Machine-Learning basierte automatische OCR-Korrektur

Betreuer: Dr. Klaus Schulz

von

Michael Strohmayer

Übersicht

- Motivation
- Ziel der Arbeit
- Vorgehensweise
- Probleme
- Evaluierung
- Ausblick

Motivation

- Schriftbilder, Grammatik und Schreibweisen verändern sich
- OCR-Systeme erkennen manche Wörter nicht zuverlässig
- Liefern eine Liste an Korrekturvorschlägen von denen der korrekte ausgewählt werden muss

Ziel der Arbeit

- Erstellung einer Software zur automatischen Nachkorrektur der eingelesenen OCR-Dokumente
- Trainieren eines Machine-Learning Systems
- Auswertung der Ergebnisse

Vorgehensweise

- Einlesen der Dokumente
- Extrahieren der gegebenen Featurewerte
- Hinzufügen neuer Features
- Training von Machine-Learning Klassifikatoren
- Evaluation

Einlesen der Dokumente

- Verfügbare Grund-Truth Dokumente „Paradiesgärtlein“ und „Curiöser Botanicus“
- RIDGES Korpus, 33 Kräuterkundetexte aus der Zeit zwischen 1484 und 1914
- Erstellt von CIS LMU in Kooperation mit Humboldt Universität in Berlin

Extrahieren der gegebenen Featurewerte

Leter:{leder+[(d:t,2)]}
+ocr[(t:h,2)],voteWeight=0.000406401,levDistance=1,ocrToken=Leher,corToken=Leber,ocrAbsFreq=1,ocrReflFreq=0.000323

Extrahieren der gegebenen Featurewerte

Leter:{leder+[(d:t,2)]}
+ocr[(t:h,2)],voteWeight=0.000406401,levDistanc
e=1,ocrToken=Leher,corToken=Leber,ocrAbsFreq
=1,ocrReflFreq=0.000323

-1 0.000406401 1

Extrahieren der gegebenen Featurewerte

Leber:{leber+[]}
+ocr[(b:h,2)],voteWeight=0.995316,levDistance=1,ocrToken=Leher,corToken=Leber,ocrAbsFreq=1,

ocrReflFreq=0.000323

Extrahieren der gegebenen Featurewerte

Leber:{leber+[]}
+ocr[(b:h,2)],voteWeight=0.995316,levDistance=1,
ocrToken=Leher,corToken=Leber,ocrAbsFreq=1,

ocrReflFreq=0.000323

1 0.995316 1

Hinzufügen zusätzlicher Features

- Längendifferenz
- Konfidenzwert des folgenden Korrekturvorschlags
- Frequenzlisten

Training von Machine-Learning Klassifikatoren

- Scikit-learn – Große Bibliothek an Machine-Learning und Data Mining Tools

  Verwendet für Gauß Naive Bayes Klassifikator
- Libsvm – verwendet Support Vector Machine zur Klassifizierung von Daten

Probleme

- Im Anfangsstadium Perfomance Probleme in der Datenverarbeitung
- Konfidenzwerte in der Ausgabe abgeschnitten, deshalb falsche Trainingswerte

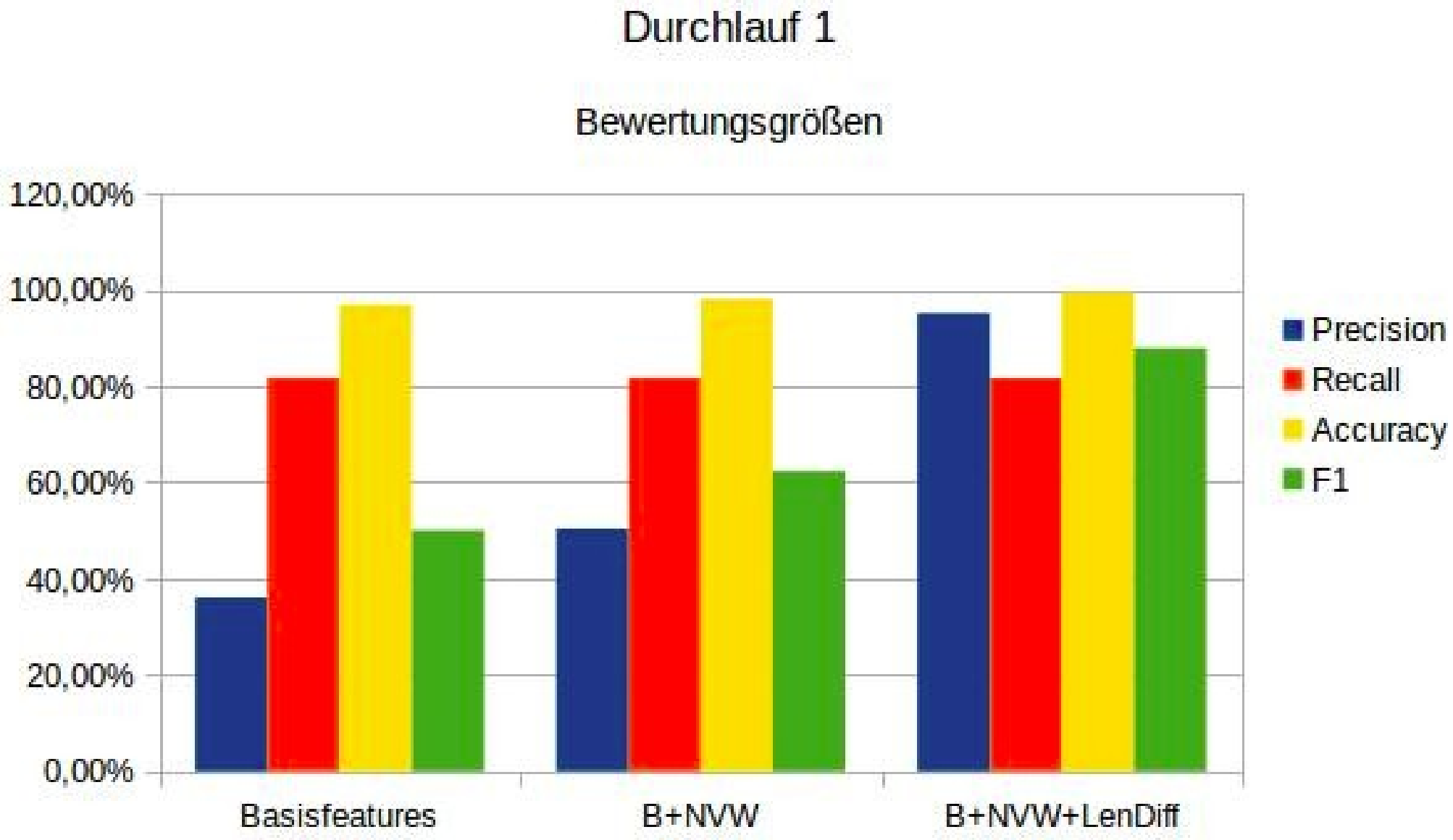
Evaluation

- Durch Kreuzevaluierung wurden deutlich bessere Ergebnisse erziehlt
- Berechnung von Naive Bayes ist sehr schnell, libsvm bietet hochwertigere Ergebnisse

Libsvm und Naive Bayes im Vergleich

	Train	Predict
Libsvm Multi-C	34.00s	4.26s
Libsvm One-C	12.048s	3.26s
NaiveBayes	0.32s	0.009s

Evaluation


Durchlauf 1
Bewertungsgrößen
120,00%
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
Basisfeatures
B+NVW
B+NVW+LenDiff
Precision
Recall
Accuracy
F1

Häufige Fehlklassifikationen

- Wenn der Profiler einen falschen Konfidenzwert geliefert hat
- Wenn kein Korrekturvorschlag in den Grund-Truth Daten vorhanden ist

Ausblick

- Automatische Nachkorrektur von OCR Dokumenten ist durchaus sinnvoll und liefert gute Ergebnisse
- Weitere Schritte wären die Kombination von beiden Klassifikatoren und die Hinzunahme von weiteren Features

Interessantes zum Weiterlesen

- Scikit-learn http://scikit-learn.org/stable/
- Libsvm https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Johannes Lächle, Support Vector Machines http://www.cogsys.cs.uni-tuebingen.de/lehre/ss06/pro_learning/JohannesLaechele.pdf
- Uwe Springmann, Anke Lüdeling,Felix Schremmer, Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten, http://gams.uni-graz.at/o:dhd2015.p.34