Cooperative and Fast-Learning Information Extraction from Business Documents for Document Archiving

Daniel Esser

Technical University Dresden
Computer Networks Group
01062 Dresden, Germany
daniel.esser@tu-dresden.de

Abstract. Automatic information extraction from scanned business documents is especially valuable in the application domain of document management and archiving. Although current solutions for document classification and extraction work pretty well, they still require a high effort of on-site configuration done by domain experts or administrators. Especially small office/home office (SOHO) users and private individuals often do not use such systems because of the need for configuration and long periods of training to reach acceptable extraction rates.
Therefore we present a solution for information extraction out of scanned business documents that fits the requirements of these users. Our approach is highly adaptable to new document types and index fields and uses only a minimum of training documents to reach extraction rates comparable to related works and manual document indexing. By providing a cooperative extraction system, which allows sharing extraction knowledge between participants, we furthermore want to minimize the number of user feedback and increase the acceptance of such a system.
A first evaluation of our solution according to a document set of 12,500 documents with 10 commonly used fields shows competitive results above 85% F1-measure. Results above 75% F1-measure are already reached with a minimal training set of only one document per template.

Keywords: Document Layout Analysis, Information Extraction, Cooperative Extraction, Few-Exemplar-Learning

1 Introduction

Today a huge amount of communication between business partners is done in a textual manner. The movement towards paperless offices all over the world and the need for archiving due to legal regulations will further increase this tendency. To handle the wealth of information, companies need fast and accessible technologies for document management and archiving. Existing solutions like smartFIX [1] and OpenText [2] automate this process. They provide functionality to scan paper documents, classify them according to business processes and

extract relevant information. This index data can later on be used for improving document search, automatical handling of correspondences or attaching them to existing ERP systems or internal databases.

While current solutions work pretty well for large and medium-sized companies, it still requires a high effort of on-site configuration to adapt such a system to the own requirements. Depending on the domain of the institution, new document types and index fields have to be considered and specialized extraction mechanisms have to be added. Especially for SOHO users and private individuals, the need for configuration and long periods of training to archive high classification and extraction rates constrain the usage of such systems.

The goal of this work is the reduction of configuration efforts in business document extraction to make such kind of systems more attractive and suitable for small companies. Therefore we plan a solution, which combines highly adaptable and fast-learning classification and extraction algorithms adapted to the domain of business documents with a cooperative approach to share extraction knowledge with other participants. While specialized algorithms reduce the need for locally available pre-annotated documents, the cooperative approach minimizes feedback by improving the system performance using foreign extraction knowledge. Altogether the combination of both procedures lowers necessary user interaction and ends up in minimal effort for manual configuration.

The local algorithms will mainly focus on incremental few-exemplar learning and work purely training-based without a need for generating rules or pre-annotated example documents. Relevance feedback provided by the user, i.e. new document types and fields or corrections of wrong extractions, is taken into account for the classification and extraction of the next document.

The cooperative approach is based on a pool of common extraction knowledge every user can contribute to. This allows to include individual extraction knowledge and profit from annotations and corrections another user has already done. Especially in the domain of business correspondences, where similar documents are exchanged between many companies, the probability for finding another participant having information on how to extract a given document is quite high. The idea of such collaborative solutions is not new at all [3]. Therefore we want to focus on problems coming along when connecting thousands of local systems to a huge distributed one. While the size of such a global pool increases very fast, we want to explore the relation between size and extraction performance and give solutions for separating distributed extraction systems.

2 Related Work

The classification of business documents and the extraction of relevant information has been tackled by a lot of researches in the last twenty years.

The categorization according to the document type is nearly solved. As an evidence for that, one can see the high spread of this functionality within com-

mercial products from ABBYY[1], Insiders Technologies[2] and OpenText[3]. Therefore a lot of novel works focus on the categorization of documents according to their template. Current solutions either use the document's text [4] or different levels of layout [5–7] as feature base. While [6, 7] only work on the density of pixels, more sophisticated approaches like [5, 8] try to transform a document into a high-level representation, i.e. trees or graphs that can later on be used by matching or learning algorithms. The machine learning approaches used in current solutions are manifold and vary from symbolic algorithms [5] over artificial neuronal networks [9] to statistical [7] and instance-based [10] techniques.

The extraction of information is mostly done on top of a classification. Knowing the type and template of a document allows to draw a conclusion on the existing index fields within that document. Solutions in the field of document extraction differ in the level of granularity of extracted information. While some authors only try to identify single field values [11, 12] – this is the kind of fields we are going to extract – other works focus on finding multiple value fields like contents of lists or tables [13, 14]. The applied techniques do not differ a lot from categorization. Current solutions either use text or layout to find relevant index data. Textual solutions try to find patterns and are mostly built upon some pattern matching [15]. Layout-based solutions try to include position and font effects into their extraction decision [12, 13].

Although current solutions produce acceptable results, [16] criticizes the high level of manual effort that is necessary to train machine learning approaches to reach good classification and extraction rates. Future research should attend to mechanisms that allow learning from very few examples. The best-case scenario will be a one-shot learning, whereby only one example of the same class will be sufficient for learning to classify and extract other documents of that class.

[17, 18] did first empirical studies on the field of few-exemplar learning. The authors compare different algorithms from Weka[4] using changing sizes of training documents. While both works are a first step towards the improvement of few-exemplar learning algorithms, they only focus on general implementations. In context of document classification and extraction only a small number of related works evaluate their approaches according to the ability to learn from few examples. [5, 8] use a minimum of ten documents per class as a training set for classification. As this size is almost too high for our requirements, a user would not accept to correct ten or more documents per class or template to reach good results, [10, 7] present approaches with one and two training documents per class. Both solutions rely on low-level layout features, which makes it hard to differentiate between very similar types of templates. [13] used only ten documents for training and reach impressive extraction rates of 92%. While this approach focuses on multiple values and uses repeating structures within a document, it is not fully comparable to our goal of single value extraction.

[1] http://www.abbyy.com
[2] http://www.insiders-technologies.com
[3] http://www.opentext.com
[4] http://www.cs.waikato.ac.nz/ml/weka

Altogether we could not find a solution that completely fulfills our requirements for adaptive and fast-learning document processing. Although some solutions were tested according to small training sets, there currently exists no solution that provides a real one-shot learning within this domain.

3 Research Hypotheses

The proposed system is based on following hypotheses:

Hypotheses 1 - Self-learning template detection: Because of the clear structure of business documents and company workflows, self-learning systems can instantly and successfully classify documents according to their template with classification rates above 95% F1-measure. In case of using self-learning approaches users will be relieved by the abolition of manual configuration.

Hypotheses 2 - Few-exemplar extraction: The usage and combination of specialized algorithms allow comparative extraction results and an improvement of learning speed especially in the starting period of an extraction system. This reduces the need for feedback and enhances the user acceptance. Altogether we want to reach a one-shot learning ensuring common extraction rates upon 80%.

Hypotheses 3 - Scalability of distributed knowledge pools: The usage of distributed knowledge pools and the combination of foreign and self-generated extraction results can improve the performance of local systems. This influence decreases with the number of training documents in the distributed knowledge pool because of a missing differentiation between similar documents. A fragmentation and specialization of the distributed knowledge pool into a hierarchical system compensates this effect.

4 Methods

To prove the hypotheses described in the previous section, we developed an information extraction process containing the single steps that can be seen in Figure 1. Until today we already implemented a prototype according to this process. Unfortunately the usage of distributed knowledge bases is not yet integrated. Nevertheless first evaluation results of template detection and few-exemplar extraction are presented in Section 5.

Template Detection: While relevant information within a document are highly dependent on the document's template, the first step of our methodology is a classification according to the template, the document was built on. Starting with such a model, the creator of a document adds relevant index data and gets a new document instance. Due to corporate identity and standardizations, companies are influenced to produce their business documents in a consistent manner. A first analysis on our set of real-world business correspondences has

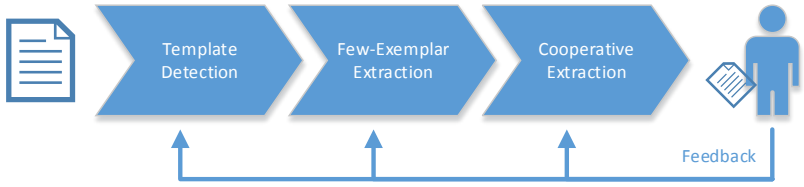

Fig. 1. The information extraction process.

shown that 97% of all documents were generated in such a way. The goal of our template detection is the classification of a document according to its template without configuration effort. Therefore we use an adapted feature generation in combination with a kNN algorithm to label a document with the template used by its nearest neighbors. To ensure fast classification behavior, we avoid to analyze the structure of a document, but rather use only word tokens and their coordinates within the documents as features. The usage of an instance-based algorithm ensures an online processing of documents and a nearly instant integration of user feedback, which either can be acknowledgments on system results or corrections for wrong classification and extraction decisions. Details on our template detection so far can be found in [19].

Few-Exemplar Extraction: The extraction of index data is mainly based on templates which were identified in the previous step. The existence of templates indicates a clear structure within documents generated out of it. This structure can be used to extract relevant information. For illustration, Figure 2 shows three instances created on top of the same template. Relevant information like sender, customer identifier, and amount share nearly the same positions within each document. Based on the information from template detection, we identify documents using the same template within the training set. On this similar looking documents, three types of extraction steps are currently executed. First, we try to identify fields with similar value across the set of similar looking documents which often apply for document type and sender. Second, we try to identify index fields with values at nearly the same position using bounding boxes. Third, we use context words in the surrounding of tagged fields in the training documents. Altogether each of these algorithms needs at least one document of the same template within the training set to produce extraction results. The combination of template detection, usage of similar documents and instant integration of feedback guarantees a very low number of training documents and allows a high adaptivity for new document types, templates and fields via user feedback.

Cooperative Extraction: To further reduce the amount of manual annotations and improve the performance of our system especially in the starting period, we focus on a cooperative approach. Therefore local knowledge pools are defined, where extraction knowledge from each individual user can be securely stored. Each of them is connected to a common distributed knowledge pool. Depending

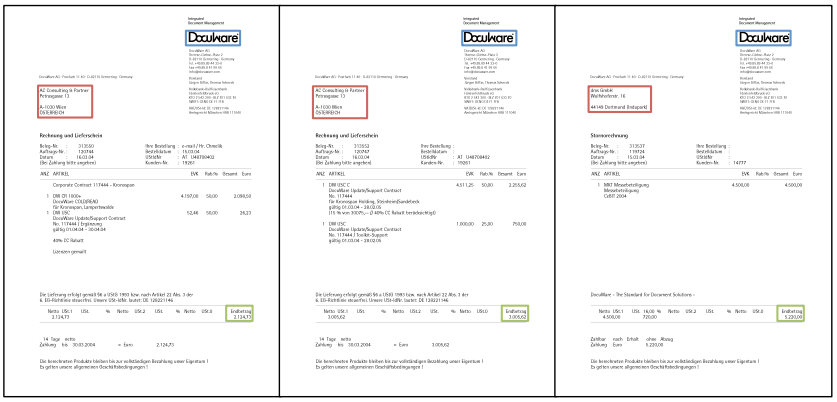

Fig. 2. Business documents using the same template. Bounding boxes show the nearly constant positions of index data over all documents.

on the quality of local results, a document is forwarded to the common pool for extraction. Afterwards distributed and local results will be combined and delivered back to the user. In order to enhance the effectivity of the common knowledge base, feedback given by the user will also be passed. For customization purposes the number of common knowledge pools and possible communication paths can be adopted to business domains and workflows.

Security and privacy issues play an important role within a distributed approach. Although they are not the main aspect of our work, we want to discuss solutions like obfuscation of documents and feedback.

Altogether the cooperation approach limits the amount of feedback by avoiding the manual tagging of documents that already have been seen and tagged by another user. While first results indicate an improvement of systems performance in the starting period, we expect limitations of this approach with a larger amount of training documents within the distributed knowledge pool. Coupling thousands of local extraction systems to interact with a distributed knowledge base will highly increase the number of documents within this pool. By implementing such a scenario, we plan to identify limits of this kind of distributed approaches and give solutions for solving such problems, i.e. finding thresholds for splitting decision and dividing a distributed knowledge pool into several specialized parts, communicating with each other to ensure constantly high extraction rates.

5 Evaluation

Our document corpus consists of 12,500 real-world business documents from the archive of our project partner DocuWare. Due to international business relations, our corpus includes German, English, and Spanish documents. We captured each with a customary scanner and tagged and corrected it according to commonly

used fields in document archiving. Beside a minimal set of fields to enable structured archiving (*document type*, *recipient*, *sender*, *date*), we added further popular fields like *amount*, *contact*, *customer identifier*, *document number*, *subject*, and *date to be paid* based on an inedited survey carried out by DocuWare.

To evaluate our system we use common metrics precision, recall and F1-measure. For classification we rely on the definition by Sebastiani [20]. For extraction we evaluate according to the metrics presented by Chinchor and Sundheim [21] for MUC-5. As the user expects only correct results, we ignore error class "partial" and tackle this kind of extractions as wrong. Overall values are calculated using a micro-averaging approach by averaging single results over all recognized labels.

For evaluation of learning behavior and speed we use an iterative testing procedure. To underline the highly adaptive character of our approach, we test our solutions with what we call "cold start metrics", i.e. a gold standard evaluation starting with an empty learning model and adding each document not recognized correctly as a training example. The system performance is evaluated depending on the current size and structure of the training set. By calculating the area under this curve, we get a single value, which represents the learning speed and allows a comparison between algorithms.

To test the performance of our template detection, we evaluated it against the document set of 12,500 documents using our "cold start metrics" approach. Therefore we manually tagged each document according to the template that was used for creation. Altogether we identified 399 different templates within our document set. The first prototype of our template detection reaches 95% F1-measure, which is already comparable to the state of the art.

Afterwards we tested our local extraction algorithms against our document set. Again we used our iterative "cold start metrics" approach. The overall and field-by-field results are shown in Figure 3. The dotted line represents the minimal rate a user reaches doing manual index data extraction. [22] identified by interviewing several companies an error rate up to 20% for manual indexing. As one can see, our overall result for extraction reaches 87%.

To test the learning behavior of our system, we determined for each processed document the number of training examples with the same template in the current training set and calculated our evaluation metrics upon this number. Figure 4 shows the results. Having only one document of the same template within the training set, our system produces overall extraction results upon 74%. With three documents of the same template we already pass the threshold of 80%.

6 Discussion and Future Work

Our first evaluations have shown the ability of our system to reach good results on basis of a minimal set of training documents. Although we pass 80% F1-measure with three documents of the same template within our training set, a user will get frustrated, if he has to provide feedback for documents of a new template multiple times. Therefore we have to increase the learning speed

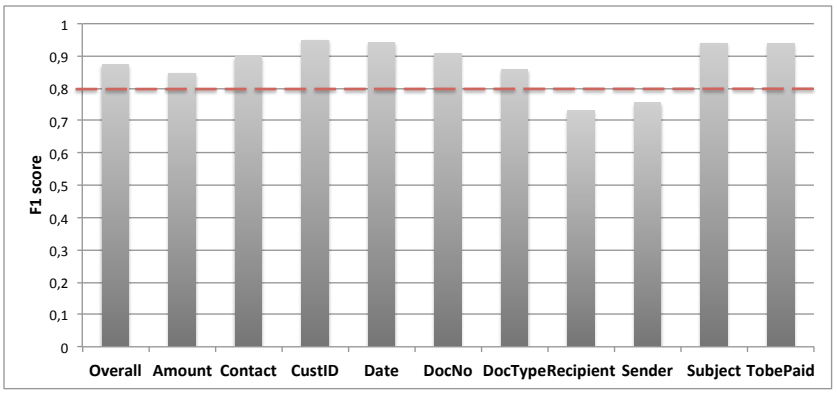

Fig. 3. Overall and field-by-field extraction results of our local extraction approach.

furthermore to reach this limit with only one similar document or in best case with no similar document in training (see shaded area in Figure 4).

A first step into this direction is the further improvement of our template detection and extraction algorithms. While they are not yet perfect, we want to analyze typical error cases, i.e. extraction and OCR errors, and improve the performance of the local system. Possible solutions are the optimization of our feature generation and modifications on our instance-based learning algorithms. Even without collaboration we want to reach a level in learning speed that lies above existing solutions.

A second step is the integration of our distributed approach. We expect to get a much higher performance in the starting period by coupling local systems following our cooperative approach. Ideally the threshold of 80% F1-measure will often already be passed without any similar document in the local training set, only by using extraction knowledge from other participants. Therefore we plan to repeat our evaluations for learning speed using the proposed distributed approach.

To get significant results for scalability, we also have to enlarge the size of our document set. While 12,500 documents are enough for local system evaluation, we want to detect the influence of very large sets of training documents in a distributed knowledge pool according to our cooperative extraction approach. Therefore we plan to reach a much higher number of business documents (i.e. 1,000,000) by generating new ones based on our current document set.

7 Conclusion

In this paper we propose a distributed information extraction system to identify index terms from scanned business documents using a highly adaptable and fast-learning approach that needs nearly no effort in on-site configuration. For that reason it is especially suitable for SOHO users and private individuals.

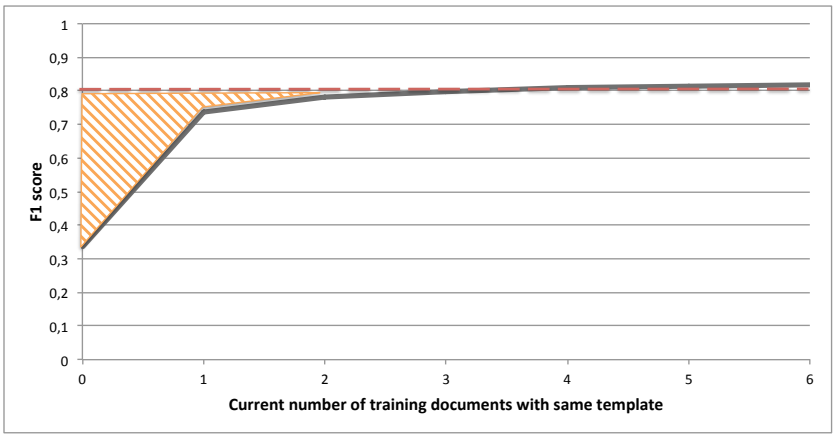

Fig. 4. Learning behavior of our local extraction approach in relation to the number of documents with same template within the current training set.

The combination of local and distributed knowledge pools improves the learning behavior of local systems and enhances the user acceptance. First evaluations on our local system have proven the ability of our methodology to reach acceptable extraction results with a minimal amount of training documents. Further work will mainly focus on integration and evaluation of our distributed approach.

Acknowledgments

This work is a result of my research efforts in the context of the ModelSpace project, which was funded by the German Federal Ministry of Education and Research (BMBF) within the research program "KMU Innovativ" (fund number 01/S12017). My special thanks to my colleagues for insightful discussions and careful prove reading and our project partners from DocuWare for providing us with the document corpus used for evaluation.

References

1. B. Klein, A. Dengel, and A. Fordan, "smartfix: An adaptive system for document analysis and understanding," *Reading and Learning*, pp. 166–186, 2004.
2. Opentext, "Opentext capture center." http://www.opentext.com/2/global/products/products-capture-and-imaging/products-opentext-capture-center.htm, 2012.
3. F. Schulz, M. Ebbecke, M. Gillmann, B. Adrian, S. Agne, and A. Dengel, "Seizing the treasure: Transferring knowledge in invoice analysis," in *10th International Conference on Document Analysis and Recognition, 2009.*, pp. 848–852, 2009.
4. H. Sako, M. Seki, N. Furukawa, H. Ikeda, and A. Imaizumi, "Form reading based on form-type identification and form-data recognition," in *Seventh International Conference on Document Analysis and Recognition, 2003.*, pp. 926–930, 2003.

5. E. Appiani, F. Cesarini, A. M. Colla, M. Diligenti, M. Gori, S. Marinai, and G. Soda, "Automatic document classification and indexing in high-volume applications," *International Journal on Document Analysis and Recognition*, vol. 4, no. 2, pp. 69–83, 2001.
6. E. Sorio, A. Bartoli, G. Davanzo, and E. Medvet, "Open world classification of printed invoices," in *Proceedings of the 10th ACM symposium on Document engineering*, DocEng '10, (New York, NY, USA), pp. 187–190, ACM, 2010.
7. M. Vila, A. Bardera, M. Feixas, and M. Sbert, "Tsallis mutual information for document classification," *Entropy*, vol. 13, no. 9, pp. 1694–1707, 2011.
8. M. Diligenti, P. Frasconi, and M. Gori, "Hidden tree markov models for document image classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 4, pp. 519–523, 2003.
9. A. Belaïd, V. Poulain D'Andecy, H. Hamza, and Y. Belaïd, "Administrative Document Analysis and Structure," in *Learning Structure and Schemas from Documents* (M. Biba and F. Xhafa, eds.), vol. 375 of *Studies in Computational Intelligence*, pp. 51–72, Springer Verlag, Mar. 2011.
10. C. Alippi, F. Pessina, and M. Roveri, "An adaptive system for automatic invoice-documents classification," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2, pp. II–526–9, 2005.
11. S. Adali, A. C. Sonmez, and M. Gokturk, "An integrated architecture for processing business documents in turkish," in *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '09)*, 2009.
12. F. Cesarini, E. Francesconi, M. Gori, and G. Soda, "Analysis and understanding of multi-class invoices," *IJDAR*, vol. 6, no. 2, pp. 102–114, 2003.
13. E. Bart and P. Sarkar, "Information extraction by finding repeated structure," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pp. 175–182, 2010.
14. Y. Belaid and A. Belaid, "Morphological tagging approach in document analysis of invoices," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004.
15. L. Likforman-Sulem, P. Vaillant, and F. Yvon, "Proper names extraction from fax images combining textual and image features," in *Seventh International Conference on Document Analysis and Recognition*, pp. 545–549 vol.1, 2003.
16. E. Saund, "Scientific challenges underlying production document processing," in *Document Recognition and Retrieval XVIII (DRR)*, 2011.
17. C. Salperwyck and V. Lemaire, "Learning with few examples: An empirical study on leading classifiers," in *The International Joint Conference on Neural Networks (IJCNN)*, 2011.
18. G. Forman and I. Cohen, "Learning from little: Comparison of classifiers given little training," in *Knowledge Discovery in Databases: PKDD 2004*, pp. 161–172, Springer, 2004.
19. D. Esser, D. Schuster, K. Muthmann, M. Berger, and A. Schill, "Automatic indexing of scanned documents - a layout-based approach," in *Document Recognition and Retrieval XIX (DRR)*, (San Francisco, CA, USA), 2012.
20. F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.
21. N. Chinchor and B. Sundheim, "Muc-5 evaluation metrics," in *Proceedings of the 5th conference on Message understanding*, MUC5 '93, pp. 69–78, 1993.
22. B. Klein, S. Agne, and A. Dengel, "Results of a study on invoice-reading systems in germany.," in *Document Analysis Systems*, 2004.