ABBA: Smart Information and Document Analyser

Özlem Uçar, Gurkan Tuna, Samsun M. Başarıcı, and Yılmaz Kılıçaslan

***Abstract*—During the last couple of decades the number of documents in digital form has grown tremendously in size. Consequently, being able to automatically organise and classify documents is highly important. The target of research in classification is to partition unstructured sets of documents into groups that give a description of the contents of the documents. Since the task is to assign documents to one or more classes or categories, it can be realised either manually or automatically, based on a set of algorithms.**

This paper presents a smart information and document analyser called ABBA. ABBA uses machine learning and natural processing techniques to provide a number of useful functions including correcting mistyped words in documents, separating intermingled documents, classifying documents into designated groups, extracting information from documents, and easing document access, document control, management and storage by processing even unstructured documents. In addition to presenting the details of ABBA, this paper also gives information about major solution areas and sectors in which ABBA can be used.

***Index Terms*—Information analysis, document analyser, machine learning techniques, natural language processing.**

I. Introduction

Physical archives and record storage facilities provide a proper environment for the purpose of storing documents, records and materials which require permanent protection. Although archiving and record storing must be done in high-performance buildings with systems designed to operate permanently with zero tolerance for failure, in reality the situation is much different in most low to medium income countries, and since archiving and record storing are delivered by the offices typically located in basements, in disasters such as flooding and fire, important documents and records are in danger. Digital archives, on the other hand, are protected against threats such as unauthorized access and data loss, and physical hazards are not a concern for such archives.

As mentioned above, traditional physical archives are open to different threats. Moreover, the documents in those archives may become obsolete over time and the writings on the documents may become unreadable. Since the archives can contain many documents, it is a difficult and time-consuming process for the employees to reach these documents, to process them on daily basis and to update them. On the other hand, digital archives do not require manpower for operations such as finding a document and updating it. These operations are performed very quickly and effortlessly. It is quite easy to create versions for updated documents and keep them together.

Document analysis can be described as the process of examining the content and structure of documents in order make information therein analysable; and, information analysis is the process of inspecting, transforming and modelling information in support of a decision-making process.

The term ‘document analysis’ refers to the processes of locating and analysing patterns or facts in existing documents and can be performed either as a standalone procedure or as a necessary forerunner to collecting new data using other approaches [1], [2]. Its time and cost requirements are minimal since it does not involve collecting new data. Basically, it involves gathering information used in a formal description of the electronic text and examining the content and structure of documents to identify and name the components of some class of documents, to specify their interrelationships, and to name their properties. Since in recent years the amount of information available in electronic publications, e-books, e-mail messages, news articles, blogs, and Web pages has increased rapidly, document classification has become a critical process.

Machine learning (ML) and natural language processing (NLP) are two widely used techniques in document classification. ML methods can be broadly categorised into two main categories: supervised and unsupervised. Supervised ML methods require predefined category labels to be assigned to instances in training sets. However, in unsupervised ML methods, training instances do not carry class labels.

In the last decade, document classification has been heavily investigated [3]-[9]. Baker *et al.* in [10] report the results of a series of experiments that consist of the integration of fully automatic document classification approaches into an existing document retrieval system. In [11], Joachims explains the use of Support Vector Machines (SVMs) for learning text classifiers from examples by analysing the particular properties of learning with text data and identifying the suitability of SVMs for document classification.

NLP is a multidisciplinary field and concerned with the interactions between computers and natural languages [12], [13]. Therefore, NLP is related to the area of human-computer interaction. In the context of text classification, NLP allows for the introduction of linguistic rules to the system. Basile *et al.* in [14] argue that to provide a large set of semantically annotated texts with formal semantics, using a bootstrapping method that comprises state-of-the-art NLP tools for parsing and semantic interpretation is a good approach.

As ABBA is mainly based on Turkish, it might be useful to go into the relevant literature for this language to some extent.

Manuscript received June 10, 2017; revised September 10, 2017. This work was supported in part by the Scientific and Technological Research Council of Turkey under Grant 7150455.

Ö. Uçar and G. Tuna are with the Department of Computer Engineering, Trakya University, Edirne, Turkey (e-mail: ozlemu@trakya.edu.tr, gurkantuna@trakya.edu.tr).

S. M. Başarıcı and Y. Kılıçaslan are with the Department of Computer Engineering, Adnan Menderes University, Aydın, Turkey (e-mail: sbasarici@adu.edu.tr, yilmaz.kilicaslan@adu.edu.tr).

doi: 10.18178/ijlll.2017.3.3.116

In [15], Kılıçaslan argues that even discourse-pragmatic factors come into play in the informational structuring of Turkish sentences. In [16], it is claimed that the specificity status of the referent of a non-partitive (NP) cannot be the ultimate criterion for this NP to carry or not to carry case morphology in this language. Moreover, it is argued that the determining factor in case marking alternations in Turkish is the informational status of the NP in the partitioning of the semantic material of the sentence between described and resource situations. Kılıçaslan *et al.* in [17] propose a three-layered morpho-phonological analyser for Turkish and design and implement three types of automata to carry out the autonomous operations. The layered approach is based on the recognition of the autonomy of the levels of a language and each automaton is responsible for combining linguistic units relevant to the level implemented into larger fragments.

In sum, although NLP is a highly useful tool in classification tasks, complementary methodologies and approaches to automatically correct mistyped words [18]-[20] and extract information from documents [21], [22] are needed to fulfil a text classification task successfully.

In contrast to the studies in the literature that focus on the use of specific ML techniques or NLP approaches for document classification, this study mainly focuses on developing fully operational information and document analyser via the integration of ML techniques. The rest of this paper is structured as follows. Section II presents an overview of ABBA. Section III explains how ABBA handles documents in detail. Finally, this paper is concluded in Section IV.

TABLE I: ABBA VS CONVENTIONAL DOCUMENT CLASSIFICATION AND ARCHIVING APPROACHES

	ABBA	Others
Cost	Not costly and effective	It is time costly to find, update, and archive documents.
Manpower	Only a computer is required to process documents.	Significant manpower required to process documents.
Practicality	Up-to-date and practical.	Out-of-date and impractical.
Classification	Documents are analysed; learning features are extracted; and rules of classification identified.	PDFs made searchable through OCR
Search	Easy to search within documents.	Hard to search within documents.
Update	It is fast to find, update, and then archive documents.	Updating takes significant time.

II. A TOP VIEW OF ABBA

ABBA is an intelligent information and document analysis tool. With its learning and NLP capabilities and with its ability to automatically process even unstructured documents, ABBA has distinct advantages over conventional document classification and archiving approaches as listed in Table I.

ABBA is equipped with the following functionalities:

- morphologically analysing words,
- spell-checking and correcting mistyped words (which are very frequent in Optical Character Recognition (OCR)ed documents) in documents,
- separating intermingled documents,
- classifying documents into designated groups,
- extracting information from documents,
- easing document access, document control, management and storage.

With its above-mentioned distinct functions, ABBA is an effective solution for many areas such as service office (scanning & data entry), image/data capture, institutional content management and archive, electronic document management, media management, process modelling andprocess management solutions, and physical archive management.

ABBA can work either as a standalone module or as part of an automated content management process and can accept structured, semi structured or unstructured forms via e -mail, mobile devices, fax, physical papers, and files. Before going into the details of its internal structure and working, let us have a quick look at the content management process where ABBA is situated. Fig. 1 depicts this process.

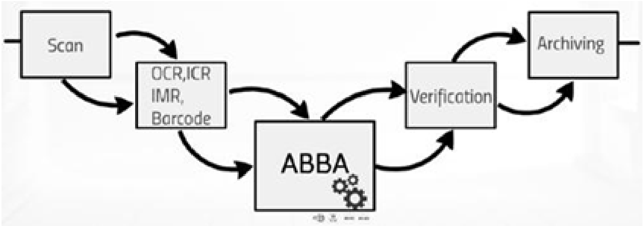


Fig. 1. ABBA as part of a content management process.

ABBA is a station in a journey of documents. As shown in Fig. 2 the journey starts with a scanning operation.

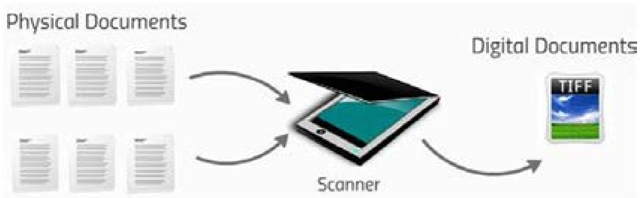


Fig. 2. Conversion of physical documents into digital images.

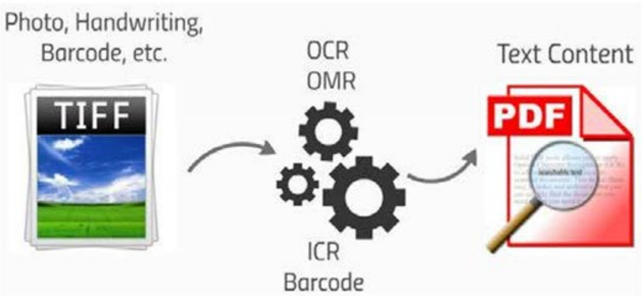


Fig. 3. Conversion of digital images into machine-encoded texts.

This is simply an operation through which the content of a physical document is converted into a digital image. As shown in Fig. 3 the scanning operation is followed by an operation of character recognition. This is an operation whereby the image of a typed, handwritten, or printed text is converted into a machine-encoded text. This operation can be carried out with software equipped with an OCR, Intelligent Character Recognition (ICR), Optical Mark Recognition (OMR), or barcode reading function.

The machine-encoded texts are fed into ABBA of which they come out as spell-checked and –corrected, grouped, classified and easily accessible documents. The output of ABBA goes to a verification stage as shown in Fig. 4. In this stage, documents are screened and checked by staff to verify that they are correctly classified into designated types. Finally, the verified documents are organised and archived for later use (see Fig. 5). In what follows, we will look at the internal structure and functionality of ABBA.

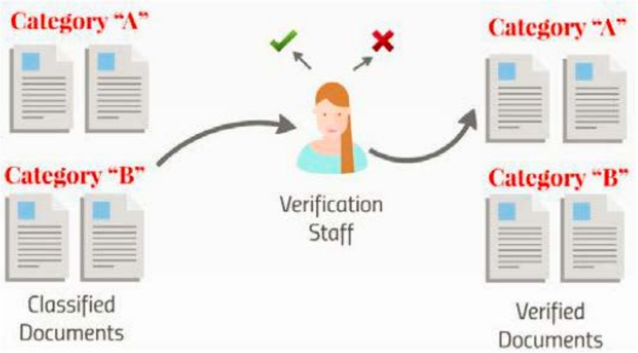


Fig. 4. Verification of the classification of documents.

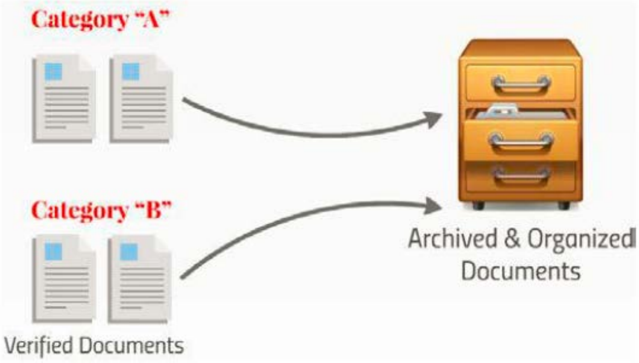


Fig. 5. Archiving the verified documents.

III. ABBA from an Internal Point of View

ABBA consists of three modules. All documents coming into ABBA first have to go through a Word Processor. Afterwards, they pass through either a Rule-based Document Processor or a Learning-based Document Processor (see Fig. 6).

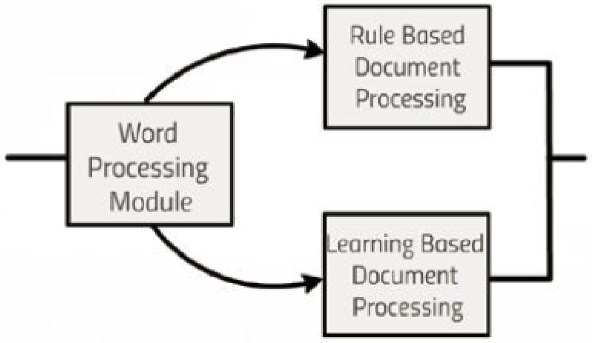

Rule Based Document Processing
Word Processing Module
Learning Based Document Processing

Fig. 6. The internal structure of ABBA.

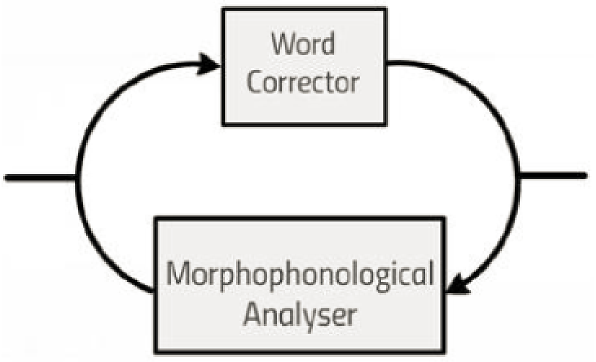


Fig. 7. Structure of the word processor of ABBA.

The structure of the word processor of ABBA is shown in Fig. 7. To improve the accuracy of the word corrector used in ABBA, a morpho-phonological analysis module has been developed. Since most of the documents in our experimental study originate from Turkish, a finite-state automaton has been designed for analysing Turkish words (see Fig. 8). Using the automaton, the words are separated from the suffixes and the stems are identified. The overall word processing stage of ABBA is shown in Fig. 9.

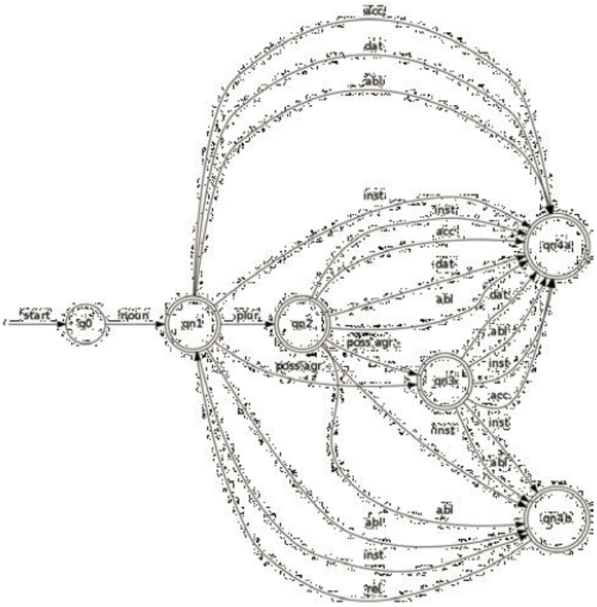


Fig. 8. Automaton used by the morpho-phonological analyser.

Document separation and classification is either realised using the rule-based document processor shown in Fig. 10 or the learning-based document processor shown in Fig. 11. Both processors presuppose a set of words to be used for discriminating documents from each other. Very roughly, if a word occurs in a certain type of documents more frequently compared to the rest, it receives a discriminating status for that type of documents. The words to be used to this effect are determined in a context-sensitive way by analysing the documents to be classified. The rule-based document processor makes use of these discriminators in some user-defined classification formulas. For the learning-based processor, the discriminators serve as learning attributes. Therefore, each instance in the training set is annotated in accordance with these attributes. The training set comprises a certain amount of documents to be processed. The process of document separation and classification of ABBA is shown in Fig. 12.

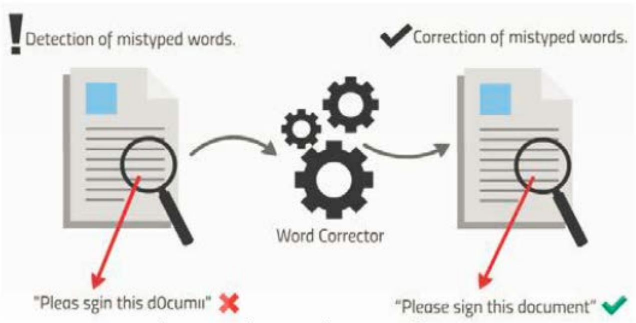

Detection of mistyped words.
Correction of mistyped words.
Word Corrector
"Pleas sgin this dOcumıı"
"Please sign this document"

Fig. 9. Word processing stage of ABBA.

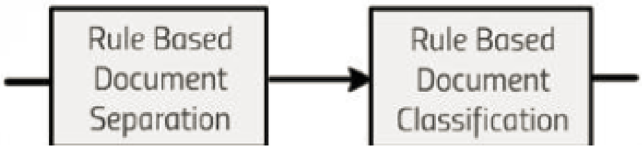

Rule Based Document Separation
Rule Based Document Classification

Fig. 10. Structure of the rule based document processor of ABBA.

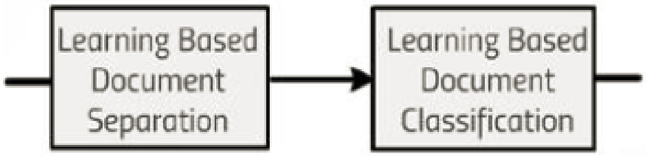

Learning Based Document Separation
Learning Based Document Classification

Fig. 11. Structure of the learning based document processor of ABBA.

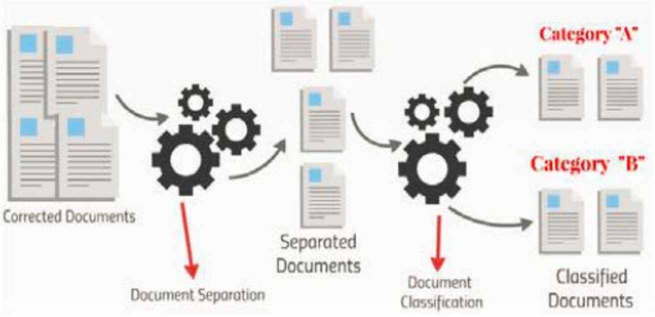


Fig. 12. Document separation and classification stages of ABBA.

IV. Conclusion

Effective business processes require fast and easy access to relevant information and documented evidence. Therefore, every organization needs a system that classifies documents to enhance its business processes. Automatic document classification appears in many applications and is attractive since manually organising document bases can be costly or unfeasible given the time constraints of the application or the number of documents involved in the classification task.

Document analysis can be described as a set of functions for automatic detection of specific objects on a page such as text blocks, barcodes, tables, pictures, and separators. In this paper, ABBA, which has been developed to speed up the document analysis process and to raise the quality of this process, is presented. ABBA bears a crucially degree of novelty as it brings together both ML and NLP techniques for document analysis.

References

[1] J. L. Pershing, "Using document analysis in analyzing and evaluating performance," *Perf. Improv.*, vol. 41, pp. 36-42, 2002.

[2] B. R. Witkin and J. W. Altschuld, *Planning and Conducting Needs Assessments: A Practical Guide.* Newbury Park, CA: Sage Publications, 1995.

[3] H. Borko and M. Bernick, "Automatic document classification," *Journal of the ACM*, vol. 10, no. 2, pp. 151-162, 1963.

[4] W. W. Cohen and Y. Singer, "Context-sensitive learning methods for text categorization," *ACM Transactions on Information Systems*, vol. 17, no. 2, pp. 141–173, 1999.

[5] L. Denoyer and P. Gallinari, "Bayesian network model for semi-structured document classification," *Information Processing & Management*, vol. 40, no. 5, pp. 807-827, 2004.

[6] D. L. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361-397, 2004.

[7] F. Li and Y. Yang, "A loss function analysis for classification methods in text categorization," in *Proc. International Conference on Machine Learning (ICML)*, 2003, pp. 472-479.

[8] R.-L. Liu and Y.-L. Lu, "Incremental context mining for adaptive document classification," in *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 599-604.

[9] T. Salles, L. Rocha, F. Mourão, G. L. Pappa, L. Cunha, M. A. Gonçalves, and W. Meira Jr., "Automatic document classification temporally robust," *Journal of Information and Data Management*, vol. 1, no. 2, pp. 199-211, 2010.

[10] K. Baker, A. Bhandari, and R. Thotakura, "An interactive automatic document classification prototype," in *Proc. the Third Workshop on Human-Computer Interaction and Information Retrieval*, Washington, D.C., 2009, pp. 30-33.

[11] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Proc. Tenth European Conference on Machine Learning*, 1998, pp. 137-142.

[12] W. Daelemans and V. Hoste, "Evaluation of machine learning methods for natural language processing tasks," in *Proc. the Third International Conference on Language Resources and Evaluation*, 2002, pp. 755-760.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An* Introduction *to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2006.

[14] V. Basile, J. Bos, K. Evang, and N. Venhuizen, *Developing a large Semantically Annotated Corpus*, 2012.

[15] Y. Kılıçaslan, "Syntax of information structure in Turkish," *Linguistics*, vol. 42, no. 4, pp. 717-765, 2004.

[16] Y. Kılıçaslan, "A situation theoretic approach to case marking semantics in Turkish," *Lingua*, vol. 116, pp. 112-144, 2006.

[17] Y. Kılıçaslan, Ö. Açıkgöz, and Ö. Aydın, "A three-layered morpho-phonological analyzer for Turkish," *Journal of International Scientific Publications: Materials, Methods and Technologies*, vol. 8, 2014.

[18] S. N. Srihari, *Computer Text Recognition and Error Correction*, Los Alamitos, CA: IEEE Computer Society Press, 1985.

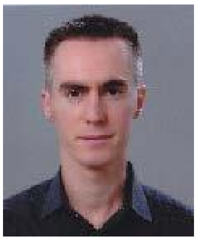
[19] F. Ahmed, E. W. de Luca, and A. Nürnberger, "Revised N-gram based automatic spelling correction tool to improve retrieval effectiveness," *Polibits*, vol. 40, 2009.

[20] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys*, vol. 24, pp. 377-439, 1992.

[21] E. Cortez and A. S. da Silva, *Unsupervised Information Extraction by Text Segmentation*, Springer, 2013.

[22] M.-F. Moens, *Information Extraction: Algorithms and Prospects in a* Retrieval *Context*, Springer, 2006.

Özlem Uçar is an assistant professor at the Department of Computer Engineering in Trakya University, Turkey. She has authored papers in refereed journals and international conference proceedings and has been actively serving as a reviewer for international journals and conferences.



Gürkan Tuna is an associate professor at the Department of Computer Programming of Trakya University, Turkey. He has authored several papers in international conference proceedings and refereed journals, and has been actively serving as a reviewer for international journals and conferences.

Samsun M. Başarıcı is an assistant professor at the Department of Computer Engineering in Adnan Menderes University, Turkey. He has authored papers in refereed journals and international conference proceedings and has been actively serving as a reviewer for international journals and conferences. He has also authored a book and edited a conference proceeding.

Yılmaz Kılıçaslan is a professor at the Department of Computer Engineering in Adnan Menderes University, Turkey. He has authored many papers in refereed journals and international conference proceedings and has been actively serving as a reviewer for international journals and conferences.