Text KI ist ein Baukasten

Dr. Hans Weber

portamis
Automatic Layout

portamis Software GmbH

portamis macht seit 20 Jahren u.a.

- **aus XML gestaltetes PDF**
- **aus XML gestaltetes InDesign**
- **aus XML anderes XML**
- **aus PDF Informationen**
- **aus PDF strukturiertes XML**

Hans Weber, einer von zwei Gründern

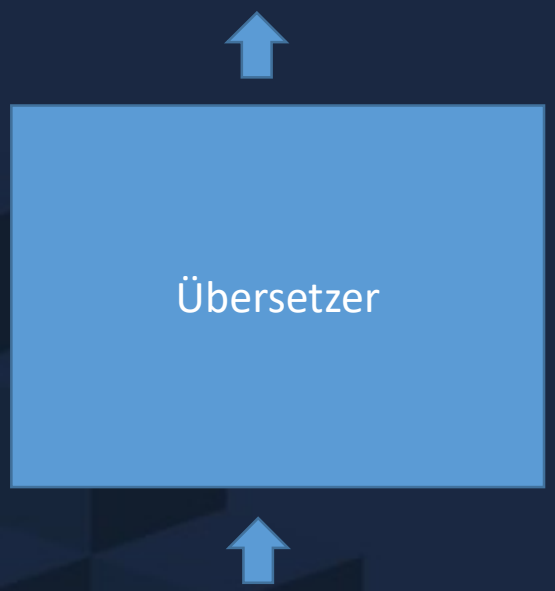
- **1990 bis 2001: Text-KI**
- **2005 bis heute: XML Anwendungen wie oben..**
- **Seit ca. 5 Jahren: Beides**

Motivation / Inhalt

Überblick und Verständnis geben ....

- **Zwei Grundtechniken in der Text-KI dargestellt**
  - **Klassifizieren (einfache Entscheidungen)**
  - **Übersetzen (komplexe Abbildungen)**
- **Für welche Aufgaben ist das geeignet**
- **Text-KI Aufgaben bei XML Beständen und Altdokumenten**
- **Aufbau von Lösungen als Baukasten von Classifiern**

Grundbausteine Text-KI
Klassifikation und Übersetzung



Classifier werden trainiert mit getaggten Daten:

Listen von Beispielen von Wortketten für eine feste Kategorie

Übersetzer werden trainiert mit Paaren von Wortketten:

Jeweils eine Wortkette als Eingabe und eine als Ausgabe

Text Klassifikation

Kategorie, Score

Klassifikator

Feature 1, Feature 2…Feature N

Tokenizer / Feature Selection

Wort1, Wort2, Wort3 …Wort K

Ein Text-Classifier wird trainiert mit Texten einer gewünschten Art

Danach gibt er dann hohe Scores aus für sehr ähnliche Texte

Er gibt niedrige Scores aus für sehr unähnliche Texte

Für die Texte aus den Trainingsdaten gibt er den maximalen Score aus

Text Klassifikation

Textart-Erkennung

Wort1, Wort2,
Wort3 .....
....
WortN

Klassifikator Textart A

Hoher Score für sehr ähnliche Texte zu **A**
Niedriger Score für alle anderen Texte

Klassifikator Textart B

Hoher Score für sehr ähnliche Texte zu **B**
Niedriger Score für alle anderen Texte

Klassifikator Textart C

Hoher Score für sehr ähnliche Texte zu **C**
Niedriger Score für alle anderen Texte

MAX

Text Klassifikation

Textart-Erkennung

Text-Classifier können für viele Zwecke verwendet werden

Dokumente in verschiedenen Sprachen -> Sprachenerkennung
(Eine Seite pro Sprache reicht zum Training)

Verwenden nur von Dokumenten-Überschriften -> Themenerkennung

Beliebige Vorverarbeitungen / Tokenizing

Ansonsten Erkennung von

- SPAM Mails
- E-Mail Kategorien für Weiterleitungen
- Dokumentenarten für Arbeitsflusssteuerung

Ist eine sehr generelle universell einsetzbare Technik zum Trennen von Textsorten

Typische Szenarien mit Übersetzern

Maschinelle Übersetzung
- Training mit Satzpaaren und Abschnittspaaren

Dialogführung
- Frage und Antwort werden als Übersetzung trainiert

Code Generieren
- Anforderungen werden übersetzt in Code Stücke

- Übersetzer Szenarien sind meistens komplexe Abbildungen
- Übersetzer haben daher oft sehr viele interne Zustände, um viel Information zu speichern
- Übersetzer benötigen extrem viele Trainingsdaten
- Übersetzer benötigen extrem viel Rechenleistung

- Sie sind nicht geeignet, wenn wenig Trainingsmaterial vorliegt

Aufgabenstellungen mit weniger Trainingsdaten (Beispiele)

portamis
Automatic Layout

XML Tagging

- Weil XML Contents sehr unterschiedlich getaggt sind, liegen für spezielles XML Tagging meistens nur wenige Trainingsdaten vor

Dokumenten-Clustering unbekannter Dokumente

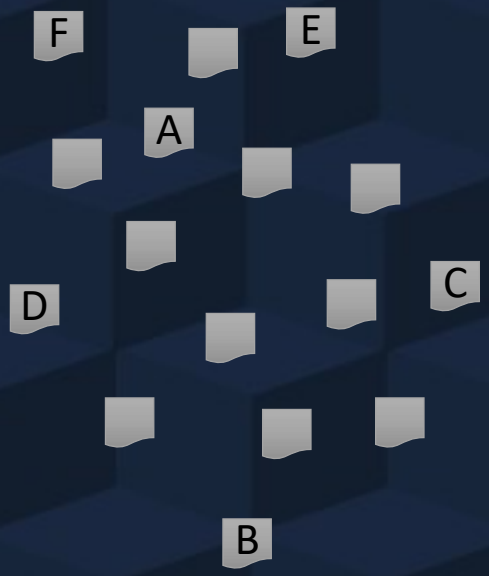
- Alte PDF Bestände sind oft unübersichtlich und keiner weiß mehr, was drin ist

Finden von falsch getaggten Contents in speziellen XML Content Datenbanken

- Verwenden von Classifiern mit wenig oder keinen vorab Trainingsdaten
- Baukastenprinzip:
  - Kombination von Bootstrapping: Sicher erkannte Bereiche prüfen und Trainingsmenge inkrementell erhöhen
  - Kombination von Classifiern mit Clusteringverfahren

Text Klassifikation

Clustering von Texten



1. Nimm ein Startdokument A und trainiere Modell A mit A als Trainingsdaten
2. Bewerte alle Dokumente und nimm das unähnlichste als B
3. Trainiere Modell {A U B} , bewerte alle Dokumente und nimm das unähnlichste als C
4. Trainiere Modell {A U B U C], bewerte …….. Und so weiter für K Cluster
5. Sortiere in die K Cluster die restlichen Dokumente ein, immer nach der größten Ähnlichkeit
6. Wenn ein Dokument hinzukommt, trainiere das Modell neu mit den neuen Daten hinzugefügt.

Nach welchen Kriterien geclustert wird, entscheidet die Vorverarbeitung

- Wenn nur Produktnummern und Produktnamen gefiltert herausgefiltert werden, wird danach geclustert.
- Bei allen Wörtern kommen tatsächlich Gattungen heraus.

Clustering mit Classifiern

portamis
Automatic Layout

Die Vorgehensweise

- Natives Verfahren lernt während des Clusterns
- erfordert sehr schnelle Trainingsläufe
- -> kein Deep Learning möglich, weil so viele wiederholte Trainings
- Verfahren wird derzeit getestet.

Verbesserungen:

Start-Cluster mit Trainigsdaten versehen

- Startcluster vorsehen, die manuell gewählte Dokumente enthalten

Bootstrapping (manuelle Kontrolle und Restart)

- Reste-Cluster generieren, bis ein nächstes entferntes Dokument näher an den Startclustern liegt als an einem Reste-Cluster
- Reste-Cluster manuell sichten und ggf. wiederholen mit zusätzlichen Start-Clustern

Automatisches XML Tagging

Funktioniert gut, wenn Daten vorsegmentiert sind

- z.B. bei Übernahme von Texten aus gelayouteten PDF Dokumenten
  - Tabellen, Listen, Überschriften sind hier schon segmentiert
- Bei semantischem Retaggen / Anreichern von vorhandenem XML Beständen

Funktioniert nicht gut bei rein Layout- oder Prozess-orientiertem Markup

- Weil der statistische Bezug von den Wörtern zur Kategorie schwach ist.

Verfahren:

- Trainingsdaten mit getaggtem Text. Für jede Kategorie wird ein Classifier trainiert auf den darin enthaltenen Wörtern.
- Vorhandene Segmente (aus dem Layout oder aus vorhandenem XML) werden klassifiziert.

Alternativ ohne Segmentgrenzen:

- Alle möglichen Segmente erzeugen und klassifizieren. Dann die beste Sequenz von Elementen auswählen.

Fazit

Bei einigen Anwendungen für Text-Dokumente und XML Contents sind Trainingsdaten nicht ausreichend vorhanden und erstellen ist teuer.
Daher sind Übersetzer (insbesondere Deep Learner) oft nicht anwendbar

Mit Classifiern in einem Baukasten von Suche oder Clustering können auch mit wenig Trainingsdaten Aufgaben mit Text-KI bearbeitet werden.

Dokumenten-Klassifikation, Automatisches XML Taggen, Dokumenten-Clustering kann damit effektiv automatisiert werden.

Das ist möglich, wenn wir nicht eine komplexe Übersetzung direkt lernen, sondern Classifier wie in einem Baukasten verwenden.

portamis: Automatic Layout for XML Content

portamis
Software GmbH

Erlangen-
Nuremberg
Germany

Founded in
2002

Dr. Hans Weber
Thimo Seitz

12 People

Siemens
LHS
Daimler
Philips
Roche
EDE
Hager
Berker
Brose
…

20 years of experience

Kontakt

Ute Heinz / Dr. Hans Weber
portamis Software GmbH

90491 Nürnberg
Äußere Sulzbacher Straße 159-161

Tel. +49 911 311 0977 0
E-Mail: info@portamis.de

www.portamis.de

portamis
Automatic Layout