5.4 Textklassifikation

Thomas Brückner

Mit der beginnenden Informationsflut des IT–Zeitalters in den 90er Jahren stiegen auch die wissenschaftlichen Aktivitäten auf dem Gebiet der Textklassifikation. Die Klassifikation natürlichsprachlicher Texte in vordefinierte Kategorien entwickelte sich in den letzten Jahren zu einem der aktivsten Forschungsthemen in der Computerlinguistik und dem Maschinellen Lernen. Ein klassisches Beispiel für Textklassifikation ist die Einordnung von Zeitungsnachrichten in Rubriken wie Politik, Wirtschaft, Kultur, Sport, usw. Dieses Beispiel zeigt auch schon einen der wesentlichen Vorteile klassifizierter Textbestände im Zusammenhang mit der Informationsflut: einen leichteren Zugang zur gewünschten Information.

5.4.1 Generische System-Architektur

Wesentliches Ziel der Textklassifikation ist es, die inhaltliche Erschließung von, vor allem großen, Textmengen zu automatisieren. Im Allgemeinen hat man es hierbei mit komplexeren Klassifikationsaufgaben als bei dem obigen Zeitungsrubriken-Beispiel zu tun. Eine wesentlich größere Anzahl von Klassen, die eventuell noch hierarchisch angeordnet sind, sowie die Möglichkeit der Einordnung eines Textes in mehrere Klassen bzw. gar keine Klasse sind typische Anwendungs-Charakteristika.

Ein **Textklassifikations–System** besteht aus einer Komponente zum Wissenserwerb und einer Komponente zur eigentlichen Klassifikation. Bei der Komponente zum Wissenserwerb werden Klassenprofile erstellt, die vom Klassifikations–Algorithmus als Wissensbasis genutzt werden. Ein Klassenprofil (oder auch Modell) besteht üblicherweise aus einer Menge von Merkmalen und Gewichtungen bzw. Beziehungen zwischen den Merkmalen. Die Merkmale basieren meist auf Wörtern oder Buchstaben–N–Grammen.

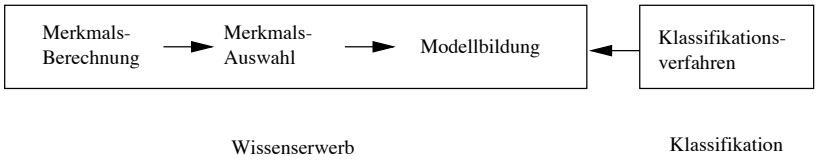

Abbildung 5.5: Prinzipieller Aufbau von Textklassifikations–Systemen

Für die Vorverarbeitung der Texte wird bei der Benutzung von Wort-Merkmalen oft eine flache syntaktische Analyse eingesetzt. Diese kann von der Lemmatisierung über das Part–of–Speech–Tagging bis hin zum NP-Parsing (siehe Unterkap. 3.3) gehen. Im Zusammenhang mit Mehrwort–Merkmalen finden auch Techniken zur Word Sense Disambiguation Anwendung. (siehe auch Unterkap. 4.3)

5.4.2 Verschiedene technologische Lösungsansätze

Die verschiedenen Typen von Textklassifikations-Systemen können auf zwei Ebenen unterschieden werden.

Unterscheidung nach der Art und Weise, wie die Klassenprofile erstellt werden

Hierbei unterscheidet man zwischen lernenden und nicht–lernenden Systemen. Bei den nicht–lernenden Systemen werden die Klassifikationsprofile manuell erstellt. Hierzu werden üblicherweise von menschlichen Experten Regeln für die einzelnen Klassen definiert.

Demgegenüber berechnen lernende Systeme die Klassifikationsprofile automatisch anhand von Trainingsbeispielen. Bei den Trainingsbeispielen handelt es sich um eine möglichst repräsentative Auswahl bereits klassifizierter Texte. Das Training besteht im Wesentlichen aus zwei Schritten: der Merkmals-Auswahl und der Berechnung des Klassifikationsprofils. Die Merkmals-Auswahl dient dazu, die Grundgesamtheit der zu berücksichtigenden Merkmale auf die wichtigsten zu reduzieren, um die Komplexität der Berechnung der Klassifikationsprofile möglichst gering zu halten. Für die Gewichtung der Merkmale sind verschiedene Maße gebräuchlich. Die wichtigsten Maße sind:

- **TF/IDF-Gewichte**, siehe hierzu Unterkap. 5.3, Seite 487.
- **Entropie-Maße** – Die Entropie ist ein Begriff aus der Informationstheorie, welcher den mittleren Informationsgehalt eines Nachrichtensymbols (hier eines Wort– bzw. N–Gramm–Merkmals für eine Klasse) in Bit bezeichnet.
- **Korrelations-Maße** geben den Grad (statistische Abhängigkeit) an, in dem das Vorkommen eines statistischen Merkmals (hier Wort– bzw. N–Gramm–Merkmal) die Wahrscheinlichkeit des Vorkommens eines anderen Merkmals (hier Klasse) beeinflusst.

Welche Merkmale dann tatsächlich ausgewählt werden, wird entweder durch eine fest vorgegebene Anzahl (die N Merkmale mit dem höchsten Gewicht), einen empirischen Schwellwert (alle Merkmale, deren Gewicht größer als ein vorgegebener Schwellwert ist) oder statistische Testverfahren (dynamisch berechnete Schwellwerte, ab denen von einer statistischen Signifikanz ausgegangen werden kann) entschieden.

Eine vergleichende Aussage über die Güte dieser Maße im Zusammenhang mit der Textklassifikation kann pauschal nicht getroffen werden. Dies hängt insbesondere auch stark von dem verwendeten Klassifikationsverfahren ab.

Unterscheidung nach dem der Klassifikation zugrunde liegenden Verfahren

Hinsichtlich des eigentlichen Klassifikationsverfahrens kann man zwei Typen von Systemen unterscheiden: Systeme, bei denen die Klassifikation durch Auswer-

tung von Regeln erfolgt, oder Systeme, bei denen die Klassifikation als Funktion in Abhängigkeit von der Gewichtung der Merkmale berechnet wird (statistische Verfahren und Neuronale Netze).

Regelbasierte Verfahren: Regeln werden meist in Form von booleschen Anfragen benutzt, bei denen der Zusammenhang zwischen Wörtern und Klasse durch die logischen Operatoren AND ($\wedge$), OR ($\vee$) und NOT ($\neg$) ausgedrückt wird. So könnte z.B. die Regel für eine Klasse „Straßenverkehr“ ansatzweise wie folgt aussehen:

(„auto“ $\vee$ „motorrad“ $\vee$ „autobahn“ $\vee$ „verkehr“ $\vee$ „straße“) $\wedge \neg$ („zug“ $\vee$ „bahn“ $\vee$ „eisenbahn“)

Das Klassifikationsverfahren ist hier bereits in den Regeln enthalten: ein Text wird nur dann einer Klasse zugeordnet, wenn er deren Regel erfüllt. Dieses Vorgehen ermöglicht auch die Mehrfachklassifikation eines Textes. Zur automatischen Erzeugung von Regeln werden häufig **Entscheidungsbäume** verwendet. Die Grundidee ist, dass jeder Knoten von der Wurzel beginnend eine einfache Bedingung über das Vorkommen eines Merkmales im Text enthält. Ist die Bedingung erfüllt, wird ein Zweig des Baumes weiterverfolgt, wenn nicht, der andere. Dies wird wiederholt, bis ein Blatt erreicht ist. Dieses ist dann mit der oder den entsprechenden Klassen markiert. Beim Lernen eines Entscheidungsbaums wird dieser von der Wurzel bis zu den Blättern aufgebaut, indem man für den jeweiligen Knoten das Merkmal mit der höchsten Entropie nimmt, das zusammen mit dem Eltern–Merkmal vorkommt bzw. nicht vorkommt.

Zur Verdeutlichung hier noch ein einfaches Beispiel:

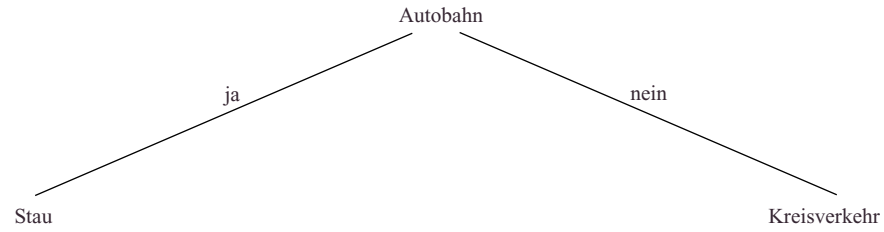

Abbildung 5.6: Ein einfacher Entscheidungsbaum

Dieser Beispielbaum entspricht der booleschen Regel („stau“ $\wedge$ „autobahn“) $\vee$ („kreisverkehr“ $\wedge \neg$„autobahn“).

Die Schwierigkeit dieses Verfahrens liegt in der Trainingsphase, die sehr aufwändig ist. Außerdem muss man verhindern, dass ein Entscheidungsbaum zu spezifisch gewählt wird, also eigentlich nur noch genau die Trainingstexte klassifizieren kann. Dieses Problem nennt man auch **Overfitting**. Auf der positiven Seite zeichnen sich gute Entscheidungsbäume durch eine hohe Güte der Klassifikation sowie hohe Effizienz in der Anwendungsphase aus (vgl. *Weiss et al.* 1999).

Statistische Verfahren: Grundlegende Lernverfahren für Klassifikationssysteme, die auf Merkmals–Gewichten beruhen, sind:

- **Rocchio–Algorithmus:** Ein sehr einfaches Verfahren, das von J. Rocchio Anfang der 70er Jahre zum Relevance-Feedback im Information-Retrieval entwickelt wurde (*Rocchio* 1971). Der **Rocchio-Algorithmus** basiert auf der Verwendung von Zentroid-Vektoren zur Repräsentation einer Klasse. Während der Trainingsphase wird zu jeder Klasse der Zentroid-Vektor aller Trainingstexte berechnet. Zur Ermittlung der Klasse eines neuen Dokumentes wird einfach die Distanz zu den Zentroid-Vektoren aller Klassen berechnet und die nahe liegendste (bzw. bei Mehrfachklassifikation die nahe liegendsten) ausgewählt. Die Vorteile dieses Verfahrens liegen in seiner Einfachheit und hohen Geschwindigkeit. Der Hauptnachteil liegt in der stark nachlassenden Güte des Verfahrens bei mehr als einigen hundert Klassen.

- **k-Nearest-Neighbour/kNN–Algorithmus:** Ein weiteres, ebenfalls recht einfaches Verfahren basiert auf der Ermittlung der k nächsten Nachbarn eines Textes. Beim Training werden lediglich die Merkmalsvektoren der Trainingstexte mit den zugehörigen Klassen abgespeichert. Zur Klassifikation eines neuen Textes wird dessen Merkmalsvektor mit allen Trainingsvektoren verglichen und die k nächsten ermittelt. Eine „Abstimmung", zum Teil wirklich in Form einer einfachen Mehrheitsentscheidung implementiert, ermittelt dann aus den Klassen der k nächsten Texte die des neuen Textes. Die Vorteile dieses Verfahrens liegen in der Geschwindigkeit des Trainings und der guten Skalierbarkeit bzgl. der Anzahl von Klassen.

- **Support-Vector-Machine/SVM:** Eine der aktuell vielversprechendsten Methoden ist die SVM. Die SVM ist ein komplexes mathematisches Verfahren zur Mustererkennung, das von T. Joachims auf das Problem der Textklassifikation angewandt wurde (*Joachims* 1999). Es ist vektorbasiert und berechnet eine Hyperebene, welche eine optimale Trennung der positiven und negativen Trainingsbeispiele darstellt. Als Ergebnis des Trainings erhält man pro Klasse die Trainingsvektoren, welche der Hyperebene am nächsten liegen. Diese Vektoren werden auch **Support-Vektoren** genannt. Die Klassifikation eines neuen Textes erfolgt dann im Wesentlichen durch die Distanzberechnung zu den Support-Vektoren. SVM's sind schnell in der Anwendung, skalieren gut, sind wenig anfällig gegenüber dem Overfitting und haben mit die besten veröffentlichten Testergebnisse bzgl. der Klassifikationsgüte. Demgegenüber steht das aufwändige Training, da der Berechnung der Hyperebenen ein Optimierungsproblem zugrunde liegt.

- **naive Bayes:** Die Anwendung der **Bayes-Formel** (siehe Unterkap. 2.4) für die Textklassifikation wurde 1992 zum ersten Mal von D. Lewis untersucht (*Lewis* 1992). Hierbei werden die bedingten Wahrscheinlichkeiten $p(\text{Klasse} = K \mid \text{Merkmal} = M)$ der Merkmale benutzt, um die Wahrscheinlichkeit zu berechnen, dass der gegebene Text zur Klasse gehört. Die

bedingte Wahrscheinlichkeit wird aus den Trainingsbeispielen geschätzt. Eine Schwäche dieser Verfahren ist, dass die statistische Unabhängigkeit der Merkmale vorausgesetzt wird. Diese ist aber im Allgemeinen nicht gegeben. Die empirischen Ergebnisse sind auch deutlich schlechter als die der anderen hier vorgestellten Verfahren.

Vergleich der Verfahren: Regeln haben offensichtlich den Vorteil, dass sie nachvollziehbar und damit auch für den Menschen änderbar sind. Dies ist in der Praxis aber gar nicht so einfach, da Regeln für umfangreichere Klassen sehr komplex werden können. Ein Nachteil vieler regelbasierter Systeme ist, dass sie keine eindeutige Klassifikation (d.h. ein Text darf höchstens einer Klasse zugeordnet werden) vornehmen können. Bei der klassischen logischen Auswertung von Regeln wird ja nur angegeben, ob ein Text zu einer Klasse gehört oder nicht. Erfüllt der Text mehrere Regeln, so hat man im Gegensatz zu den auf Statistik beruhenden Systemen keine Information dazu, bei welcher Klasse sich das System am sichersten ist.

5.4.3 Evaluierung von Textklassifikations–Systemen

Evaluierungs–Ergebnisse von Textklassifikations–Systemen werden meistens in Recall und Precision angegeben. Um einen einzelnen Wert zum direkten Vergleich mehrerer Systeme zu haben, werden auch oft der Break-Even-Point bzw. das F-Maß verwendet. Für die Definition dieser Maße siehe Unterkapitel 5.3.

Für eine vergleichbare Evaluierung von Textklassifikations–Systemen gibt es verschiedene frei verfügbare Standard–Test–Kollektionen. Am häufigsten wird die Reuters–21578–Kollektion (*Lewis* 1992) verwendet, bei der es sich um eine Sammlung klassifizierter Wirtschaftsnachrichten handelt. Weiterhin erwähnenswert ist die Ohsumed-Kollektion (*Hersh et al.* 1994), basierend auf den Medical Subject Headings (MeSH). MeSH ist das kontrollierte Vokabular der National Library of Medicine zur Klassifikation medizinischer Informationsquellen.

5.4.4 Stand der Technik

Technologisch haben sich die Verfahren in den letzten Jahren enorm weiterentwickelt. Während die ersten Systeme auf der Reuters–Kollektion einen Break–Even–Point von ca. 0,65 erreichten, liegen die besten aktuellen Systeme bei ca. 0,87. Bedenkt man, dass ein Vergleich zweier Menschen im Normalfall auch keine 100%-ige Übereinstimmung ergibt und ein Ende der 80er Jahre bei Reuters entwickeltes und produktiv eingesetztes nicht–lernendes System einen Break–Even–Point von ca. 0,90 hatte (allerdings nicht auf derselben Test-Kollektion), so sind diese Ergebnisse recht brauchbar.

Man muss sich aber auch darüber im Klaren sein, dass diese guten Resultate noch nicht bei allen Klassifikations–Anwendungen erzielt werden können und es immer sensible Anwendungen geben wird, bei denen nur sehr geringe Fehlerraten akzeptabel sind. Nichtsdestotrotz gilt die Textklassifikation heute als eine der

zentralen Technologien für das Knowledge–Management und findet schon vielfältig Anwendung in Dokument–Management–Systemen und im Internet–Umfeld (z.B. Routing von E-Mails).

5.4.5 Perspektiven

Da Textklassifikations-Systeme normalerweise komplizierte Berechnungen auf sehr großen Merkmalsmengen durchführen, werden sie alleine schon durch die ständige Weiterentwicklung der Hardware profitieren. Das gilt vor allem auch für die Untersuchung tiefergehender semantischer Analysen im Bereich der Textklassifikation, die vor allem auch aus Performance–Gründen noch nicht so zum Zuge gekommen sind.

Was mögliche Erweiterungen angeht, so bietet es sich an, die Textklassifikation mit der Textzusammenfassung und der Informations-Extraktion (Unterkapitel 5.6 und 5.5) zu kombinieren. Der Vorteil nach bereits erfolgter Klassifikation liegt auf der Hand: Unter der Voraussetzung, dass die Klasse richtig ist, besitzt man aus dem Klassifikationsverfahren gesichertes Wissen über den Inhalt des Textes. Dieses dynamische Wissen sowie vordefiniertes klassenspezifisches Wissen (z.B. Attributierungs-Schemata) kann dann mit bestehenden Verfahren zur Textzusammenfassung und Informations–Extraktion kombiniert werden um bessere Ergebnisse zu erzielen.

5.4.6 Literaturhinweise

Als Einstieg in die Thematik eignet sich ganz gut die Sonderausgabe zum Thema Textklassifikation der Zeitschrift ACM Transactions on Information Systems (*ACM* 1994). Eine gute Übersicht über die Merkmals-Auswahl bietet *Yang und Pedersen* 1997, während man in *Goller et al.* 2000 auch näheres über einige Klassifikationsverfahren nachlesen kann. Zur Vertiefung der Evaluierung von Textklassifikations–Systemen empfiehlt sich *Lewis* 1995. Eine Sammlung von Arbeiten, die auf einem Workshop der AAAI–Konferenz 1998 vorgestellt wurden, findet man in *AAAI* 1998.