Automatische Spracherkennung

2 Grundlagen: der statistische Ansatz der Spracherkennung

Soweit vorhanden ist der jeweils englische Fachbegriff, so wie er in der Fachliteratur verwendet wird, in Klammern angegeben.

Beispiele von Computerkommandos und Skizzenanweisungen sind in `Schreibmaschinenschrift` wiedergegeben.

Statistik in der automatischen Spracherkennung

Ganz zu Beginn dieses Kurses wurde gesagt, dass sich ASR nicht programmieren lässt, sondern Verfahren des Maschinenlernens (machine learning) verwendet, um die Beziehung von kontinuierlichen Signalen und linguistischen Symbolen herzustellen.

Eine Möglichkeit für Maschinenlernen (nicht die Einzige!) ist die Vorstellung, dass das kontinuierliche Signal von einer Quelle kommt, deren Verhalten zwar nicht vorhersehbar ist, aber deren Eigenschaften sich mit Mitteln der Statistik beschreiben lassen.

Um eine Analogie aus der Physik zu bemühen: der Aufenthaltsort von sehr kleinen Einheiten wie Elektronen lässt sich nicht mehr mit Methoden der Makrophysik beschreiben. Daher verwendet die Quantenmechanik statistische Modellierungen, um den Aufenthaltsort eines Elektrons zu beschreiben: anstatt die Position genau anzugeben, gibt man für jeden Raumpunkt eine Wahrscheinlichkeit an, mit der sich das Elektron dort befindet.

Wenn ein unbekannter Sprecher ein 'a' spricht, können wir nicht voraussagen, wie das Sprachsignal dieses 'a' aussehen wird, weil die konkrete Form von zuvielen (größtenteils unbekannten) Faktoren abhängt. Wir können aber durch Beobachtung ein statistisches Modell aufbauen, welches die Wahrscheinlichkeit für bestimmte 'a'-Formen im Signal berechnet.

Im folgenden besprechen wir daher einige Grundlagen zum Thema 'Statistische Modellierung'.

Exkurs: Die Bedingte Wahrscheinlichkeit

Siehe beigefügte Folienausdrucke bzw. Präsentation mit Titel *Die Bedingte Wahrscheinlichkeit in der Spracherkennung.*

Bayes Formel

Alle statistisch basierten Spracherkenner lassen sich auf die berühmte Bayessche Formel zurückführen:

$$P(A|B)P(B) = P(B|A)P(A)$$

Dabei ist $P(A)$ die (diskrete) Wahrscheinlichkeit, dass das Ereignis A eintritt (Wert zwischen 0 und 1).
$P(A|B)$ ist die bedingte Wahrscheinlichkeit, dass das Ereignis A eintritt, unter der Voraussetzung, dass auch das Ereignis B eintritt bzw. eingetreten ist. In der englisch-sprachigen Literatur werden $P(A)$ und $P(A|B)$ oft als 'a priori' und 'a posteriori probability' bezeichnet.

$p(s)$ dagegen ist die Wahrscheinlichkeitsdichtefunktion (WDF) für eine kontinuierliche Zufallsgröße (z.B. ein Sprachsignal $s(t)$).

Übertragen auf den Fall der statistischen Spracherkennung:

Gesucht ist die wahrscheinlichste Wortkette[1] W' (diskret) unter der Voraussetzung (Bedingung), dass das Signal s (kontinuierlich) aufgenommen wurde:

$$W' = argmax_W(P(W|s))$$

(Die Funktion $argmax_W(\ldots)$ berechnet den Wert der Variablen W, für den die Formel $(\ldots)$ maximal wird.)

$P(W|s)$ lässt sich nicht direkt in einem statistischen Modell schätzen, weil man dazu für alle möglichen Signale s alle Wortketten W beobachten müsste (s ist ein kontinuierlicher Raum, wogegen W diskrete Wortketten sind, wenn auch unendlich viele). Man könnte zwar eine große Menge von Signalen s als Bedingungen beobachten, aber die Frage ist, wie man daraus Abschätzungen für die Wahrscheinlichkeit der Wortketten W ableitet[2]. Eine Messgröße s als Bedingung für eine bedingte Wahrscheinlichkeit ist ein theoretisches Konstrukt, dass sich in der Praxis nicht modellieren läßt, weil s nicht abzählbar ist wie die Menge der Wortketten.

Nach Bayes lässt sich $P(W|s)$ aber umformen zu

$$P(W|s) = \frac{p(s|W)P(W)}{p(s)}$$

und somit wird

$$W' = argmax_W(\frac{p(s|W)P(W)}{p(s)})$$

$p(s)$ ist die Auftretenswahrscheinlichkeit eines bestimmten Signals s. Da wir bei der Spracherkennung immer ein bestimmtes Signal betrachten, ist $p(s)$ für alle in Frage kommenden Wortketten W konstant und kann also bei der Max-Entscheidung weggelassen werden. Man könnte auch einfach sagen, $p(s)$ ist von der Variable W unabhängig. Damit reduziert sich unser Problem zu:

$$W' = argmax_W(p(s|W)P(W))$$

Um W' zu bestimmen, müssen wir also zwei Terme $p(s|W)$ und $P(W)$ berechnen können, und zwar für beliebige Signale s und Wortketten W.

Akustische Modellierung (acoustic modeling)

Der Term $p(s|W)$, also die Auftretenswahrscheinlichkeit eines Signals s unter der Bedingung, dass es von der Wortkette W stammt, können wir in Form von Wahrscheinlichkeitsdichtefunktionen (symbolisch klein $p()$) abschätzen; das geschieht normalerweise durch Beobachtung eines Trainingskorpus, bei dem ja durch geeignete Labelung oder Transkription die Kette der Worte für jedes Signal s bekannt ist. Hier ist die Bedingung, dass eine bestimmte Wortkette W auftritt, abzählbar (wenn auch theoretisch unendlich) und kann somit als Bedingung eingesetzt werden.

[1] Statt einer Wortkette kann es sich auch um beliebige andere Ketten von linguistischen Einheiten handeln, z.B. Silben oder Phoneme.

[2] außer der trivialen Beobachtung, dass die gesprochene Wortkette W' die Wahrscheinlichkeit 1 bekommt und alle anderen 0: $P(W = W') = 1$, $P(W \neq W') = 0$

Beispiel:

In einem Sprachkorpus finden wir 200mal den Satz 'Heute ist schönes Frühlingswetter' gesprochen von lauter verschiedenen Sprechern. Durch Beobachtung der dabei auftretenden 200 Merkmalsvektorfolgen könnten wir ein statistisches Modell generieren, dass die Wahrscheinlichkeit für beliebige Signale berechnet, wenn die Wortkette 'Heute ist schönes Frühlingswetter' gesprochen wird. (Die Beobachtungen werden dabei i.A. in parametrische statistische Modelle umgerechnet, z.B. Gauß-funktionen oder HMM.)
Nehmen wir nun an, wir haben solche Wortkettenmodelle nicht nur für diesen Satz generiert, sondern für viele Sätze, z.B. die 450 Sätze des PD1 Sprachkorpus $W_{1...450}$. Nun wird ein unbekannter Satz s gesprochen und wir wollen die Wahrscheinlichkeit berechnen, dass dieser Satz s von einem der bekannten $W_{1...450}$ stammt. Mit Hilfe der 450 trainierten Modelle berechnen wir also 450 Wahrscheinlichkeiten $p(s|W_{1...450})$. (Technisch heißt das, dass man das konkrete Signal s in 450 verschiedene Modellformeln einsetzt und jeweils einen Wahrscheinlichkeitswert berechnet.)

Nun bleibt aber noch der Term $P(W)$ in der Maximierungsformel.

Sprachmodellierung (language modeling, LM)

Das Sprachmodell bzw. language model $P(W)$, d.h. die diskrete Wahrscheinlichkeit, dass die Wortkette $W = w_1, w_2, w_3, ...w_n$ gesprochen wurde, wird in fast allen statistischen Verfahren zur automatischen Spracherkennung eingesetzt, um brauchbare Ergebnisse zu erzielen. Anschaulich repräsentiert das Sprachmodel das Wissen, welche Wortketten 'wohlgeformt' bzw. 'sinnvoll' sind und welche nicht (Syntax, Semantik, Pragmatik).

Beispiel:

Der Satz W_1 ='Heute ist schönes Frühlingswetter' ist syntaktisch richtig und semantisch sinnvoll; er wird also ein relativ hohes $P(W_1)$ haben.
Der Satz W_2 ='Niemand Alien für kein gestern liegen' ist syntaktisch falsch und auch semantisch sehr unwahrscheinlich; $P(W_2)$ wird also sehr klein sein (aber niemals Null!).

Dieses Wissen ist absolut unabhängig von der akustischen Realisierung, d.h. auch geschriebene Wortketten habe eine bestimmte Wahrscheinlichkeit, dass sie 'sinnvoll' sind.

$$P(W) = P(w_1, w_2, w_3, w_4, ..., w_n)$$

(n-gram Modell)

Vollständige n-gram-Modelle sind nicht realistisch, weil man jede nur denkbare Wortkette zunächst beobachten müsste, um ihre Auftretenswahrscheinlichkeit $P(W)$ zu schätzen. Daher schätzt man $P(W)$ durch eine geeignete Kombination von kürzeren Bigram- bzw. Trigram-Modellen:

Bigram:

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2)P(w_4|w_3)...$$

wobei $P(w_2|w_1)$ die bedingte Wahrscheinlichkeit ist, dass das Wort w_2 (Ereignis) nach dem vorherigen Aussprechen des Wortes w_1 (Bedingung) ausgesprochen wird.

Trigram:

$$P(W) = P(w_1, w_2)P(w_3|w_1, w_2)P(w_4|w_2, w_3)...$$

Bi- und Trigram-Modelle lassen sich durch Abzählen von Wortpaaren oder -trippeln in großen Textsammlungen abschätzen (sog. Konkordanzanalyse). Ein Bigram-Modell ist in der einfachsten Form eine riesige quadratische Matrix mit Wahrscheinlichkeitswerten für jede mögliche Wortpaarung.

In beiden Fällen werden allerdings statistische Bindungen, die weiter als ein (Bigram) oder zwei Wörter (Trigram) reichen, vernachlässigt (z.B. im Deutschen die syntaktischen Beziehungen zwischen Hilfsverben und Verben).

Skizze

`Blockschaltbild der ASR mit MU (p(s|W)) und LM (P(W))`

Um also automatische Spracherkennung mit Hilfe von Statistik zu realisieren, brauchen wir zwei Modelle, die auf realen (beobachteten) Daten trainiert wurden:
Ein akustisches Modell, welches uns für beliebige Kombinationen von Wortkette und Signal die bedingte Wahrscheinlichkeit abschätzt, dass dieses Signal von einer Person stammt, welche die Wortkette gesprochen hat, und
ein Sprachmodell, welches für jede beliebige Wortkette einer Sprache abschätzt, wie wahrscheinlich es ist, dass diese gesprochen wird.