Gander, Lezuo, Unterweger :

Rule based document understanding of historical books using a hybrid fuzzy classification system

SS 12 - Logische Dokumentenanalyse: Andreas Neumann

Datum

Überblick

- Anwendungsgebiet
- Techniken / Ansätze
- Umsetzung
- Auswertung

Anwendungsgebiete

- Massendigitalisierung: Bücher
- Szenario, Beispielanwendungen:
  - Inhaltsverzeichnisse zur Navigation nutzen
  - Inhaltsverzeichnisse erzeugen, wenn nicht vorhanden
  - Inderdependenzen zw. Büchern sichtbar / navigierbar machen
  - Indexierung verbessern (Ranking boosten, gezielte Suche)

Was wird extrahiert

- Überschriften
- Fussnoten
- Seitenzahl
- Fließtext
- Signature Mark

Beschreibung des Verfahrens

- Mehrstufiger Bottom-Up-Ansatz
- Soll den menschlichen Verstehensprozess eines Dokuments mit dessen Regelbildung nachahmen

Techniken zum Dokumentverstehen

- Vorverarbeitung:
  - OCR mit physischen Strukturangaben
  - Featureextraktion
- Interpretation / Verarbeitung
  - Handkodierte Regeln
  - Machine Learning : Fuzzy-Logic + Learning Algorithm
  - Verbesserung der Ergebnisse

Vorverarbeitung

- Bildverbesserung
- OCR

OCR mit physischen Strukturangaebn

- Dokument wird einer OCR unterzogen
- Koordianten
- physische Strukturen: Blöcke, Tables, Pictures, Zeilen, Wörter
- Formatierung, Schrifgewichte (Bold, Italic, normal)

Features finden / berechnen

1. Aus OCR- Ergebnis direkt

2. Aus einem Nachverarbeitungsschritt, basierend auf OCR-Daten

- x1 = Distance to the previous line
- x2 = Distance to the subsequent line
- x3 = Left indent of the line
- x4 = centring of the line
- x5 = Length of the line
- x6 = Number of Lines within the same text block
- x7 = Distance of the text block, containing the re- viewed line, to the previous text block
- x8 = Distance of the text block, containing the re- viewed line, to the subsequent text block
- x9 = Average surface area of a character within the reviewed line

Machine Learning

- Input annotieren
- Regeln lernen
- Labeling
- Refinement

Input

- x1 ... xi => Annotierte Eingabe
- y => Label (manuell vergeben)

(x1,x2, ...xi,y)

(36, 41, 5, 0.3, 951, 4, 35, 38, 25.324, footnote)

- Distance of 36 pixel to the previous line,
- Distance of 41 pixel to the subsequent line,
- Left indent of 5 pixel,
- 0.3 degree of centering,
- Length of 951 pixel,
- Sharing the same text block with 4 other lines,
- Distance of 35 pixels of the text block, containing the reviewed line, to the previous text block,
- Distance of 38 pixels of the text block, containing the reviewed line, to the subsequent text block,
- Average surface area of 25.324 square pixel needed for one character,
- Manually labelled as footnote

4 stufiges Lernverfahren - Variation Wang - Mendel - Algorithmus

- Werte fuzzifizieren
- Fuzzy-Rules generieren
- Zuverlässigkeits-/Sicherheitswertswert für Regeln berechnen
- Regelbasis verbessern

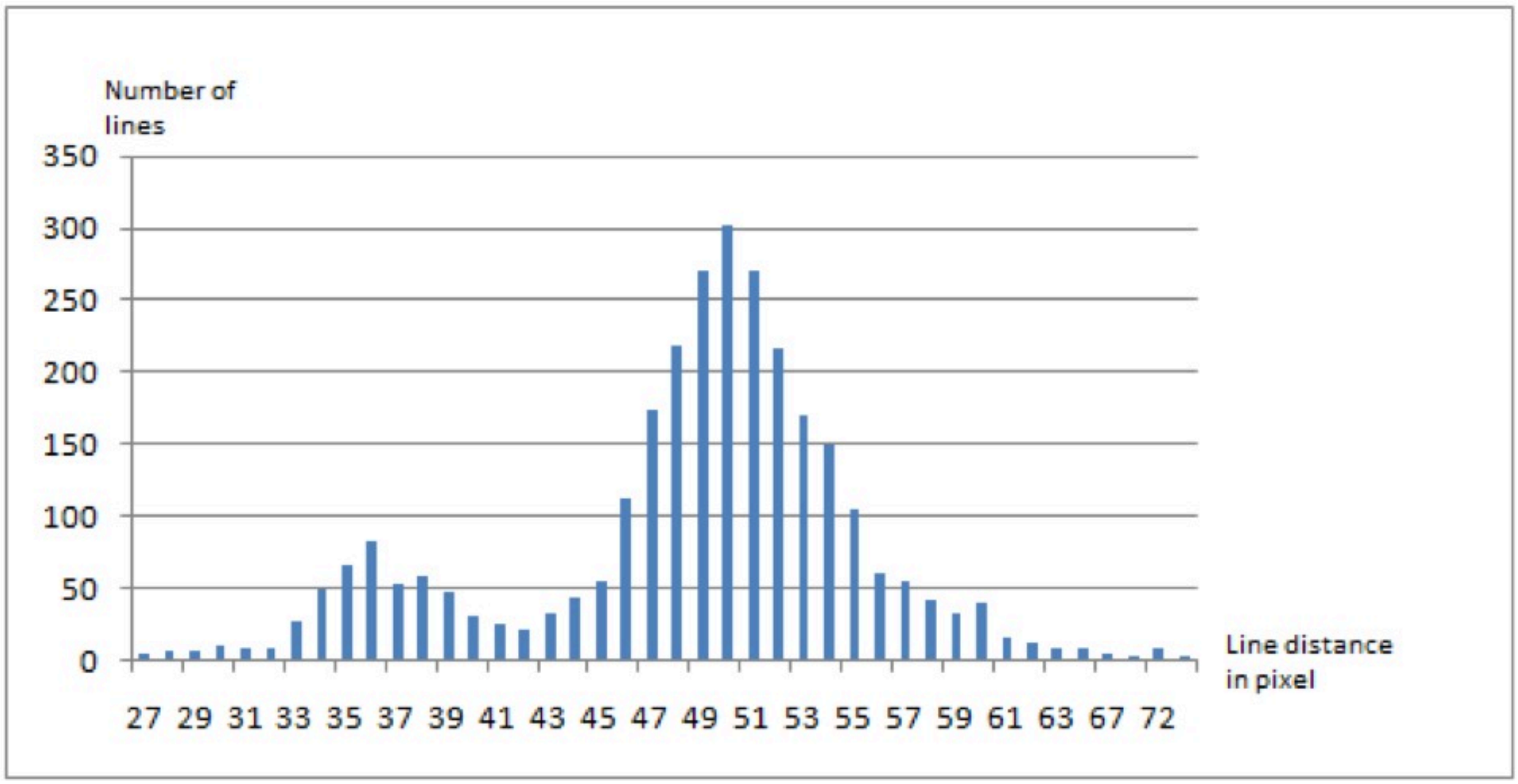
Schritt 1 : Werte fuzzifizieren

- Werte fuzzifizieren, für jedes Dokument von neuem
- „normale Größe“ kann von Dokument zu Dokument variieren

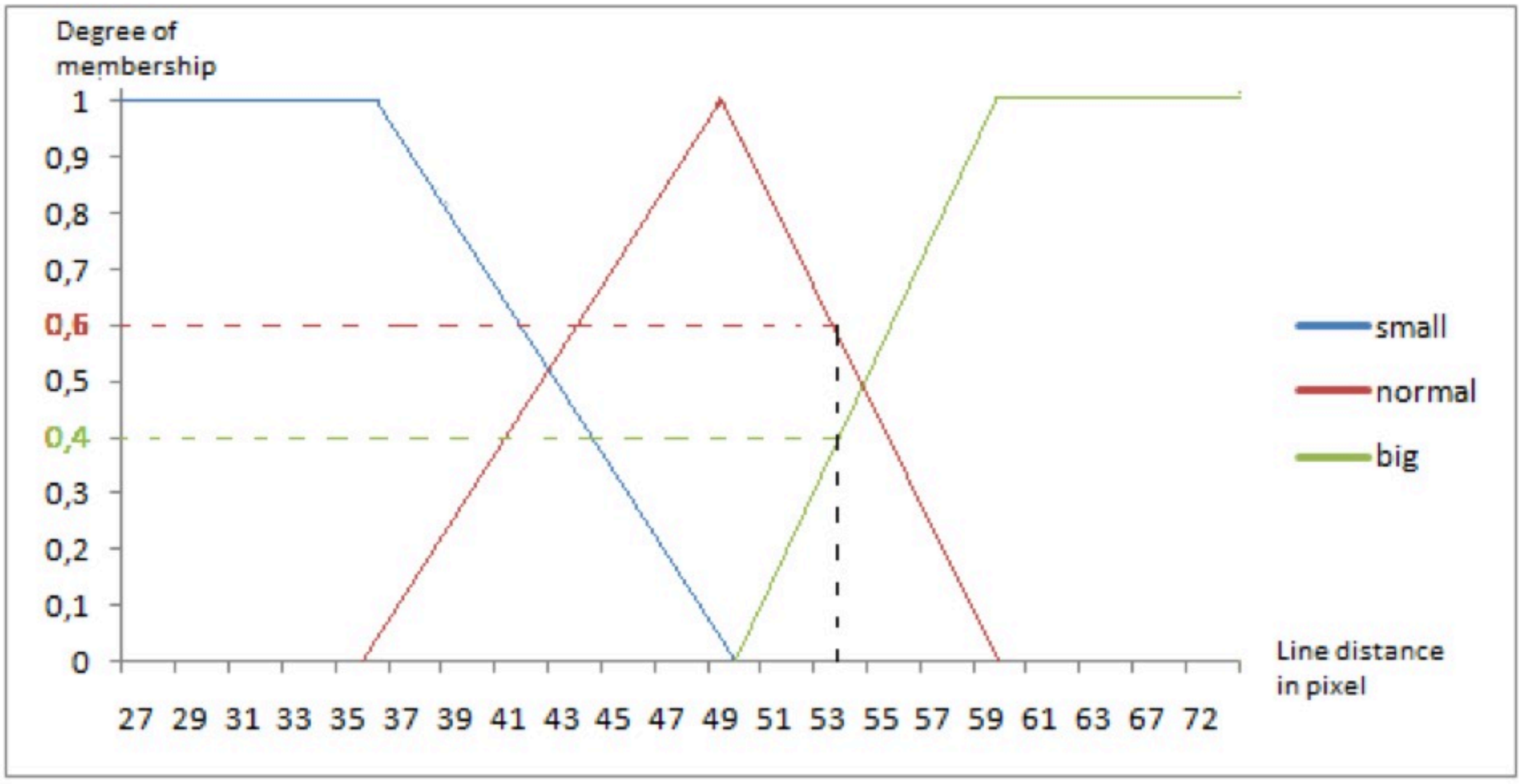
Fuzzy - Logic

- Daten aus OCR-Prozess sind selten exakt
- Beispielsweise ist es unwahrscheinlich, dass alle Zeilehöhen pixelgenau erkannt werden
- Die Daten werden sich einem optimalen Wert aber annähern
- Fuzzy-Logic eignet sich um mit diesem Phänomen umzugehen

Beispiel Zeilenhöhe



Beispiel Zeilenhöhe



Schritt 2 : Regeln lernen

- Aus gelabelten Input-Daten
- Kopf mit Bedingung und AND-Verknüpfung für jedes Feature
- Label im THEN-Teil

IF
  x′1 is small AND x′2 is small AND
  .
  .
  x′9 is small
THEN
  y = footnote

Gelernte Regeln

- Zero-Order Takagi-Sugeno-Rules
- Ergebnis: Eine Funktion, die einem Input eindeutig ein Label zuweist
- x´1 ... x´1 => Fuzzifizierte Input-Werte
- y => Label

$$(x'1,x'2, \ldots x'i) \Rightarrow y$$

Schritt 3 : Qualitätsmaß an Regeln vergeben

- Durch Schritt 2 können widersprüchliche Regeln erzeugt werden
- Zum berechnen wird die minimum T-Norm benutzt
- damit wird jeder Regeln ein Wert zwischen 0 und 1 zugewiesen, basierend auf der Zugerhörigkeit einzelner fuzzifizierter Inputwerte
- D_{rule} => Qualität der Regel 0 ... 1
- $m_{A1}(x_1)$ => Zugehörigkeit x1 zur fuzzifizierten Menge

$$D_{rule} = \min(m_{A_1}(x_1), m_{A_2}(x_2), \ldots m_{A_j}(x_i))$$

Schritt 4: Regelbasis schaffen

- Ambiguitäten in der Regelbasis auflösen
- Votingprozedur führt zu einer Regel (hier anders als Wang & Mendel!)
- Regeln werden nach Antezedent (Kopf/IF-Teil) gruppiert
- jeder Body (THEN-Teil) stimmt mit seinem Qualitätswert
- Für die (am höchsten gerankte/gestimmte) Regel wird ein Wahrscheinlichkeit berechnet
- Stimmen für Regel / Gesamtanzahl Stimmen in Gruppe

Beispiel für Lernverfahren - Beispielinput

- Ein Zeile L1
- mit folgenden fuzzifizierten Zugehörigkeiten

- x′1(L1) is
  small with a degree of **0.65** and
  normal with a degree of **0.35**
- x′2(L1) is
  small with a degree of **0.45** and
  normal with a degree of **0.55**

Beispiel für Lernverfahren - Regeln

- Zwei Regeln
- 1. Fußnote mit 0,9 Wahrscheinlichkeit / 2. Text mit 0,6 Wahrscheinlichkeit

1. IF
   x′1 is small AND
   x′2 is small
   THEN
   y=footnote
   with precision P = 0.9

2. IF
   x′1 is small AND
   x′2 is normal
   THEN
   y=text
   with precision P = 0.6

Beispiel für Lernverfahren - Anwendung der Regeln auf Beispielinput

- Zugehörigkeit der Beispielzeile zu den Regeln.

$$DR1\ (L1) = \min(0.65, 0.45) = 0.45$$
$$DR2\ (L1\ ) = \min(0.65, 0.55) = 0.55$$

- Berechnung der Confidence über ein Kombination der Regelzugehörigkeit und und des Konfidenzwerts der Regel

$$CR1\ (L1) = 2 * 0.45 * 0.9/(0.45 + 0.9) = 0.60$$
$$CR2\ (L1) = 2 * 0.55 * 0.6/(0.55 + 0.6) = 0.57$$

Beispiel für Lernverfahren - Ergebis

- L1 ist mit 0.6 ein Fußnote
- L1 ist mit 0.57 Text

Auszeichnung / Labeling

- Zweistufiges Verfahren
- Handkodierte Regel anwenden
- erlernte Regeln anwenden

Handkodierte Regeln

- Basieren auf Domänenwissen
- hoher Aufwand beim kodieren
- gute Ergebnisse
- hier für: Signaturmarken, Seitennummern, Page-Headers

```
/*****************************************************************************************
*                                                                                        *
*                              Rules Module 1                                            *
*                                                                                        *
*                 o initializes Outputfact                                               *
*                 o detects Strings which are physically placed in multiple Paragraphs   *
*                                                                                        *
*****************************************************************************************/

(defrule m1::init
    (MAIN::FEP_Paragraph (fep_Document_ID ?docid)(fep_Page_ID ?pid))
    (not (MAIN::OCR_Quality_Evaluation))

    =>
    (assert
        (MAIN::OCR_Quality_Evaluation(fep_document_ID ?docid)(fep_Page_ID ?pid)(string_multiple_block 0))
        )
)

(defrule m1::identify_Strings_in_multiple_Blocks

    ;paragraph1
    (MAIN::FEP_Paragraph (readingOrder ?order1)(left ?left_p1)(right ?right_p1)(top ?top_p1)(bottom ?bottom_p1))

    ;paragraph2
    (MAIN::FEP_Paragraph (readingOrder ?order2&:(neq ?order1 ?order2))
                         (left ?left_p2)(right ?right_p2)(top ?top_p2)(bottom ?bottom_p2))
    ;string physically placed in both paragraphs
    (MAIN::FEP_String    (readingOrder ?id)
                         (top ?t&:(and (> ?t ?top_p2)(> ?t ?top_p1)))
                         (bottom ?b&:(and (< ?b ?bottom_p2)(< ?b ?bottom_p1)))
                         (right ?r&:(and (< ?r ?right_p2)(< ?r ?right_p1)))
                         (left ?l&:(and (> ?l ?left_p2)(> ?l ?left_p1)))
                         )
=>
    ;create new fact
    (assert
         (MAIN::String_multiple_block (id ?id))
        )
)
```

erlernte Regeln anwenden

- nur auf Elemente (hier Zeilen), die von den handkodierten Regeln nicht erfasst wurden

- alle möglichen Ergebnisse werden gespeichert: Text:0.65, Footnote 0.10

- Gibt es zu einem möglichen Label keine Regel erhält das Element den Wert 0 für dieses Label

Verbesserung

- Bottom-Up

- Jedes Element hat in der Ausganssituation mehre Labels mit verschiedenen Konfidenzwerten

- Ziel : Jedem Element ein einziges Label zuorden

- Zweistufiges Verfahren

Verbesserung Vorannahmen

- Inderdependenzen zwischen Elementen (Zeilen) wurden bei Labeling nicht berücksichtigt

- OCR hat physische Einheiten (Seite > Block > Paragraph > Line) korrekt erkannt

- Innerhalb einer physischen Einheit haben alle Elemente das gleiche Label

Verbesserungsschritt 1: Paragraphenebene

- Falsche Labels auf Paragraphenebene erkennen und entfernen
- Voting innerhalb eines Paragraphen für wahrscheinlichstes Label

Line 1: footnote with C=0.60, text with C=0.57 and heading with C=0
Line 2: footnote with C=0.55, text with C=0.77 and heading with C=0.05
Line 3: footnote with C=0.30, text with C=0.75 and heading with C=0.15

```
1.9.2p290 :003 > (0.60 + 0.55 + 0.3) / 3.0 #Footnote
 => 0.48333333333333334

1.9.2p290 :004 > (0.57 + 0.77 + 0.75) / 3.0 #Text
 => 0.6966666666666667

1.9.2p290 :005 > (0.15 + 0.05) / 3.0 #Heading
 => 0.06666666666666667
```

Verbesserungsschritt 2: Seitenebene

- Seitenaufbau ist relativ einheitlich

- Grammatikalischer Ansatz

- DFA zum validieren

- Grammatik wurde anhand von Beispielseiten der annotierten Groundtruth erlernt

Verbesserungsschritt 2: Reguläre Ausdrücke

- Beschreiben Seitenstruktur

(label) : Ein Elemente
(Label)+ : Ein- oder mehrmaliges auftreten des Elements

Beispiel:
(text)+ ((heading)+(text)+)+ (footnote)+

Verbesserungsschritt 2: DFA

- Auf Basis von 60.000 handannotierten Eingabeseiten gewonnen
- 80 reguläre Ausdrücke, händisch nachgeprüft
- Reguläre Ausdrücke werden in einen DFA übersetzt

Verbesserungsschritt 2: Anwendung

- Input aus Verbesserungsschritt 1 wird von Automaten geprüft
- bei Nichtannahme wird die Seite als „suspicious“ markiert
- „suspicious“-Seiten werden mit verschiedenen Heuristiken bearbeitet bis sie vom Automaten akzeptiert werden
- (Wird in Paper nicht beschrieben: Abbruch nach bestimmter Zeit, bleibt suspicious)

Ergebnis

Table 1: Experimental Results

	recall		precision		f-measure	
	E	T	E	T	E	T
page number	0.97	0.97	1.00	0.99	0.98	0.98
page header	0.97	0.97	1.00	0.99	0.98	0.98
signature mark	0.68	0.67	0.89	0.91	0.77	0.77
text	0.99	0.99	0.98	0.99	0.98	0.99
footnote	0.83	0.93	0.89	0.91	0.86	0.92
heading	0.85	0.87	0.80	0.81	0.82	0.84

Table 2: Experimental Results after human correction

	recall		precision		f-measure	
	E	T	E	T	E	T
signature mark	0.82	0.77	0.92	0.94	0.87	0.85
text	1.00	1.00	0.99	0.99	0.99	0.99
footnote	0.94	0.98	0.96	0.98	0.95	0.98
heading	0.88	0.9	0.89	0.93	0.88	0.91

- Basis: 200 Bücher
- Traingsset: 160
- Testset:40