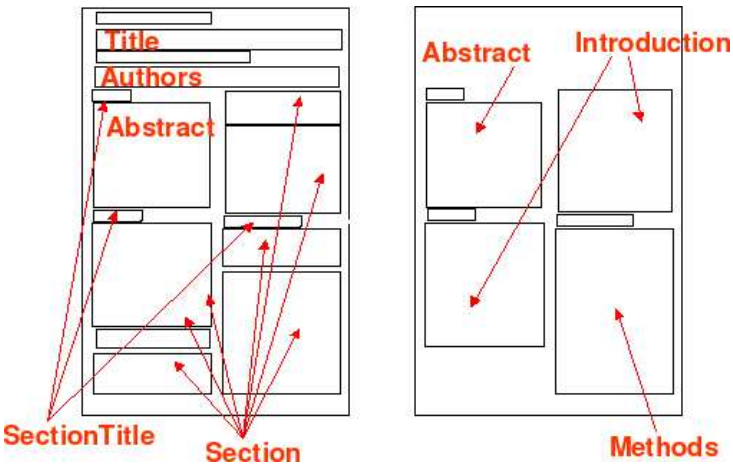

Figure 1. Example of a logical structure: logical and semantic components

By moving towards this higher level of abstraction, we have to consider that the same "semantic component" might be composed by several "logical components" (possibly belonging to different document pages). In our approach, the user can manually specify the reading order [1] of logical components. Moreover, the user can specify the logical components that play the role of separators. In our context, examples of separators are layout components labeled as "section title". By following the detected reading order, text regions labeled with the same logical label are clustered into the same semantic component until a separator is found. Finally, OCR is applied to semantic components and text is used in the text classification task in order to automatically classify semantic components.

The text categorization method implemented in the new version of WISDOM++ is based on an optimized approach for learning SVM. The classifier is built on the basis of a set of training semantic components, and applied to new semantic components. The classification is performed by estimating the posterior probability that a semantic compo-

nent belongs to a category. Details of the learning method are reported in the following section.

3. Content-Based Document Understanding

3.1 The feature selection process

In WISDOM++, each semantic component is considered as an example in the text categorization algorithm. Training examples are associated to a semantic category and are represented as a numerical feature vector, where each feature corresponds to the occurrence of a particular word in the semantic component. In this representation, also called *bag-of-words*, no ordering of words or any structure of text is used. All training examples are initially tokenized, and the set of tokens (words) is filtered, in order to remove punctuation marks, numbers and tokens of less than three characters. Standard text pre-processing methods are embedded in WISDOM++. In particular, *stopwords* are removed, such as articles, adverbs, prepositions etc. (taken from Glimpse-glimpse.cs.arizona.edu), *Stemming* is performed (e.g. 'topolog' is used instead of the words 'topology' and 'topological') by means of Porter's algorithm for English texts [12].

The feature selection process is based on the measure $maxTF \times DF^2 \times ICF$ (eq 1) [3]:

$$v_i = TF_{c'}(w_i) \times DF^2_{c'}(w_i) \times \frac{1}{CF_c(w_i)} \qquad (1)$$

that rewards common words used in semantic components belonging to a category c', and penalizes words common to both c' and other categories. In eq. (1), $TF_{c'}(w_i)$ is the maximum value of the *relative* frequency of the token w_i in an example belonging to c'. $DF_{c'}(w_i)$ is the *document frequency*, that is, the percentage of examples of category c' in which the feature w_i occurs. $CF_c(w_i)$ is the *category frequency*, that is, the number of categories c" $\neq c'$ such that w_i occurs in an example $d \in c$".

The category dictionary of c', $Dict_{c'}$, is the set of the best n_{dict} terms with respect to v_i, where n_{dict} is a user defined parameter. In WISDOM++, the feature set is unique and is obtained as the union of all category dictionaries.

3.2 The learning process

The feature set, built in the previous step, is used to build the SVM classifier. In WISDOM++, we use SVMs because of their well established good performances in text categorization. The problem of learning SVMs is defined for two-classes problems as follows: Given a set of positive and negative examples $\{(\vec{x}_1,y_1), (\vec{x}_2,y_2), \ldots, (\vec{x}_N,y_N)\}$, where $\vec{x}_i \in R^m$ ($\vec{x}_i$ is a feature vector) and $y_i \in \{-1,+1\}$, an SVM identifies the hyperplane in R^m that linearly separates positive and negative examples with the maximum margin (*optimal separating hyperplane*).

In general, the hyperplane can be constructed as the linear combination of all training examples, however, only some examples, called *support vectors*, do actually contribute to the optimal separating hyperplane, which can be represented as:

$$f(x) = \sum_{i=1}^{N^*} y_i \alpha_i \vec{x}^*_i \cdot \vec{x} + b$$

where $\vec{x}^*_i$, i=1,2, $\ldots N^*$, are the support vectors. The coefficients α_i and b are determined by solving a large-scale quadratic programming problem, for which efficient algorithms exist, which are guaranteed to find the global optimum.

SVMs are based on the *Structural Risk Minimization* principle: a function that can classify training data accurately and which belongs to a set of functions with the lowest capacity (particularly in the VC-dimension) [15] will generalize best, regardless of the dimensionality of the feature space m. Therefore, SVMs can generalize well even in large feature space, such as those used in text categorization. In the case of the separating hyperplane, minimizing the VC-dimension corresponds to maximizing the margin.

The linear separability appears to be a strong limitation, however, as experimentally observed by [7], most text categorization problems are linearly separable. In any case, SVMs can be generalized to non-linearly separable training data by mapping the data into another *feature space* F via a non-linear map Φ :$\mathrm{R}^m \rightarrow F$ and then performing the above linear algorithm in F. Generally the map introduces new features that take into account the p-order correlation between the input features. Since the solution has the form:

$$f(x) = \sum_{i=1}^{N^*} y_i \alpha_i \Phi(\vec{x})^*_i \cdot \Phi(\vec{x}) + b \qquad (2)$$

it is non linear in the original feature set. [16] report that they tested the linear and non-linear models offered by the SVM^{light} system [7], and obtained "a slightly better result with the linear SVM than with the non-linear models". Therefore, in WISDOM++ we will use only linear models.

The SVM embedded in Wisdom++ is a modified version of the Sequential Minimal Optimization classifier (SMO) [11]. It is very fast and is based on the idea of breaking the large quadratic programming (QP) problem down into a series of smaller QP problems that can be solved analytically. The same system has been applied by [5] in text categorization.

However, a modification of Platt's original method is necessary in WISDOM++, since each semantic component

can be associated only to one category. This means that the learned classifier is of the kind $one - of - r$ (multi-class classifier). More precisely, a binary classifier is learned for each couple of categories and afterwards, the probability that a semantic component belongs to a category is computed by means of a probabilistic pair-wise coupling classification [6]. In the classification process, the semantic component is associated to the most probable category.

4. Experimental results

In this section we study the performance of the content-based classifier on a set of thirty-four paper documents reporting studies on mitochondrial diseases associated to mtDNA mutations (HmtDB resources). Initially, document images are processed by performing layout analysis and document classification. The layout structure is then mapped into a corresponding logical structure on the basis of automatically learned rules that take geometric information into account (see figure 2). On layout components labelled as generic logical components the OCR is applied. Training examples are generated by asking domain experts (biologists) to manually categorize textual components into five categories (Abstract, Introduction, Methods, Results and Discussion). We evaluated the performance of the content-based classifier by means of a 5-fold cross-validation, that is, the dataset is first divided into five *folds* of near-equal size, and then, for every fold, the classifier is trained on the remaining folds and tested on it. The classifier performance is estimated by averaging classification accuracy on the five cross-validation folds. Figure 3 shows a WISDOM++ screenshot representing the extracted semantic components. In table 1, results varying the fold and the dictionary size n_{dict} are reported.

Results show that when $n_{dict} = 30$, the classifier reaches the best performance. This is mainly due to the relatively small size of training examples (between 200 words and 1900 words). Concerning the accuracy of the classifier, if we compare the results with those obtained with the trivial classifier than returns the most frequent category (accuracy = 20%), we note that the SVM classifier is much more accurate. This good performance is also due to the feature selection method that is able to extract discriminant features (tokens). For example, it selects relatively generic tokens for the "Abstract" category (e.g. 'percentag','mitochon') and relatively specific tokens for the "Methods" category (e.g. 'tRNA','tissu','enzym').

5. Conclusions

In this paper we presented a hybrid strategy for document image understanding which takes full advantage of both layout and textual features in order to meet challenges coming from documents of the biomedical domain. In particular, the proposed method aims to exploit the textual content of layout components in order to enrich the logical structure with "semantic components". The method is based on an SVM classifier and it has been implemented in a new version of WISDOM++. We also report an application of the new functionality of the system in the specific field of biomedical documents where, this hybrid strategy is strongly demanded. For future work we are investigating the possibility to learn rules for automatic reading order detection from the user interaction.

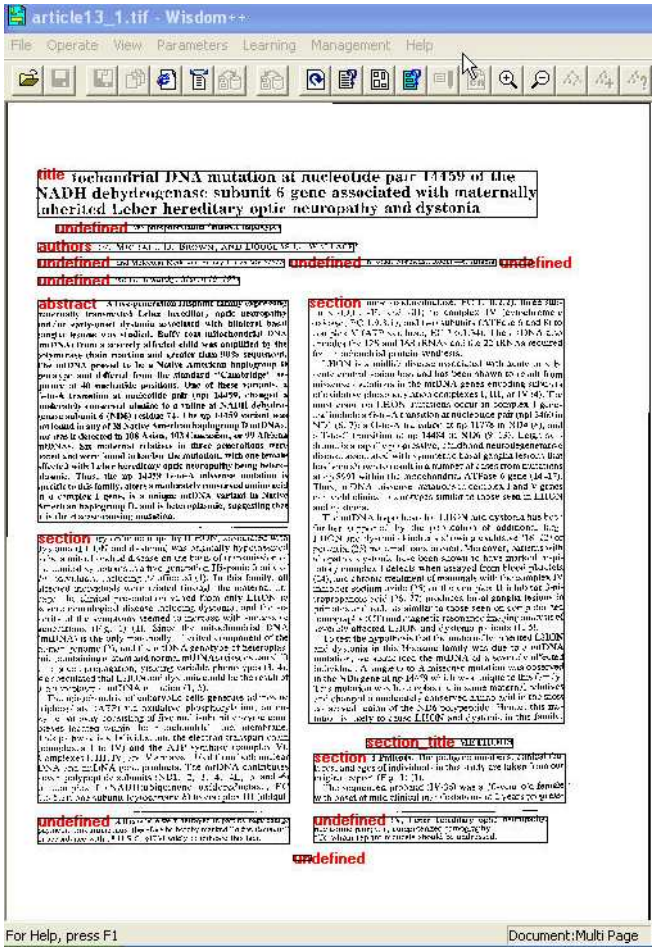

Figure 2. WISDOM++ screenshot: logical components

Acknowledgements

This work has been funded by the IBM Faculty Award 2004 received from IBM Corporation to promote innovative, collaborative research in disciplines of mutual interest.

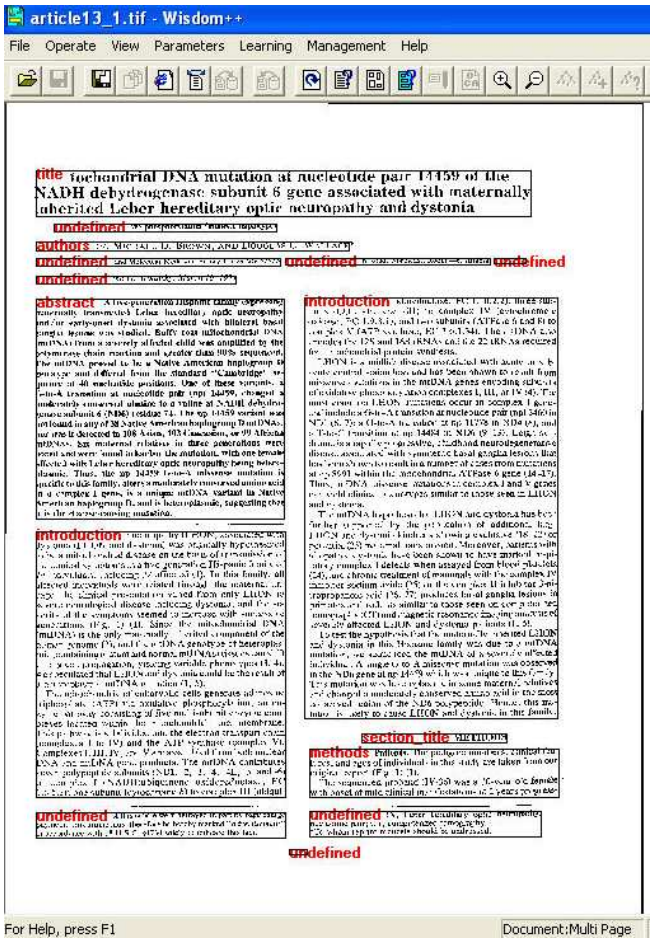

Figure 3. WISDOM++ sceenshot: semantic components

	n_{dict}			
Fold	**10**	**20**	**30**	**40**
fold1	70	73.33	80	80
fold2	60	80	83.33	80
fold3	63.33	66.67	70	70
fold4	36.67	56.67	56.67	56.67
fold5	42	52	58	58
AVG	*54.4*	*65.73*	*69.6*	*68.93*

Table 1. Accuracy percentage

References

[1] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *IJDAR*, 5(1):1–16, 2002.

[2] O. Altamura, F. Esposito, and D. Malerba. Transforming paper documents into XML format with WISDOM++. *IJDAR*, 4(1):2–17, 2001.

[3] M. Ceci and D. Malerba. Hierarchical classification of HTML documents with WebClassII. In *ECIR-03, European Conference on Information Retrieval*, pages 57–72, 2003.

[4] A. R. Dengel. Making documents work: Challenges for document understanding. In *ICDAR 03*, pages 1026–1036, 2003.

[5] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. of ACM-CIKM98*, pages 148–155, 1998.

[6] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *Proc. of Advances in neural information processing systems*, pages 507–513. MIT Press, 1998.

[7] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of Eur. Conf. on Machine Learning*, pages 137–142, 1998.

[8] S. Klink, A. Dengel, and T. Kieninger. Document structure analysis based on layout and textual features. In *Proc. of DAS2000*, pages 99–111, 2000.

[9] D. Malerba. Learning recursive theories in the normal ilp setting. *Fundamenta Informaticae*, 57(1):39–77, 2003.

[10] M. F. Perutz. Will biomedicine outgrow support? *Nature*, (399):299–301, 1999.

[11] J. Platt. *Advances in kernel methods - support vector learning*, chapter Fast training of support vector machines using sequential minimal optimization. MIT Press, 1998.

[12] M. F. Porter. *Readings in information retrieval*, chapter An algorithm for suffix stripping, pages 313–316. 1997.

[13] K. P. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4(1), 2003.

[14] S. Tsujimoto and H. Asada. Understanding multi-articled documents. In *Proc. of the 10th ICPR*, pages 551–556, 1990.

[15] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[16] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99*, pages 42–49, 1999.