***smartFIX*: A Requirements-Driven System for Document Analysis and Understanding**

Andreas R. Dengel and Bertin Klein

German Research Center for Artificial Intelligence (DFKI)
P.O.Box 2080
D-67608 Kaiserslautern, Germany
{dengel,klein}@dfki.de

Abstract. Although the internet offers a wide-spread platform for information interchange, day-to-day work in large companies still means the processing of tens of thousands of printed documents every day. This paper presents the system *smartFIX* which is a document analysis and understanding system developed by the DFKI spin-off INSIDERS. It permits the processing of documents ranging from fixed format forms to unstructured letters of any format. Apart from the architecture, the main components and system characteristics, we also show some results when applying *smartFIX* to medical bills and prescriptions.

1 Introduction

About 1.2 million printed medical bills arrive at the 35 German private health insurance companies every day. Those bills amount to 10% of the German health insurance market and they are actually maintained by printed paper bills. Figure 1 shows examples of such printed bills. Until recently the processing of these bills was done almost completely manually. In addition to the tedious task of initiating every single payment by transcribing a number of data fields from varying locations on the paper documents into a computer, this had the serious disadvantage that only a small number of inconsistent and overpriced bills were discovered. Conservative estimates predict savings in the range of several hundred million Euros each year if this process could be automated reliably.

About two years ago, a consortium of German private health insurance companies ran a public benchmark test of systems for Document Analysis and Understanding (DAU) for the respective private insurance sector. The benchmark was clearly won by *smartFIX* (smart For Information eXtraction). *smartFIX* was developed by a spin-off company of the DAU group at DFKI, INSIDERS (www.insiders-ag.de • founded in 1999 by the first author), thinking that the available DAU technology after a decade of focused research was ready to construct a versatile and adaptive DAU system [1, 2, 3]. After the premise of the identification of a viable type of application scenarios, the well-directed research on a feasible combination of available methods and completion with new methods could be successfully accomplished. This background probably explains the clear suitability of *smartFIX* for the project, which INSIDERS and four insurance companies started after the benchmark: the development of a standard product for the analysis of printed medical bills: *smartFIX healthcare*. *smartFIX* was

the result of several man-years investment. The brainpower in *smartFIX healthcare* can be estimated in a larger dimension.

Prof. Dr. med. Anja Beck 75285 Ebrinnen

Kinderärztin
Rechenbuchnummer
Bundesbank BIC: 55120333
Konto 1554520

Rechnung vom 15.12.1954

Patient: Schmidt, Klaus-Peter, geb. 15.12.1954

Für ärztliche Beratungen
einschließlich DM 136,36 zu berechnen.

Datum	QNR	Factor	Gesamt	Gebühr
14.05.98	1	2,300	20,98	
14.05.98	2	2,300	20,98	
14.05.98	3	2,300	20,98	
14.05.98	4	2,300	20,98	
14.05.98	5	2,300	20,98	
14.05.98	6	2,300	20,98	
14.05.98	7	2,300	20,98	
14.05.98	8	2,300	20,98	
14.05.98	9	2,300	20,98	
14.05.98	10	2,300	20,98	
14.05.98	11	2,300	20,98	
14.05.98	12	2,300	20,98	
14.05.98	13	2,300	20,98	
14.05.98	14	2,300	20,98	
14.05.98	15	2,300	20,98	
14.05.98	16	2,300	20,98	
14.05.98	17	2,300	20,98	
14.05.98	18	2,300	20,98	
14.05.98	19	2,300	20,98	
14.05.98	20	2,300	20,98	
14.05.98	21	2,300	20,98	
14.05.98	22	2,300	20,98	
14.05.98	23	2,300	20,98	
14.05.98	24	2,300	20,98	
14.05.98	25	2,300	20,98	
14.05.98	26	2,300	20,98	
14.05.98	27	2,300	20,98	
14.05.98	28	2,300	20,98	
14.05.98	29	2,300	20,98	
14.05.98	30	2,300	20,98	
14.05.98	31	2,300	20,98	
14.05.98	32	2,300	20,98	
14.05.98	33	2,300	20,98	
14.05.98	34	2,300	20,98	
14.05.98	35	2,300	20,98	
14.05.98	36	2,300	20,98	
14.05.98	37	2,300	20,98	
14.05.98	38	2,300	20,98	
14.05.98	39	2,300	20,98	
14.05.98	40	2,300	20,98	
14.05.98	41	2,300	20,98	
14.05.98	42	2,300	20,98	
14.05.98	43	2,300	20,98	
14.05.98	44	2,300	20,98	
14.05.98	45	2,300	20,98	
14.05.98	46	2,300	20,98	
14.05.98	47	2,300	20,98	
14.05.98	48	2,300	20,98	
14.05.98	49	2,300	20,98	
14.05.98	50	2,300	20,98	
14.05.98	51	2,300	20,98	
14.05.98	52	2,300	20,98	
14.05.98	53	2,300	20,98	
14.05.98	54	2,300	20,98	
14.05.98	55	2,300	20,98	
14.05.98	56	2,300	20,98	
14.05.98	57	2,300	20,98	
14.05.98	58	2,300	20,98	
14.05.98	59	2,300	20,98	
14.05.98	60	2,300	20,98	
14.05.98	61	2,300	20,98	
14.05.98	62	2,300	20,98	
14.05.98	63	2,300	20,98	
14.05.98	64	2,300	20,98	
14.05.98	65	2,300	20,98	
14.05.98	66	2,300	20,98	
14.05.98	67	2,300	20,98	
14.05.98	68	2,300	20,98	
14.05.98	69	2,300	20,98	
14.05.98	70	2,300	20,98	
14.05.98	71	2,300	20,98	
14.05.98	72	2,300	20,98	
14.05.98	73	2,300	20,98	
14.05.98	74	2,300	20,98	
14.05.98	75	2,300	20,98	
14.05.98	76	2,300	20,98	
14.05.98	77	2,300	20,98	
14.05.98	78	2,300	20,98	
14.05.98	79	2,300	20,98	
14.05.98	80	2,300	20,98	
14.05.98	81	2,300	20,98	
14.05.98	82	2,300	20,98	
14.05.98	83	2,300	20,98	
14.05.98	84	2,300	20,98	
14.05.98	85	2,300	20,98	
14.05.98	86	2,300	20,98	
14.05.98	87	2,300	20,98	
14.05.98	88	2,300	20,98	
14.05.98	89	2,300	20,98	
14.05.98	90	2,300	20,98	
14.05.98	91	2,300	20,98	
14.05.98	92	2,300	20,98	
14.05.98	93	2,300	20,98	
14.05.98	94	2,300	20,98	
14.05.98	95	2,300	20,98	
14.05.98	96	2,300	20,98	
14.05.98	97	2,300	20,98	
14.05.98	98	2,300	20,98	
14.05.98	99	2,300	20,98	
14.05.98	100	2,300	20,98	
14.05.98	101	2,300	20,98	
14.05.98	102	2,300	20,98	
14.05.98	103	2,300	20,98	
14.05.98	104	2,300	20,98	
14.05.98	105	2,300	20,98	
14.05.98	106	2,300	20,98	
14.05.98	107	2,300	20,98	
14.05.98	108	2,300	20,98	
14.05.98	109	2,300	20,98	
14.05.98	110	2,300	20,98	
14.05.98	111	2,300	20,98	
14.05.98	112	2,300	20,98	
14.05.98	113	2,300	20,98	
14.05.98	114	2,300	20,98	
14.05.98	115	2,300	20,98	
14.05.98	116	2,300	20,98	
14.05.98	117	2,300	20,98	
14.05.98	118	2,300	20,98	
14.05.98	119	2,300	20,98	
14.05.98	120	2,300	20,98	
14.05.98	121	2,300	20,98	
14.05.98	122	2,300	20,98	
14.05.98	123	2,300	20,98	
14.05.98	124	2,300	20,98	
14.05.98	125	2,300	20,98	
14.05.98	126	2,300	20,98	
14.05.98	127	2,300	20,98	
14.05.98	128	2,300	20,98	
14.05.98	129	2,300	20,98	
14.05.98	130	2,300	20,98	
14.05.98	131	2,300	20,98	
14.05.98	132	2,300	20,98	
14.05.98	133	2,300	20,98	
14.05.98	134	2,300	20,98	
14.05.98	135	2,300	20,98	
14.05.98	136	2,300	20,98	
14.05.98	137	2,300	20,98	
14.05.98	138	2,300	20,98	
14.05.98	139	2,300	20,98	
14.05.98	140	2,300	20,98	
14.05.98	141	2,300	20,98	
14.05.98	142	2,300	20,98	
14.05.98	143	2,300	20,98	
14.05.98	144	2,300	20,98	
14.05.98	145	2,300	20,98	
14.05.98	146	2,300	20,98	
14.05.98	147	2,300	20,98	
14.05.98	148	2,300	20,98	
14.05.98	149	2,300	20,98	
14.05.98	150	2,300	20,98	
14.05.98	151	2,300	20,98	
14.05.98	152	2,300	20,98	
14.05.98	153	2,300	20,98	
14.05.98	154	2,300	20,98	
14.05.98	155	2,300	20,98	
14.05.98	156	2,300	20,98	
14.05.98	157	2,300	20,98	
14.05.98	158	2,300	20,98	
14.05.98	159	2,300	20,98	
14.05.98	160	2,300	20,98	
14.05.98	161	2,300	20,98	
14.05.98	162	2,300	20,98	
14.05.98	163	2,300	20,98	
14.05.98	164	2,300	20,98	
14.05.98	165	2,300	20,98	
14.05.98	166	2,300	20,98	
14.05.98	167	2,300	20,98	
14.05.98	168	2,300	20,98	
14.05.98	169	2,300	20,98	
14.05.98	170	2,300	20,98	
14.05.98	171	2,300	20,98	
14.05.98	172	2,300	20,98	
14.05.98	173	2,300	20,98	
14.05.98	174	2,300	20,98	
14.05.98	175	2,300	20,98	
14.05.98	176	2,300	20,98	
14.05.98	177	2,300	20,98	
14.05.98	178	2,300	20,98	
14.05.98	179	2,300	20,98	
14.05.98	180	2,300	20,98	
14.05.98	181	2,300	20,98	
14.05.98	182	2,300	20,98	
14.05.98	183	2,300	20,98	
14.05.98	184	2,300	20,98	
14.05.98	185	2,300	20,98	
14.05.98	186	2,300	20,98	
14.05.98	187	2,300	20,98	
14.05.98	188	2,300	20,98	
14.05.98	189	2,300	20,98	
14.05.98	190	2,300	20,98	
14.05.98	191	2,300	20,98	
14.05.98	192	2,300	20,98	
14.05.98	193	2,300	20,98	
14.05.98	194	2,300	20,98	
14.05.98	195	2,300	20,98	
14.05.98	196	2,300	20,98	
14.05.98	197	2,300	20,98	
14.05.98	198	2,300	20,98	
14.05.98	199	2,300	20,98	
14.05.98	200	2,300	20,98	
14.05.98	201	2,300	20,98	
14.05.98	202	2,300	20,98	
14.05.98	203	2,300	20,98	
14.05.98	204	2,300	20,98	
14.05.98	205	2,300	20,98	
14.05.98	206	2,300	20,98	
14.05.98	207	2,300	20,98	
14.05.98	208	2,300	20,98	
14.05.98	209	2,300	20,98	
14.05.98	210	2,300	20,98	
14.05.98	211	2,300	20,98	
14.05.98				

smartFIX were refined with the needs of the insurance companies. Actually, many of the requirements only became clearer to us whilst the project progressed. The insurance companies required from *smartFIX healthcare*:

1. Verified economic advantages (qualitative, quantitative)
2. Economic reliability (error recovery, error rate, error statistics)
3. Scalability
4. Adaptability to their workflow (people, tasks, databases, archives)

The advantages remained to be verified. Scalability had already been targeted and was almost available: *smartFIX* is a distributed system, CORBA on networked PCs, which can easily spread its single-analysis processes over many CPUs and still be controlled from one desktop. One aspect of reliability is achieved with a central transaction memory database. Thus, the only remaining tasks were to extend and adapt *smartFIX* to the insurers workflow requirements and to learn about reliability in terms of stable error-rates.

For the design of a DAU system we came to the following guiding principles. But still the most basic finding is that the design should not be technology-driven, but explicitly requirements-driven, and thus finally user-driven [4]:

1. Compositionality: A versatile system has to have a significant spectrum of different basic DAU components, one paradigm alone will almost surely fail to solve a real problem. – Complex methods, which at first glance are very hard to conceive, are obtained from deliberate, small combinations of simple methods. [3]
2. Practice principle: There is no way to get around errors. Thus it is important to help users to discover, judge, and either tolerate, or to correct errors. This is especially critical as the user in every day practice is left alone with the control of the system and the responsibility for its results. – The real DAU technology must be powerfully accompanied by technology for logging, tracing, visualizing, interactive testing, statistics.
3. Epistemological adequacy: The basic DAU components must be bundled to greater analysis scripts which are made available to the user. The user has to perceive document characteristics and map those to the scripts. Thus the scripts must be meaningful to the user, easy to memorize and use. Good script metaphors also guide the perception of the document characteristics. Later these scripts might report success or error messages. – This long standing AI principle implies that so-called “syntactic sugar” can actually make a difference.
4. Constructivism: The philosophic principle of constructivism says that every knowledgeable agent, human or machine, has a different account of reality and there is no objective truth at all. For a DAU system it means: (a) That every scenario will always require at least one aspect, which is outside of the capabilities of the system at hand. Thus it is important to allow users access to manipulate intermediate results. (b) That two persons at a time (and even often one person at two times) disagree on facts significantly often (cf. TREC human evaluators cross comparison [5]). Thus it is important that the DAU system gives feedback and continuously makes transparent, which information it receives from a user and how this information is used.

In general, *smartFIX* is not limited to the domain of medical bills but rather can be applied to other forms and unstructured documents as well. In the following, we will

describe the major characteristics of the system including architecture, main components as well as some technical aspects. At the end of the paper, we will show runtime results of *smartFIX* PKV applied to medical bills based on an recent evaluation of a large private health insurance company processing several tens of thousands of bills and prescriptions every day.

2 System Architecture

The capabilities of *smartFIX* PKV are diverse, and so is its architecture. The architecture, i.e. the main components, are shown in Figure 2. This section's overview is succeeded in the following sections by a sketch of the *DocumentManager* through which the system is instructed and after that an explanation of the most central DAU component, the *analyzer*.

To apply *smartFIX* PKV to a new DAU problem, it is necessary to pin down the specific document types and the “what and how” of information extraction. The resulting “document information” –one synonym is “processing knowledge”– is configured with the *DocumentManager*.

Fig. 2. System architecture of *smartFix*

Then the system can be started with the *coordinator*. This program provides a control panel, from which DAU-processes, i.e. *analyzers*, can be started on a freely-configurable network of PCs; one *analyzer* per CPU as a rule of thumb. The *coordinator* also starts some other processes, first now, two database processes used by the analyzers: the *Matching-Server* provides very fast retrieval (“matching”) on company knowledge from the *matching data base*. The *matching data base*, mostly contractual data, is a working copy, which can be updated with a handy tool, the *matching data base importer*, from company databases. The *control data base* is the central working memory of *smartFIX*. It could be called the blackboard of the whole architecture.

Fig. 3. The *DocumentManager*

The *coordinator* controls an *importer* process, which transfers document images from a predefined source into the *control data base*, together with possibly known information. Any idling *analyzer* will check out, process and check in again documents in the *control data base*. Successful information extraction provided, human agents who are logged on with a *verifier* process, are prompted the DAU results. With the database it is assured, that not a single bit is lost, even in the event of a power cut. Finished documents are transferred out of the system by an *exporter*, typically into archive systems.

This is the main architecture of *smartFIX*. There are some more components, many of which implement much of the user-driven principle, but space prohibits really elaborating on them. The complicated information for spotting and extracting tables is handled by a *TableTool* which follows the approach we proposed in [10] to keep the *DocumentManager* simpler. *Verifiers* can have different priorities and rights assigned and can log the user-actions. They provide online access to the fast databases, aiding the manual completion of missing data. Amongst other things, the freely configurable workflow logic allows intermediate processing results to be externalized with a *user*

exit process. The format is configurable, mostly XML. The intermediate results can thus be checked and changed with external software and then be reintegrated into regular processing. The workflow also allows for tricky DAU processes, where, for example, documents are analyzed to a first level and in a second run from the first level to a second level. Online and offline, a number of tools enable the overview and tracing the status, progress and statistics of the involved processes.

3 DocumentManager

Before using *smartFIX*, the system has to be trained. The *DocumentManager* is a special editor with which the user teaches the document classes and appropriate reference patterns are stored in a document definitions knowledge base. The *DocumentManager* is composed of five windows shown in Figure 3. On the right is a user-chosen sample document image, on the left:

- The directory of the different document classes and their aggregations
- The classification features and appropriate types
- The labels and the type of logical objects which capture the information to be extracted and verify if they occur in unknown documents
- So-called *Label Object Groups (LOG)* and their corresponding type

A *LOG* represents a named set of scripts which depend on each other, e.g. a LOG `<SERVICE SUBJECT>` could be composed of `<insurance number>`, `<person>`, and `<address>`, while the latter two again consist of `<first name>`, `<last name>` as well as `<zip code>`, `<city>`, and `<street name>`.

In order to teach *smartFIX*, example documents are loaded. In a first step, the user defines and links the corresponding document classes. He or she may select a region in the image around a logical object of interest. This region of interest (ROI) can be described by the mouse or the entire page if it is not possible to narrow down the position. To name the logical object, the user may either select the label from the existing list in the window or could add a new label. In addition, the user has to define the corresponding type which addresses the analysis agent to be applied for extracting the contained information. For that purpose, *smartFIX* offers various agents. The most significant are:

- ADDRESS describes an appropriate grammar of how German addresses typically look like.
- CHECK BOX
- BAR CODE
- SEARCH PATTERN addresses a *key string* (including synonyms and acronyms from an associated thesaurus) to which a corresponding *value string* has to be found. For example (see Fig. 4), the value of the logical object `<bill no.>` (in German `<rechnungs-nr.>`) is a numerical expression located within a certain word distance in the image.
- ASSOCIATION relates to information which might not be in a document but rather stored in a data base.
- TABLE allows the description of the logical structure of tables including optional columns.

- TEXT addresses logical objects which can be defined as regular expressions and are captured within free text, e.g. “my insurance number is BG/1495-H”.

All logical objects which are defined are related to ranges of values, e.g. lexical knowledge, regular expressions, or numerical constraints, depending on their type. For example, each column of a table is linked to the corresponding set of values, i.e. the logical object `<product>` which represents a table column name is linked with all valid product names in a data base.

Another issue which holds for all names of logical objects is the fact that they are all only representatives, i.e. variables with possible instantiations. Instantiations may be synonyms and acronyms stored in an attached thesaurus, e.g.,

```
<total amount>=
{ "amount", "sum", .. } := NUM{1-5} I.I NUM{2} I { "DM", "€" }
<product> =
{ "artefact", "item", .. } := ref [product data base/names]
```

After defining all logical objects and the corresponding ROIs, it is possible to support the analysis further by defining LOGs. Consequently, the reference pattern is stored in a knowledge base for later analysis.

Fig. 4. Typical constellation of numerical data close to a meaningful keyword.

4 Analyzer

The *analyzer* processes do the real DAU processing. Every running *analyzer* pulls the *control data base* for documents to process. After finishing with a document it is labeled and put back into the *control data base* (the label is like the address of the next necessary processing stage, e.g. *verifier*, and the labels flexibly implement the internal workflow).

For one page the analysis progresses in roughly four stages:

1. Image pre-processing,
2. Classification,
3. Information extraction,
4. Improvement.

OCR is called lazily, i.e. it is called for segments only when their character content is needed. As a matter of fact, the four phases are not always clearly distinguished. E.g. after the classification it is possible to have found that the image must be scaled.

4.1 Image Pre-processing

Image pre-processing consists of five optional steps. Each step can be individually switched on and off and tuned with parameters:

1. Removal of graphical lines (vertical and horizontal)
2. Optimization of matrix printed characters
3. Despeckle
4. Deskew ($\pm 10^\circ$ landscape and portrait)
5. Correction of upside-down

The pre-processed images are stored for the later visualization of the extraction results (with *verifiers*).

4.2 Classification

The classification runs in predefined phases and searches the following features until a class is determined:

1. Layout similarity [7, 8]
2. Recognized tables [6]
3. User-defined or machine-learned patterns (including Not-patterns)
4. Machine-learned semantic similarity [11]
5. Document size

The known classes and their features are configured using the *DocumentManager*. The classes can be set up so that non-classified documents are sent to a *verifier* for manual classification, or run through a default information extraction first.

The classification also has to deal with the collation of pages to documents. The most reliable method, interspersing different documents with easy-to-recognize dividing pages at scanning, is sometimes not possible. Therefore *smartFIX healthcare* provides a collation-machinery, which is not really sophisticated but confusing to explain without using many examples.

4.3 Information Extraction

Information extraction is performed according to the “scripts” selected and configured for the respective document class with the *DocumentManager*. Scripts have a type and a name. According to their type scripts perform a specific action referring to the page at hand, after which they provide a value under their name. Where this is important, scripts can be assigned priorities to constrain the sequence in which they are run. Simple extraction scripts search the image of the page. The complex extraction scripts use complicated user-provided and built-in knowledge. The data scripts are simply providing access to logical page data. These three categories of script types are explained in the following.

Simple extraction scripts types.

1. “Text” and “Combination”: extract characters in a specified area. The combination script sends the extracted characters later to have it split into components with the support of database lookups. It is possible to have the area initially cleared from form preprints.
2. “Checkbox”: two different algorithms can be chosen, to determine whether a checkbox area is ticked.
3. “Pattern search”: either on the whole page or in an area, a regular expression is either searched directly, or searched in the configurable surroundings (direction and distance) of a prior regular expression match or thesaurus synonyms match.
4. “Address”: a specific script for the extraction of German addresses, split into standard pieces. A regular expression may be used to exclude strings from the search space.
5. “Anchor”: does a pattern search and after matching it triggers other linked scripts, which then run in an area relative to the position of the match.
6. “Format”: currently simply measures and evaluates to A4, A5, or A6.

Value script types. With the choice of one of the following scripts it is possible to access page information: “Document class”: Evaluates to the result of the classification. “Document ID”: evaluates to the unique identifier of the document (which often allows access to further related external archive data). “Task ID”: evaluates to the task which the document belongs to.

A “free value script” does nothing, but can be set later with a value from a database lookup. This might be helpful as a variable for further database lookups, for automatic checks, or as an additional information for human verifiers.

Complex extraction script types. “Association”: skims for occurrences of patterns from a possibly huge database. Several fields of the database can be mixed. Prior successful scripts can be used to restrict the search space of database field combinations. This is a powerful tool e.g. to find exceptionally formatted senders’ addresses, if they are only known in a customer database.

“Fixed table”: extracts a table according to user-defined layout rules.

“Free table”: is a comprehensive analyzer in its own right. According to a complex knowledge configuration with the *TableTool*, this script extracts tables of different layouts, even spanning more than one page, rather intelligently. Extracts also table-related data like sum total. [6]

4.4 Improving

The improving module, newly based on a constraint solver [9], checks consistencies, searches optima, and fills in deducible information. It exploits the user defined *Labeled Object Groups* (LOG), user provided auxiliary conditions and finishes the “Association” scripts. The main idea is that the results of all the single scripts, which are fuzzy and unsafe, can be constrained step by step to more specific and reliable results, with local auxiliary conditions and interdependency conditions. The improving module heavily uses the *matching server* process.

With the local auxiliary conditions the system estimates a level of reliability for every single piece of extracted information, which is used in the *verifier* as colored backgrounds green, light blue, blue, to gain human attention.

LOGs relate different scripts (i.e. their values) together, by a reference to databases. The results of two scripts for a name and a contract number are rather unsafe after simple extraction. But their combination most often allows the determination of the one right name and contract number. Also the “Combination” scripts, introduced before, are processed in this module. In their case, the difference is only, that name and number and perhaps more data are concatenated in one string, possibly permuted.

The “Association” scripts are not much different. Mainly for their calculation it is possible and sometimes necessary to consider their position on the page.

Auxiliary conditions need not only be local, but can also relate two or more script values analytically, in which case they work like LOGs. E.g. all the fields in a summation column sum up to the sum total below the column. Available predicates and functions address arithmetic, dates, checkboxes, strings. Some of the auxiliary conditions can generate correction hypotheses or fill empty fields (e.g. a wrong or empty OCR result in a summation column) others only work as a filter. Via CORBA or COM users can also call user-implemented condition functions (returning a string).

Note: this intelligence is also automatically available in the *verifier*, so that after the typing of missing information, deducible information is provided by the system.

This should be enough explanation on the improving module. There are some more details, like e.g. the role and configuration of thresholds.

5 Results

smartFIX was successfully installed at the insurances’ sites and connected to their workflow. It was configured to distinguish about 50 different types of mostly unstructured documents, like hospital bills, dentist bills, medicament bills etc. All bills and prescriptions capture valuable information for establishing a patient history, for guiding internal processes, or for generating statistics. Information items that *smartFIX* was configured for comprise: (a) insurance number, (b) name of patient, (c) date, (d) service period, (e) diagnosis, (f) medical services table, (g) duration of treatment, (h) name and address of doctor, altogether a number of 107, out of which about 20 are to be found on average depending on the respective document type.

For an evaluation, we considered a representative stack of 525 documents as they arrived at one of the largest health insurance companies in Germany. The test comprised of classification into classes reimbursement form (RF), prescription (P), dentists bills (D), hospital bills (H), bills from medical doctors (M), and others (O) while the latter ones had to be sub-classified into another 20 classes. The following Table 1 shows the classification results:

Table 1. Classification results of *smartFIX* healthcare.

RF	95%	D	100%	M	92%
P	96%	H	94%	O	73%

The classification rate over all classes was 92%. Note that rejects are no problem because they are treated by the user at the verifier. The error rate was less than 0.1%.

In addition, the various ROIs were searched and the captured information extracted. For the evaluation, two aspects were of importance: The extraction rates as well as savings in time. The extraction for all fields (a) to (h) was 81 % and saved an average of 64% of time compared to manual input.

In addition to the results of information extraction, the special modules included in *smartFIX healthcare* verify the correctness of the service information shown on the bill. This means, to do a numerical cross check of single service fees, the so called factors as well as the various amounts shown on the bill. And finally, *smartFIX healthcare* checks whether the service positions shown on each bill fit into the German scale of charges and fees, the so-called GOÄ which is law.

This result is just one example presented to the reader. *smartFIX* runs at more than a dozen of customers sites, with very different configurations. A huge number of very different documents were classified and their information extracted to date. Note for example, that at one company the classification rate went down to 30 % last year. They were happy. Without changing the system configuration, they strained the system with a variety of new and complicated input documents, which could of course not be classified; and which should not be classified. And this is what *smartFIX* did not do, correctly. Statistics are not important, the customers satisfaction is.

6 Summary and Future Work

We have given an overview of a commercial document analysis system, *smartFIX*, which is the result of the exploitation of research results from the DFKI, their non-trivial combination and additional research by the start-up company INSIDERS. The system is able to process stacks of mixed-format documents including forms, invoices, letters, all of which may contain machine-written or hand-printed information. The system is installed at about two dozens of customers analyzing several hundred thousands of document pages every day. The purpose of *smartFIX* is in the distribution (classification) of scanned incoming mail and the extraction of information relevant for the user or customer. *smartFIX* has some special features. It processes unsorted incoming mail of any format and independent of the degree of structure or preprints. Furthermore it is not limited to a small number of sorted and separated document types. *smartFIX* classifies all documents and therefore is independent of pre-sorting. Images are combined automatically into single- or multi-page documents.

Recently, we have been working on some improvements and additional features of *smartFIX*. The first addresses the restructuring of the system into smaller modules which then allow for more flexible mixing, which in turn allows to serve new and more complex application scenarios.

The second addresses a new paradigm for the *DocumentManager* with full access to all *analyzer* features, so that the effect of every single configuration-information can be immediately tried out interactively. This will look and feel like a software engineering environment with full debugging support.

A third important improvement is dedicated to what we call “adaptivity”. “Adaptivity” aims to reduce the human effort to maintain applications and to configure the system for new applications. We work on learning tools and domain assessment tools.

Learning tools aid, because it is simpler to just provide examples, than determine the right features and explicate them to the system. Domain assessment tools aid determination of relevant features, by allowing the evaluation of e.g. the average distance between two consecutive words on 1000 sample documents.

References

1. A. Dengel, R. Bleisinger, R. Hoch, F. Hönes, M. Malburg and F. Fein, OfficeMAID — A System for Automatic Mail Analysis, Interpretation and Delivery, Proceedings DAS94, Int'l Association for Pattern Recognition Workshop on Document Analysis Systems, Kaiserslautern (Oct. 1994), pp. 253–276.
2. S. Baumann, M. Ben Hadj Ali, A. Dengel, T. Jäger, M. Malburg, A. Weigel, C. Wenzel, Message Extraction from Printed Documents A Complete Solution. In: Proc. of the 4.th International Conference on Document Analysis and Recognition (ICDAR), Ulm, Germany, 1997.
3. A. Dengel and K. Hinkelmann, The Specialist Board – a technology workbench for document analysis and understanding. In M. M. Tanik, F.B. Bastani, D. Gibson, and P.J. Fielding, editors, Integrated Design and Process Technology – IDPT96, Proc. of the 2nd World Conference, Austin, TX, USA, 1996.
4. G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. Van de Velde, and B. Wielinga. Knowledge Engineering and Management – The CommonKADS Methodology. The MIT Press, Cambridge, Massachusetts, London, England, 1999.
5. <http://trec.nist.gov/>
6. B. Klein, S. Gökkus, T. Kieninger, A. Dengel, Three Approaches to “Industrial” Table Spotting. In: Proc. of the 6.th International Conference on Document Analysis and Recognition (ICDAR), Seattle, USA, 2001.
7. A. Dengel and F. Dubiel, Computer Understanding of Document Structure, International Journal of Imaging Systems & Technology (IJIST), Special Issue on Document Analysis & Recognition, Vol. 7, No. 4, 1996, pp. 271–278.
8. F. Dubiel and A. Dengel, FormClas — OCR-Free Classification of Forms, in: J.J. Hull, S. Liebowitz (eds.) Document Analysis Systems II, World Scientific Publishing Co. Inc., Singapore, 1998, pp. 189–208.
9. A. Fordan, Constraint Solving over OCR Graphs. In: Proc. Of the 14th International Conference on Applications of Prolog (INAP), Tokyo, Japan, 2001.
10. T. Kieninger and A. Dengel, A Paper-to-HTML Table Converting System, Proceedings DAS98, Int'l Association for Pattern Recognition Workshop on Document Analysis Systems, Nagano, Japan, Nov. 1998, pp. 356–365.
11. M. Junker and A. Dengel, Preventing overfitting in learning text patterns for document categorization, ICAPR2001, 2nd Intern'l Conference on Advances in Pattern Recognition, Rio de Janeiro, Brazil, March 2001.