

Rheinische Friedrich Wilhelms Universität Bonn
Institut für Kommunikationsforschung und Phonetik
Hauptseminar: Information Extraction
Leitung: Prof. Dr. W. Lenders
Wintersemester 2004/05
Protokoll vom 10.02.2005
Referent: Sebastian Kirsch
Protokoll: Lijian Huang

Thema: Kern-Methoden zur Extraktion von Informationen

Drei Verfahren zur Extraktion der Informationen:

- 1) Ad-hoc-Verfahren (endliche Automaten, reguläre Ausdrücke) ist ein von Hand erzeugte Muster, nach denen der Text abgesucht wird.
- 2) generative Modelle (Hidden Markov Models, Conditional Random Fields) generieren aus Beispielen Modell für die Erzeugung eines Textes und schließen daraus auf die Funktion seiner Bestandteile.
- 3) diskriminative Methoden (Maschinelles Lernen, Support Vector Machines, Voted Perceptron) unterscheiden zwischen Beispielen mit oder ohne der gewünschten Informationen.

Die Repräsentation für natürlichsprachliche Texte

- 1) *bag-of-words-Modell*:
 - Standardmodell für Texte bei maschinellen Lernverfahren;
 - Worthäufigkeit wird in einem Vektor gespeichert und miteinander verglichen.
 - leicht mit gängigen maschinellen Lernverfahren zu verarbeiten
 - gute Erfolg bei der Kategorisierung von längeren Texten
 - Struktur der Sätze geht verloren
 - für einzelne Sätze und damit für Extraktion von Beziehungen ungeeignet
- 2) Flache Syntaxbäume
 - Standardrepräsentation in der Computerlinguistik
 - Lassen Rückschluss auf Struktur des Satzes zu
 - Mit maschinellen Lernverfahren üblicherweise nicht zu verarbeiten

Maschinelles Lernen auf Bäume

1) Konventionelle Methode

- Bestimmte Strukturen in Syntaxbaum suchen und in Vektor notieren, ob diese vorkommen.
- Diese Vektor als Eingabe für Lernverfahren benutzen

2) Kern-Methode

- arbeiten direkt auf Baum-Repräsentation
- zwei Komponenten:

Lernverfahren, das Kernfunktionen einsetzt, ist unabhängig von Eingabeformat und Ausgabe.

Kernfunktion für die Eingabe-Repräsentation, ist abhängig von Eingabeformat und Ausgabe.

Zwei Lernverfahren auf Basis von Kernfunktionen

Support Vector Machines

- versucht Beispiele durch Hyperebene mit maximalem Abstand zu den beiden Klassen zu trennen.
- Die Form der Funktion $f(x) = \text{sign}(w \cdot x) + b$
- Quadratische Programmierung ist für die Lösung notwendig.

Voted Perceptron

- hat ähnliche Charakteristiken wie der SVM-Algorithmus
- basiert auf dem Perceptron-Algorithmus
- die Form der Entscheidungsfunktion: $f(x) = \text{sign}(w \cdot x)$

Kernfunktionen

Bestimmte Lernverfahren auf Vektoren lassen sich so formulieren, dass sie nur Skalarprodukte benutzen, und Merkmale in Vektor extrahieren und dann Vektoren multiplizieren.

Kernfunktionen berechnen implizit das Skalarprodukt zwischen Vektoren mit extrahierten Merkmalen. Jede symmetrische und positiv definierte Funktion ist eine Kernfunktion. Der

Zahl der implizit verwendeten Merkmale kann sehr hoch sogar möglicherweise auch unendliche groß sein.

Kernfunktionen auf Bäumen

Kernfunktion dient als eine Art verallgemeinertes Abstandsmaß zwischen den Instanzen. Sie soll auf eine gewisse Weise die Ähnlichkeit zwischen Bäumen ausdrücken.

Beispiel für Bäume:

- Kernfunktion zählt die gemeinsam vorkommenden Unterbäume.
- Kernfunktion wird beim Parsing eingesetzt, um korrekt geparste Sätze von nicht korrekt geparsten Sätzen zu unterscheiden.
- Kernfunktion kann die Beziehungen zwischen bestimmten Unterbäumen extrahieren, die in syntaktischer Kategorie und möglicher Rolle übereinstimmen, und zählt bei diesen Unterbäumen, auf wie vielen Wörtern sie übereinstimmen.
- Dabei sind Einschübe in den Unterbäumen erlaubt, und Einschübe werden als geringere Ähnlichkeit gewertet.