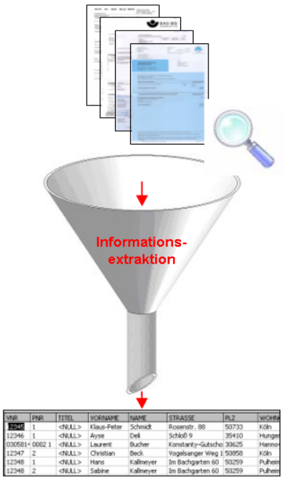
7 Information Extraction - Automated Indexing

Information Extraction



VNR	PNR	TITEL	VORNAME	NAME	STRASSE	PLZ	WOHN
12345	1	<NULL>	Klaus-Peter	Schmidt	Rosenstr. 88	50733	Köln
12346	1	<NULL>	Ayse	Deli	Schloß 9	35410	Hunge
0305814	0002 1	<NULL>	Laurent	Bucher	Konstanty-Gutscho	30625	Hanno
12347	2	<NULL>	Christian	Beck	Vogelsanger Weg 1	50858	Köln
12348	1	<NULL>	Hans	Kallmeyer	Im Bachgarten 60	50259	Pulheim
12348	2	<NULL>	Sabine	Kallmeyer	Im Bachgarten 60	50259	Pulheim

- Information Extraction is the ***automatic identification*** and ***structured representation*** of ***relevant information*** in documents
  - extract well-defined pieces of relevant information from collections of document
  - goal: populate a database (e.g. metadata)
- General Functionality
  - Input
    - Templates coding relevant information, e.g. metadata atributes
    - set of real world texts
  - Output
    - set of instantiated templates filled with relevant text fragments

Application Scenarios for Information Extraction

- Indexing: Creating indexes for information retrieval systems
  - Automated determination of metadata of documents
- Question Answering
  - Answer an arbitrary question by using textual documents as knowledge base
- Mail distribution
  - Identification of recipients in incoming letters of a company
- Converting unstructured text to structured data
  - automatic insertion of data into operative application systems and databases
- Evaluation of surveys
  - Capturing and analysis of questionnaires

Information extraction depends on …

… structural degree of input data

- **structured**: tables with typed data like numbers
- **semi-structured**: XML, tables with text
- **non-structured:** text

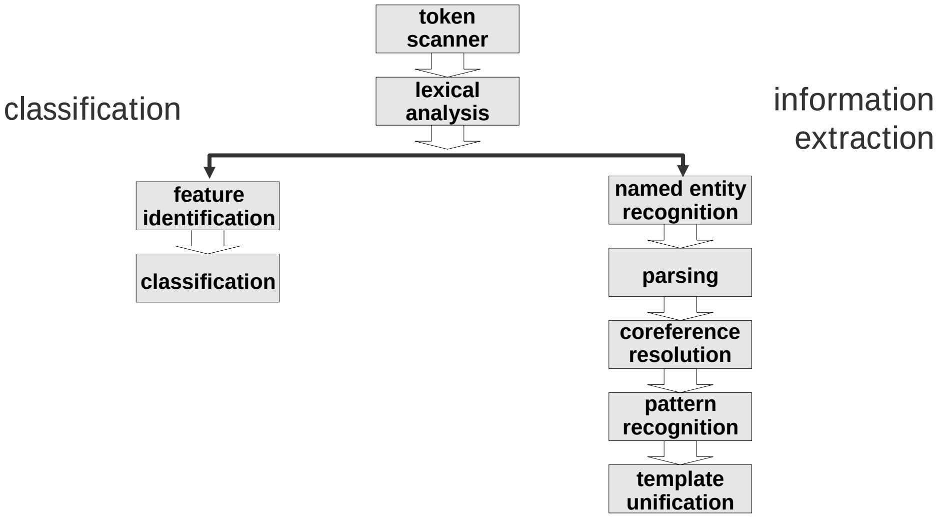
… format

- electronic information
  - coded
  - non-coded
- paper documents

… structural degree of output data

- text summary
- fulltext index
- structured data: database, attributes, classification

7.1 Information Extraction from Text Documents



classification

information extraction

token scanner

lexical analysis

feature identification

classification

named entity recognition

parsing

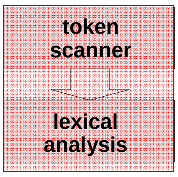
coreference resolution

pattern recognition

template unification

Lexical Analysis

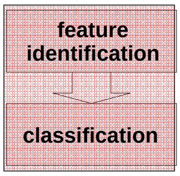
- **Token scanner:**
  - Identification of text structure (e.g. paragraphs, title etc.) and special strings (tokens) like date, time, punctuations
  - HTML or XML-parsers can be applied for markup documents
- **Lexical analysis (morphology):**
  - Determination of word forms (singular-plural)
  - Determination of the kind of word (verb,noun)
    - Part of Speech tagging, POS
  - in German: composita analysis (in German)

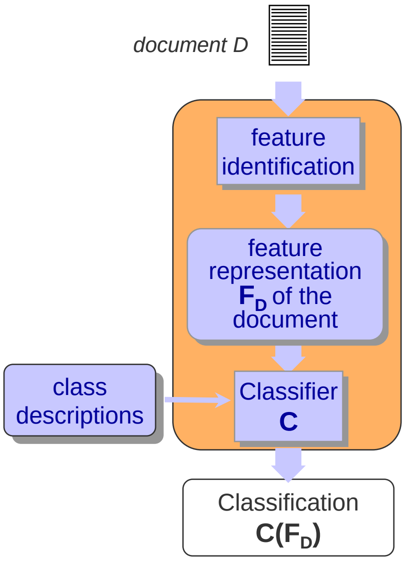


token scanner

lexical analysis

Automatic Classification





- Each document is described by a set of features
- Each class is described using the same kind of features
- A document is associated to the class(es) where the features are most similar. This can be tested using rules or similarity measures.

Rule-based Text Classification

- The features are keywords that are either associated to a document as metadata or that occur in the documents
- Example: Assume there are three classes: business
  computer science
  information systems

  The keywords in this example are: process
  OOP
  accounting
  ERP
  database
- The classifier can be represented as a set of rules:

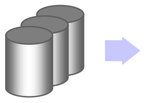
  IF a documents has the keywords process, accounting, and ERP
  THEN the document belongs to class „business“

  IF a documents has the keywords OOP and database
  THEN the document belongs to class „computer science“

  IF a documents has the keywords process, database, and ERP
  THEN the document belongs to class „information systems“

Fulltext Classification

- In the full text classification, the features are the terms occuring in the documents (fulltext index)
- The classes are represented as vectors

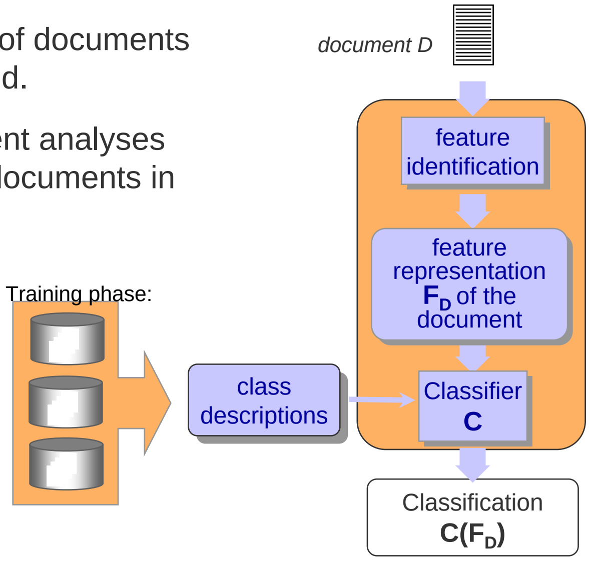


	c_1	c_2	c_3
t_1	w_{11}	w_{21}	w_{31}
t_2	w_{12}	w_{22}	w_{32}
t_3	w_{13}	w_{23}	w_{33}
t_4	w_{14}	w_{24}	w_{34}
t_5	w_{15}	w_{25}	w_{35}
t6	w_{16}	w_{26}	w_{36}

- The classification of a document is computed using a well-known ranking function well-known from inforamtion retrieval (cosinus).

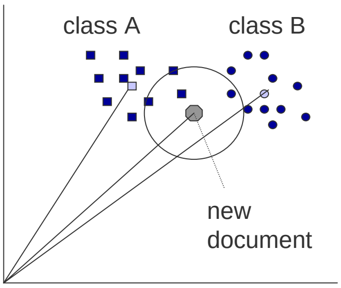
Automatic Learning of Classification Rules

- A characteristic set of documents is manually classified.
- A learning component analyses the features of the documents in the classes



Classification Methods

- Specific Document classifiers, e.g.
  - Linear Least Square Fit (LLSF)
  - Latent Semantic Analysis (LSA)
- Adaptation of general Classifiers, e.g.
  - Decision Trees
    - Explicit rules to test document features
  - K Nearest Neighbor
    - Documents are represented as vectors
    - A new document is compared with all documents of the training set
    - The majority of the k most similar documents gives the classification
  - Zentroid
    - Each class is represented by a prototypical vector
  - Neural Network



Information Extraction

- Example: From business news information about job changes should be extracted
- Sample text:

Peter Smith left Arconia Ltd. The former director retired on 31 March 2007. His successor is Susan Winter. At the same time George Young became sales manager. He followed John Kelly.

Template Instances that should be extracted from the sample text

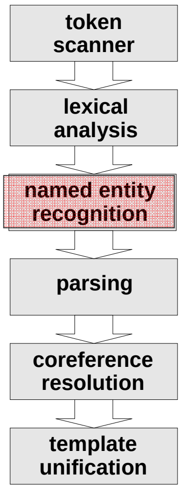
PersonOut	Peter Smith
PersonIn	Susan Winter
Position	director
Organization	Arconia Ltd
Date	31 March 2007

PersonOut	John Kelly
PersonIn	George Young
Position	sales manager
Organization	Arconia Ltd
Date	31 March 2007

Named Entity Recognition

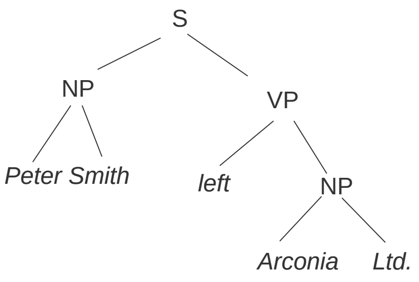
- Mark into the text each string that represents a person, organization, or location name, or a date or time, or a currency or percentage figure.
- Example:

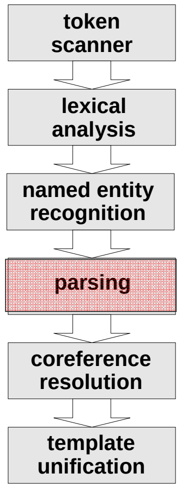
<name type=person>Peter Smith</name>, left <name type=organisation>Arconia Ltd. </name>. The former director retired on <date>31 March 2007</date>. His successor is <name type=person>Susan Winter</name>. At the same time <name type=person>George Young</name> became sales manager. He followed <name type=person>John Kelly</name>.


token scanner
lexical analysis
named entity recognition
parsing
coreference resolution
template unification

Parsing

- Parsing: Identification of phrase structures: noun phrase (NP), verb phrase (VP), ..

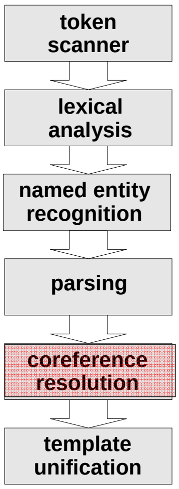



token scanner
lexical analysis
named entity recognition
parsing
coreference resolution
template unification

Coreference Resolution

- Capture information on corefering expressions, i.e. all mentions of a given entity, including those marked in NE and TE (nouns, noun phrases, pronouns).
- Example:
  - „the former director“ refers to „Peter Smith“
  - „His“ refers to „Peter Smith“
  - „He“ refers to „Georgs Young“
  - „At the same time“ refers to „31 March 2007“

<name type=person>Peter Smith</name>, left <name type=organisation>Arconia Ltd. </name>. The former director retired on <date>31 March 2007</date>. His successor is <name type=person>Susan Winter</name>. At the same time <name type=person>George Young</name> became sales manager. He followed <name type=person>John Kelly</name>.



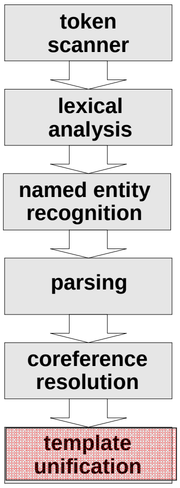
Template Unification

- Information for instantiating a single template often is distributed over multiple sentences. This information has to be collected and unified.
- Template Unification can comprise multiple tasks:
  - **Template Element Recognition (TE)**

    Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text
  - **Scenario Template Recognition (ST)**

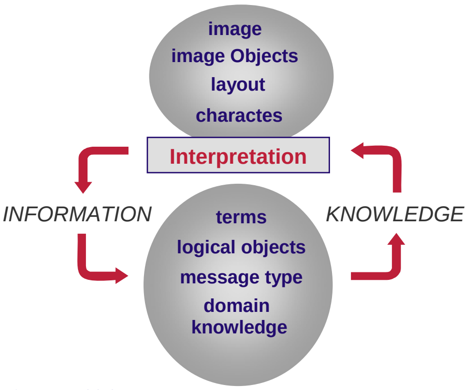
    Extract prespecified event information and relate the event information to particular organization, person, or artifact entities.
  - **Pattern Recognition (PR)**

    Identification of domain specific patterns (“Microsoft founder” = “Bill Gates”)

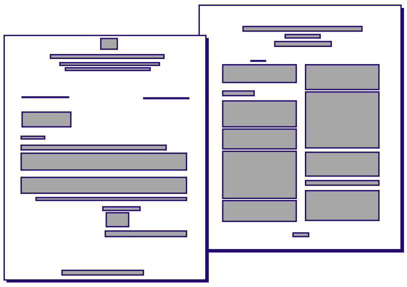


7.2 Information Extraction from (semi-)structured Document

- Integrated consideration of
  - layout structure
  - logical structure
  - content (semantics)

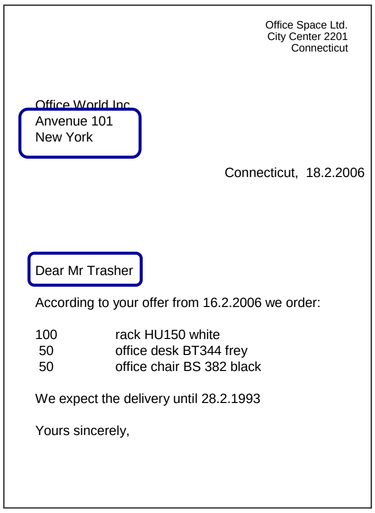


Example:



Source: A. Dengel, DFKI

Information Extraction using Layout, Logical Structure and Content



Office Space Ltd.
City Center 2201
Connecticut

Office World Inc
Anvenue 101
New York

Connecticut, 18.2.2006

Dear Mr Trasher

According to your offer from 16.2.2006 we order:

100	rack HU150 white
50	office desk BT344 frey
50	office chair BS 382 black

We expect the delivery until 28.2.1993

Yours sincerely,

Example: Letter

- Address of Recipient

  Layout: General Rules for position of address block

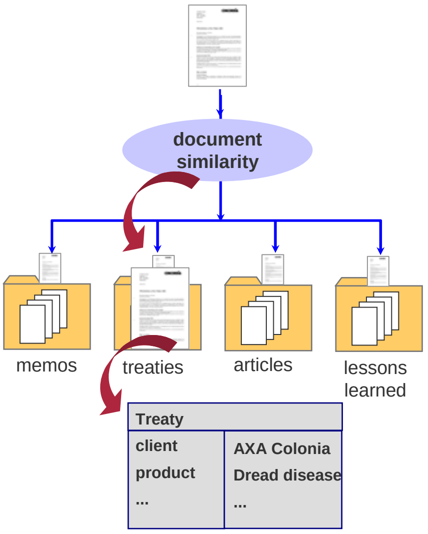
  Structure: Recipient consists of name and address

- Recipient

  Content: Knowledge aboutnamed entities and context

  „Dear Mr Trasher“

Guiding Extraction by Classification



Treaty	
client	AXA Colonia
product	Dread disease
...	...

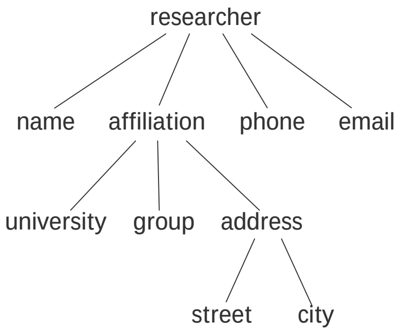
Knowledge about document structure can target information extraction

1. Classification:
   - Assigning documents to predefined document classes
   - For the document classes the structural objects are defined
2. Information Extraction
   - Identification of relevant information
   - Targeted seach in structural elements

Information Extraction from Markup Documents: XML

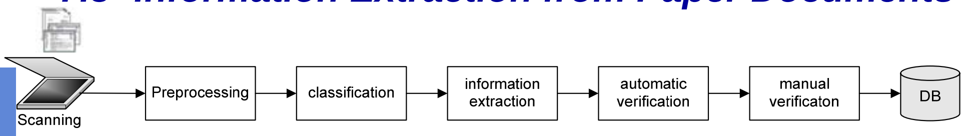
Predefined markup guides information extraction and recognition:

- Elements (tags, attributes)
- Structure



```
<researcher>
<name> Knut Hinkelmann </name>
<affiliation>
    <university> Fachhochschule
          Nordwestschweiz</university>
    <group> Wirtschaftsinformatik</group>
    <address>
          <street> Riggenbachstrasse 16 </street>
          <city> 4600 Olten </city>
    </address>
</affiliation>
<phone > ++41 62 286 00 80 </phone>
<email> knut.hinkelmann@fhnw.ch </email>
</researcher>
```

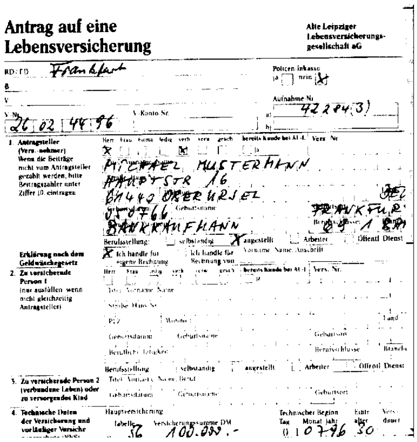
7.3 Information Extraction from Paper Documents



- **Scanning**
  - Result: Image of the document (non-coded information)
- **Preprocessing**
  - **Correction**
  - **Optical Character Recognition OCR**
    Intelligent Character Recognition ICR (advanced OCR e.g. hand writing)
  - Result: Content as text (coded information)
- **Classification**
  - Result: Document class (e.g. invoice of Hamilton Inc., ...)
- **Information extraktion**
  - Result: Relevant information in structured form (e.g. amount invoiced)

Information Extraction from forms

:



Antrag auf eine Lebensversicherung

Alte Leipziger Lebensversicherungsgesellschaft aG

RD / FD Frankfurt

Policen-Inkasso ja □ nein ☒

Aufnahme-Nr. 42284(3)

V.-Nr. 26 02 44 96

V.-Konto Nr.

1. **Antragsteller (Vers.-nehmer)** Wenn die Beiträge nicht vom Antragsteller gezahlt werden, bitte Beitragszahler unter Ziffer 10. eintragen

Herr ☒ Frau □ Firma □ ledig □ verh ☒ verw □ gesch □ bereits Kunde bei AL-L □ ja Vers. Nr.

MICHAEL MUSTERMANN

HAUPTSTR 16

61440 OBERURSEL

Geburtsdatum 050766 Geburtsname FRANKFURT

BANKKAUFMANN Berufsklasse 09 1 BM

Berufsstellung: □ selbständig ☒ angestellt □ Arbeiter □ Öffentl. Dienst

Erklärung nach dem Geldwäschegesetz ☒ Ich handle für eigene Rechnung □ Ich handle für Rechnung von Vorname, Name, Anschrift

2. **Zu versichernde Person 1** (nur ausfüllen, wenn nicht gleichzeitig Antragsteller)

Herr □ Frau □ ledig □ verh □ verw □ gesch □ bereits Kunde bei AL-L □ ja Vers. Nr.

Titel, Vorname, Name

Straße, Haus-Nr.

PLZ Wohnort Land

Geburtsdatum Geburtsname Geburtsort

Berufliche Tätigkeit Berufsklasse Branche

Berufsstellung □ selbständig □ angestellt □ Arbeiter □ Öffentl. Dienst

3. **Zu versichernde Person 2 (verbundene Leben) oder zu versorgendes Kind**

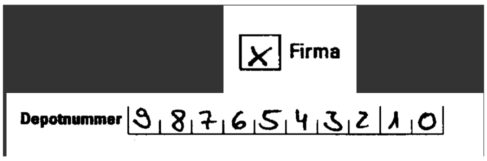
Titel, Vorname, Name, Beruf

Geburtsdatum Geburtsname Geburtsort

4. **Technische Daten der Versicherung und vorläufiger Versiche**

Hauptversicherung Tabelle 56 Versicherungssumme DM 100.000,-

Technischer Beginn Tag Monat Jahr 0 1 07 96 Eintr. alter 30 Vers.-dauer



☒ Firma

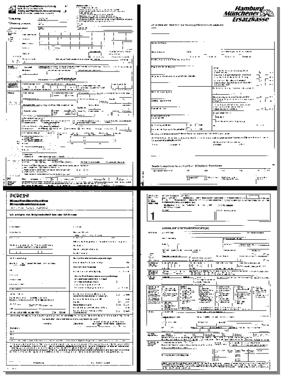
Depotnummer 9 8 7 6 5 4 3 2 1 0

- In forms the layout (position) determines the meaning of information
- The layout must be known to the recognition system
- The form must be sparated from the entries (content)

Types of documens

Fixed form

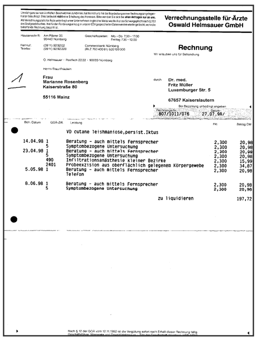
space for entries fixed



Hamburg Münchener Ersatzkasse

Dynamic form

forms with space for free entries (text, tables)



Verrechnungsstelle für Ärzte
Oswald Helmsauer GmbH

Hausanschrift: Am Plärrer 35
90443 Nürnberg

Geschäftszeiten: Mo.–Do. 7.30–17.00
Freitag 7.30–12.00

Fernruf: (0911) 92 92 02
Telefax: (0911) 92 92 220

Commerzbank Nürnberg
(BLZ 760 400 61) 520109000

Rechnung

Wir erlauben uns für Behandlung

O. Helmsauer · Postfach 22 22 · 90009 Nürnberg

Herrn/Frau/Fräulein

1

Frau
Marianne Rosenberg
Kaiserstraße 80

55116 Mainz

durch Dr. med.
Fritz Müller
Luxemburger Str. 5

67657 Kaiserslautern

Bei Bezahlung unbedingt angeben

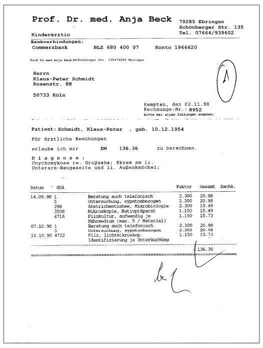
Rechnungs-Nr. 807/1011/076 Datum 27.07.98/

Beh.-Datum	GOÄ-Ziff.	Leistung	Fkt.	Betrag DM
		VD cutane leishmaniose,persist.Iktus		
14.04.98	1	Beratung - auch mittels Fernsprecher	2,300	20,98
	5	Symptombezogene Untersuchung	2,300	20,98
23.04.98	1	Beratung - auch mittels Fernsprecher	2,300	20,98
	5	Symptombezogene Untersuchung	2,300	20,98
	490	Infiltrationsanästhesie kleiner Bezirke	2,300	15,99
	2401	Probeexzision aus oberflächlich gelegenem Körpergewebe	2,300	34,87
5.05.98	1	Beratung - auch mittels Fernsprecher Telefon	2,300	20,98
8.06.98	1	Beratung - auch mittels Fernsprecher	2,300	20,98
	5	Symptombezogene Untersuchung	2,300	20,98
		zu liquidieren		197,72

Nach § 12 der GOÄ vom 12.11.1982 ist die Vergütung sofort nach Erhalt dieser Rechnung fällig.

Free documents

no predefined layout



Prof. Dr. med. Anja Beck

79285 Ebringen
Schönberger Str. 135
Tel. 07664/939602

Kinderärztin

Bankverbindungen:
Commerzbank BLZ 680 400 07 Konto 1966620

Prof Dr.med.Anja Beck*Schönberger Str. 135*79285 Ebringen

Herrn
Klaus-Peter Schmidt
Rosenstr. 88

50733 Köln

1

Kempten, den 02.11.98
Rechnungs-Nr.: 8952
Bitte bei allen Zahlungen angeben.

Patient: Schmidt, Klaus-Peter , geb. 10.12.1954

Für ärztliche Bemühungen

erlaube ich mir DM 136.36 zu berechnen.

Diagnose:
Onychomykose re. Großzehe; Ekzem am li.
Unterarm-Beugeseite und li. Außenknöchel;

Datum	GOÄ		Faktor	Gesamt	Sachk.
14.09.98	1	Beratung auch telefonisch	2.300	20.98	
	5	Untersuchung, sypmtombezogen	2.300	20.98	
	298	Abstrichentnahme, Mikrobiologie	2.300	10.49	
	3508	Mikroskopie, Nativpräparat	1.150	10.49	
	4716	Pilzkultur, aufwendig je Nährmedium (max. 5 / Material)	1.150	15.73	
07.10.98	1	Beratung auch telefonisch	2.300	20.98	
	5	Untersuchung, sypmtombezogen	2.300	20.98	
16.10.98	4722	Pilz, lichtmikroskop. Identifizierung je Untersuchung	1.150	15.73	
				136.36	

Dokumentklassen

- Um Informationen extrahieren zu können, muss der Aufbau der Dokumente bekann sein.
- Dokumentklassen sind Dokumente mit gleichartigem Aufbau
- Dokumentklassen steuern die Informationsextraktion
  - Zu jeder Dokumentklasse ist definiert, wo welche Information extrahiert wird
  - Beispiel: Rechnung:

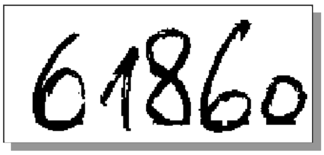
> Adresse	> Kunden.-Nr.
> Bank	> Bankleitzahl
> Kontonummer	> Betrag

- Dokumentklassen können sehr spezifisch sein
  - z.B. Rechnungsformular der Firma Meyer GmbH
  - in diesem Fall ist genau bekannt, wo die gesucht Information zu finden ist
- Dokumentklassen können sehr allgemein sein
  - z.B. allgemeine Arztrechnung
  - in diesem Fall ist mehr Aufwand bei der Suche nach Information auf dem Dokument notwendig

Phase 1: Preprocessing

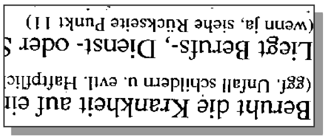
Elimination of lines:

lines negatively influence OCR results



Noise elimination

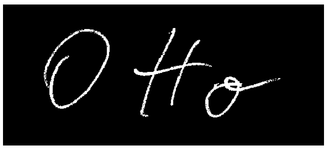
Uside-down-correction

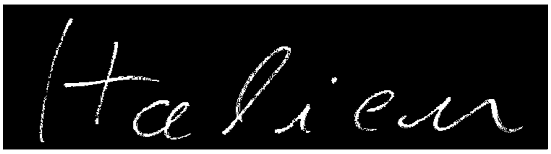

(wenn ja, siehe Rückseite Punkt 11)
Liegt Berufs-, Dienst- oder S
(ggf. Unfall schildern u. evtl. Haftpflic
Beruht die Krankheit auf ei

Rotation correction

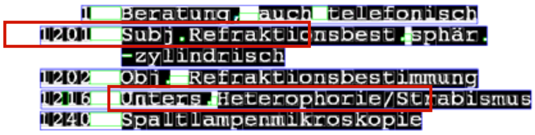
Problems with OCR/ICR

- Errors in
- Ambiguities





- Wrong segmentation

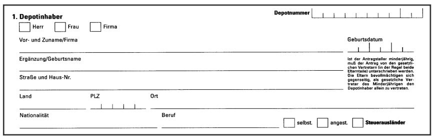


```
1    Beratung, auch telefonisch
1201 Subj.Refraktionsbest.sphär.
     -zylindrisch
1202 Obj.Refraktionsbestimmung
1216 Unters.Heterophorie/Strabismus
1240 Spaltlampenmikroskopie
```

Phase 2: Clasification

Using layout and logic structure as additional features for classification

Layout: lines, tables, ...



1. Depotinhaber

Depotnummer

□ Herr □ Frau □ Firma

Vor- und Zuname/Firma

Geburtsdatum

Ergänzung/Geburtsname

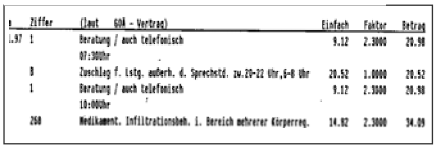
Ist der Antragsteller minderjährig, muß der Antrag von den gesetzlichen Vertretern (in der Regel beide Elternteile) unterschrieben werden. Die Eltern bevollmächtigen sich gegenseitig, als gesetzliche Vertreter des Minderjährigen den Depotinhaber allein zu vertreten.

Straße und Haus-Nr.

Land PLZ Ort

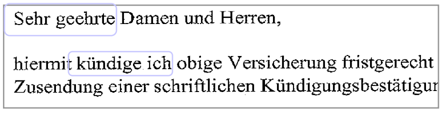
Nationalität Beruf □ selbst. □ angest. □ **Steuerausländer**

table structure and content ...



t	Ziffer	(laut GOÄ - Vertrag)	Einfach	Faktor	Betrag
1.97	1	Beratung / auch telefonisch 07:30Uhr	9.12	2.3000	20.98
	8	Zuschlag f. Lstg. außerh. d. Sprechstd. zw.20-22 Uhr,6-8 Uhr	20.52	1.0000	20.52
	1	Beratung / auch telefonisch 10:00Uhr	9.12	2.3000	20.98
	268	Medikament. Infiltrationsbeh. i. Bereich mehrerer Körperreg.	14.82	2.3000	34.09

predefined search patterns
(regular expressions)



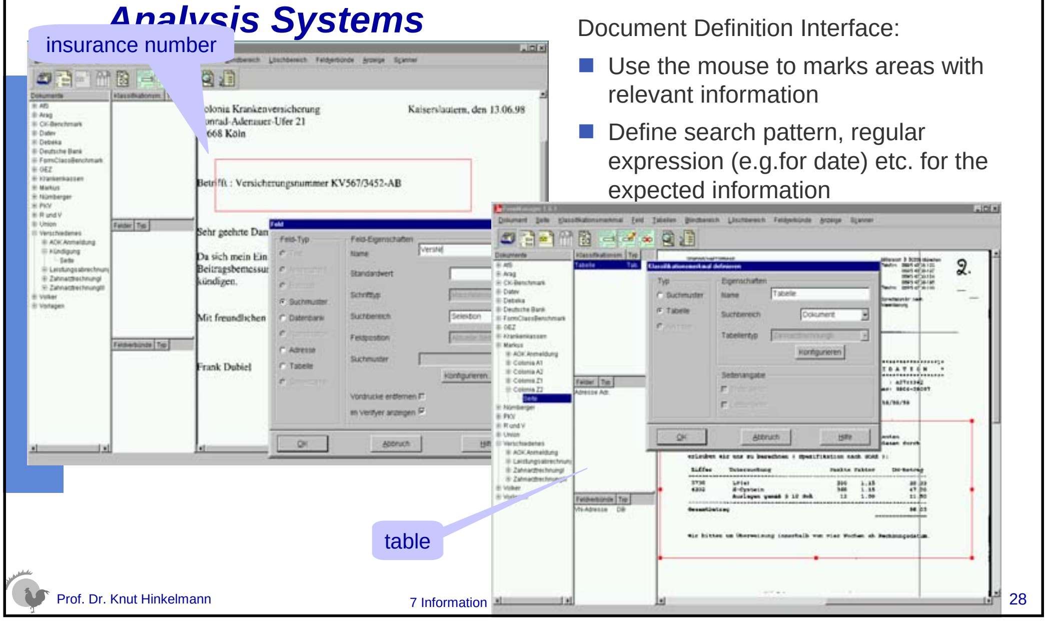
Sehr geehrte Damen und Herren,

hiermit kündige ich obige Versicherung fristgerecht
Zusendung einer schriftlichen Kündigungsbestätigur

Definition of Document Classes in Document Analysis Systems

Document Definition Interface:

- Use the mouse to marks areas with relevant information
- Define search pattern, regular expression (e.g.for date) etc. for the expected information



insurance number

table

olonia Krankenversicherung
onrad-Adenauer-Ufer 21
668 Köln

Kaiserslautern, den 13.06.98

Betrifft : Versicherungsnummer KV567/3452-AB

Sehr geehrte Dan

Da sich mein Ein
Beitragsbemessur
kündigen.

Mit freundlichen

Frank Dubiel

Feld-Typ: ● Suchmuster
Name: Versnr

Suchbereich: Selektion

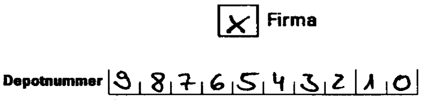
Vordrucke entfernen ☐
im Verifyer anzeigen ☑

Klassifikationsmerkmal definieren
Typ: ● Tabelle
Name: Tabelle
Suchbereich: Dokument

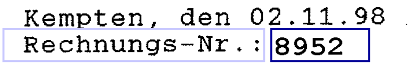
Phase 3: Information Extraction

Extract relevant Information from

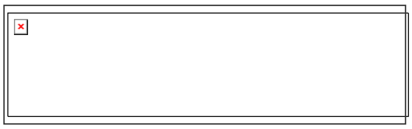
- Form fields with fixed position


☒ Firma
Depotnummer 9 8 7 6 5 4 3 2 1 0

- Search patterns


Kempten, den 02.11.98
Rechnungs-Nr.: 8952

- Tables

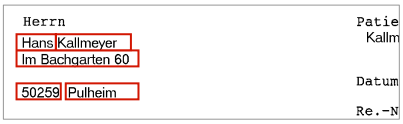


- Regular expression

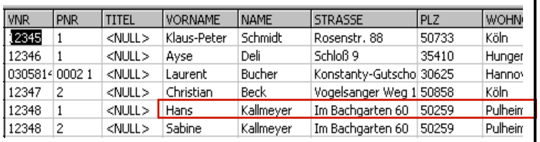
hiermit kündige ich zum 31.12.2003
mein Abonnement ...

Phase 4: Automatic Verification

- Database matching: Compare extracted ifnormation with content of a database (Levensthein distance)

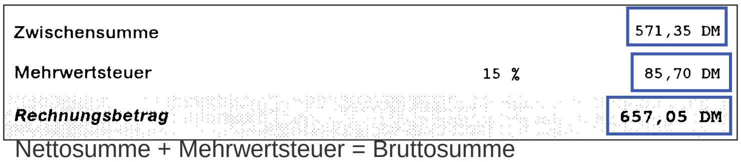


```
Herrn                                   Patie
Hans Kallmeyer                          Kallm
Im Bachgarten 60
                                        Datum
50259 Pulheim
                                        Re.-N
```



VNR	PNR	TITEL	VORNAME	NAME	STRASSE	PLZ	WOHN
12345	1	<NULL>	Klaus-Peter	Schmidt	Rosenstr. 88	50733	Köln
12346	1	<NULL>	Ayse	Deli	Schloß 9	35410	Hunger
0305814	0002 1	<NULL>	Laurent	Bucher	Konstanty-Gutscho	30625	Hannov
12347	2	<NULL>	Christian	Beck	Vogelsanger Weg 1	50858	Köln
12348	1	<NULL>	Hans	Kallmeyer	Im Bachgarten 60	50259	Pulheim
12348	2	<NULL>	Sabine	Kallmeyer	Im Bachgarten 60	50259	Pulheim

- Logical verification: Checking logical or mathematical conditions



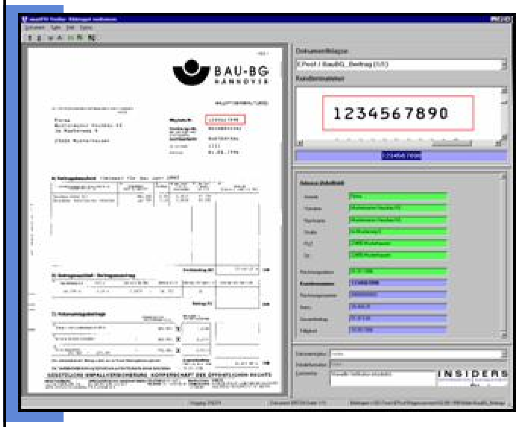
Zwischensumme		571,35 DM	Field \`Netto´
Mehrwertsteuer	15 %	85,70 DM	Field \`Mwst´
Rechnungsbetrag		657,05 DM	Field \`Brutto´

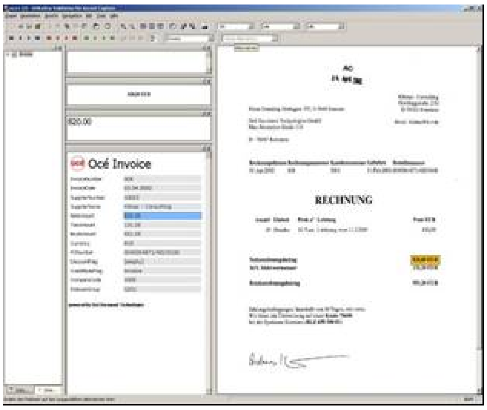
Nettosumme + Mehrwertsteuer = Bruttosumme

Expression: EQUAL(ROI(\`Brutto´), SUM(ROI(\`Netto´), ROI(\`Mwst´)))

Phase 5: Manual Verification

Document Analysis Tools provide an interface for manual verifcation


BAU-BG
HANNOVER
1234567890
INSIDERS


820.00
Océ Invoice
RECHNUNG