Automatische Spracherkennung

4 Wissenschaftliches Arbeiten

In diesem abschließenden (hoorray!) Kapitel werden die wichtigsten Regeln und möglichen Fehlerquellen für das wissenschaftliche Arbeiten im Bereich der automatischen Spracherkennung besprochen.

Wissenschaftliche Arbeiten im Bereich der automatischen Spracherkennung zielen meistens auf eine *Verbesserung* der Leistung (Performanz, performance) eines bereits existierenden Spracherkennungsalgorithmus, seltener auch auf die Entwicklung eines neuen Algorithmus an sich, wenn dieser eine Verbesserung gegenüber traditionallen Verfahren verspricht.

Mit anderen Worten: wissenschaftlicher Fortschritt in diesem Gebiet orientiert sich an *messbaren* und *nachvollziehbaren* Verbesserungen der bisherigen Hypothesen, wie Spracherkennung funktionieren könnte. Da Verbesserungen immer relativ sind, impliziert dieses Vorgehen, dass man eigentlich immer zwei Experimente auf der gleichen Datengrundlage durchführt, und deren Ergebnisse vergleicht: Zum einen ein schon bekanntes (State-of-the-art-)Experiment und zu anderen ein neu entwickeltes/modifiziertes Experiment, bei dem die Hypothese besteht, dass es zu einer statistisch signifikanten Verbesserung kommt.

4.1 Quantifizierbare Messwerte

Messbare und damit vergleichbare Performanz ist sehr schwer zu ermitteln, weil bei einem ASR-System sehr viele Faktoren eine Rolle für die Performanz spielen. Die folgenden Beispiele sind mehr oder weniger anerkannte Messwerte:

- Satz-/Äusserungserkennungsrate:
  Auszählen der vollkommen korrekt erkannten Äußerungen in Bezug auf alle Äußerungen
- Worterkennungsrate (word accuracy)[1] mit Berücksichtigung von Einfügungen:

  $$WA[\%] = \frac{N - R - D - I}{N} 100\%$$

  mit:

  N : Anzahl der Wörter im Korpus
  R : Anzahl der vertauschten Wörter in der Erkennung
  D : Anzahl der fehlenden Wörter in der Erkennung
  I : Anzahl der eingefügten Wörter in der Erkennung
- Worterkennungsrate ohne Berücksichtigung von Einfügungen

  $$WC[\%] = \frac{N - R - D}{N} 100\%$$

  Begründung: Einfügungen sind für viele ASR-System, die nur bestimmte Schlüsselwörter erkennen wollen, nicht relevant (umstritten). Dieses tolerantere Fehlermaß wird in der Literatur of 'word correctness' genannt, um es von der strengeren 'word accuracy' zu unterscheiden.

[1] In der Literatur auch oft Wortfehlerrate (word error rate) = 100 - Worterkennungsrate

- Testsatzperplexität
  Maß für die 'Voraussagekraft' eines language models (LM); i.A. ist die Leistung eines LM besser je niedriger die Perplexität ist.
  Die Testsatzperplexität ist die durchschnittliche Anzahl der Wortsukzessoren, die ein LM in einem gegebenen Testkorpus berechnet. Beim statistischen LM wird diese durch die Entropien der beteiligten n-gram-Wahrscheinlichkeiten berechnet.

- Erfolgsrate (bei Dialogssystemen; umstritten)

4.2 Typische ASR Studie

Um sich auf diesem Gebiet wissenschaftlich korrekt zu verhalten, muss man sich an bestimmte Regeln halten, damit Ergebnisse valide und veröffentlichbar sind. Leider machen viele Nachwuchswissenschaftler immer wieder die gleichen typischen Fehler. Nachfolgend ein typisches Ablaufschema einer wissenschaftlichen Studie in der ASR:

Schema	*Beispiel*
Ausgangspunkt: Existierendes System	Worterkenner basierend auf der Modellierung von Wörtern durch Verkettung von Phonemen anhand eines linearen Lexikons
Validierung des bestehenden Systems anhand eines festgelegten Trainings- und Test-Korpus	Training des Systems, Optimierung der Parameter, Test ergibt Worterkennungsrate, diese gilt als Vergleichswert (oft 'baseline system' genannt)
Aufstellung einer Hypothese zur Verbesserung	Erweiterung der Aussprache im Lexikon durch nichtlineare Aussprachevarianten verbessert Worterkennung
Durchführung	Ermittlung von Aussprachevarianten aus handannotiertem Material (dieses darf nicht aus dem Testmaterial stammen!), Übersetzung in geeignete Struktur, Erweiterung des baseline systems
Experiment zur Falsifizierung der Hypothese	Wiederholung des baseline-Versuchs mit exakt demselben Test-Material und denselben Parametern; die einzige Änderung ist das Lexikon.
Ergebnis des Performanztests	Worterkennungsrate hat sich signifikant verbessert $\rightarrow$ Hypothese nicht falsifiziert. (Erfolg) Worterkennungsrate hat sich nicht signifikant verbessert $\rightarrow$ Hypothese falsifiziert. (Misserfolg)

Bei der Veröffentlichung von Ergebnissen (auch Falsifizierungen sind Ergebnisse!) sollte streng darauf geachtet werden, dass ein Leser der Publikation alle Informationen und Daten zur Verfügung hat, so dass er das Experiment nachgestalten kann. Der springende Punkt hierbei sind oft die verwendeten Daten, d.h. das Trainings- und Test-Korpus (s.u.). Vorzugsweise sollten nur Ergebnisse veröffentlicht werden, die auf Daten basieren, die öffentlich zugänglich sind. In den meisten Fällen ist dies leider nicht gegeben.

4.3 Typische Fehler beim wissenschaftlichen Arbeiten

Manchmal werden in der der Literatur Ergebnisse von Studien präsentiert, die beim genaueren Hinschauen nicht haltbar sind, weil typische 'Anfängerfehler' gemacht werden. Das ist sehr peinlich, weil eine Publikation nicht einfach zurückgezogen oder nachträglich korrigiert werden kann.
Im Folgenden die wichtigsten Fehler, die passieren können:

4.3.1 Falsche Aufteilung der Sprachdaten in Test-, Entwicklungs- und Trainings-Korpus

Gewöhnlich steht für ein Experiment ein Sprachkorpus zur Verfügung, das noch nicht in Test-, Entwicklungs- und Trainings-Korpus aufgeteilt ist. Bei der Einteilung der Daten ist Folgendes zu beachten:

- Alle drei Korpora müssen vollständig und nach den selben Prinzipien orthographisch transkribiert sein.
- Alle drei Korpora müssen sprecherdisjunkt sein, d.h. kein Sprecher darf in mehr als einem Teilkorpus vorkommen.
- Alle drei Korpora sollten Sprache desselben Typs (spontan, gelesen), aus derselben Domäne und unter den selben Aufnahmebedingungen enthalten (Akustik, Aufnahmetechnik). Wenn z.B. der Sprachkorpus sowohl Studio- als auch Telefon-Sprachaufnahmen enthält, müssen diese in gleichen Proportionen über die drei Korpora verteilt sein.
- Alle drei Korpora sollten die vom ASR-System modellierten linguistischen Einheiten einigermaßen representativ enthalten (z.B. sollten alle Phoneme oft genug vorkommen, dass statistische Aussagen möglich sind). Typisch sind mindestens 1 Stunde Sprache; für das Trainingskorpus oft sehr viel mehr.
- **Trainingskorpus (training set):** wird zur Abschätzung der akustischen Modellierung (z.B. Training der HMM oder ANN) **und** zum Training des Sprachmodells (language models) verwendet. (Für die Verbesserung des Sprachmodells können weitere Textkorpora hinzugezogen werden.)
- **Entwicklungskorpus (development set):** wird als Testmaterial zur Berechnung des Abbruchkriteriums im Training und zur Einstellung von Test-Parametern verwendet (z.B. Einfluss des Sprachmodells, Pruning Faktoren etc.).
- **Testkorpus (test set):** wird als **abschließendes** Testmaterial zur Beurteilung der Forschungshypopthese verwendet (Signifikanz-Test). Eine Optimierung auf dieses Datenset ist nicht zulässig.

4.3.2 Sprachmaterial entspricht nicht der Realität

Der häufigste, oft unvermeidbare Fehler beim praktischen Aufbau eines Spracherkennungssystems: Das Sprachmaterial, mit dem trainiert wird, wurde unter anderen Bedingungen aufgenommen, als sie in der realen Anwendung vorkommen.

Beispiele:

- Laborsprache vs. Feldaufnahmen

  Unter Laborsprache (lab speech) versteht man Sprachmaterial, dass unter stark kontrollierten Bedingungen aufgezeichnet wurden. Zum Beispiel wurden Hintergrundgeräusche eliminiert, der Abstand des Mikros zum Sprecher immer exakt eingehalten, die Lautstärke bei der Aufzeichnung von Hand ausgesteuert und die Situation, in der die Versuchsperson sich befand, ist artifiziell.

  Als Feldaufnahmen (field recordings) dagegen bezeichnet man Sprachmaterial, das im wirklichen Einsatz des Systems aufgezeichnet wurde.

Beide Arten von Material unterscheiden sich natürlich erheblich; daher ist vom statistischen Ansatz her zu erwarten, dass Modelle, die mit Laborsprache trainiert wurden, nicht die gewünschten Schätzwerte $p(s|W)$ liefern werden und damit den Musterkennungsprozess falsch modellieren. (Dasselbe gilt natürlich auch für Language-Modelle, die mit Textmaterial trainiert werden, dass nicht aus Feldaufnahmen stammt.) Es ist allerdings sehr schwierig, Feldaufnahmen in einer konkreten Situation der Benutzung eines Spracherkenners zu erzeugen, wenn es diesen Erkenner noch gar nicht gibt (er soll ja mit dem erhobenen Material erst trainiert werden!).

Eine Lösung dieses Dilemmas ist die sog. 'Wizard-of-Oz'-Methode (benannt nach dem bekannten amerikanischen Kinderbuch 'The Wizard of Oz'), bei der ein 'Wizard' (zu deutsch Zauberer), also ein Mensch, der aber der Versuchsperson verborgen bleibt, die Rolle des noch nicht existenten Spracherkennungssystems simuliert. Je nach Einsatzgebiet kann diese Simulation sehr schwierig und aufwändig werden, so dass in den meisten Fällen aus Kostengründen darauf verzichtet wird.

Dieser Fehler wird bei praktisch allen derzeit üblichen ASR-Systemen in Kauf genommen. Mit verschiedenen Methoden der Adaption während des Betriebes lassen sich die daraus resultierende schlechtere Performanz des Erkenners teilweise wieder ausgleichen (Adaption an den Kanal, an den Sprecher, an Hintergrundgeräusche, etc.)

- Gelesene vs. spontane Sprache (prompted vs. non-prompted vs. spontaneous speech)

  Der Einsatz eines ASR-Systems, das mit gelesener Sprache trainiert wurde, in einem Bereich, in dem die Benutzer frei sprechen (also praktisch alle Situationen, ausser dem Diktieren von Texten) führt fast immer zu drastischen Einbußen in der Performanz. Grund hierfür ist natürlich die starke Reduktion und Verschleifung der einzelnen Sprachlaute in spontaner Sprache. Besonders sog. Funktionswörter (function words), die wenig semantischen Gehalt tragen ('ein', 'ist', 'der', 'die', ...), werden in flüssiger Spache weitgehend reduziert bis hin zur vollständigen Elimination (z.B. 'und' in 'neunundneunzig'). Als Notlösung wird versucht, bei der Erhebung der Daten, die Spontanität wieder einzuführen, indem zwei Sprecher zusammen eine Aufgabe lösen. Manche Literaturstellen unterscheiden zwischen 'prompted speech' (= gelesene oder vorgesprochene Sprache), 'non-prompted speech' (= quasispontan in Spielsituation oder durch direkte Fragen provoziert) und 'spontaneous speech' (= Sprache, die ohne Wissen der Versuchspersonen aufgezeichnet wurde). Letztere ist ethisch nicht tragbar und kommt daher kaum vor.

- Unterschiedliche Raumakustik / Hintergrundgeräusche

  Raumechos verursachen in erster Näherung eine Faltung des Sprachsignals mit der Impulsantwort des jeweiligen Raumes (die Impulsantwort kann man 'hören', indem man z.B. kräftig in die Hände klatscht). Es ist sehr aufwändig, eine solche Faltung rückgängig zu machen, da sich die Impulsantwort je nach Position des Sprechers und des Mikrophons drastisch ändern kann.

  Häufigste Abhilfe zur Vermeidung von verschiedenen Raumakustiken in Trainings- und Testmaterial ist die Verwendung von Nahbesprechungsmikrophonen, bei welchen die Raumechos gegenüber dem primären Signal kaum noch eine Rolle spielen.

  Konstante Hintergrundgeräusche (z.B. ein Lüftergeräusch) bewirken eine additive Verformung des mittleren Spektrums und kann daher am einfachsten schon mit einer zeitlichen Ableitung der Merkmale im front-end kompensiert werden.

  Beispiel: Kanaladaption im front end:

  In erster Näherung wirkt sich der sog. Übertragungskanal vom Mikrophon zum A/D-Wandler (Mikrophon, Kabel, Verstärker, Anti-Aliasing-Filter (Nyquist!)) als eine Kette von linearen

Übertragungssystemen aus. Die gesamte Übertragungsfunktion ist die Multiplikation all dieser einzelnen Übertragungsfunktionen. Normalerweise ist diese über die Dauer einer Äusserung (1-20 sec) konstant. Hinzu kommen noch Hintergrundgeräusche z.B. Lüfterrauschen, Plattengeräusch, Maschinen, die ebenfalls in erster Näherung für die Dauer der Äusserung ein konstantes Spektrum aufweisen.

Das Problem hier ist, die additiven (z.B. konstante Geräusche) von den multiplikativen (z.B. Übertragungsfunktion eines Mikros) zu trennen. Wenn das möglich ist, können solche Einflüsse durch Subtraktion des mittleren Langzeitspektrums bzw. durch Normierung ausgeglichen werden.

- Verschiedene Korpora für das Training des Language Models

  Das Language Model (LM) dient der Abschätzung der Auftretenswahrscheinlichkeit von Wortketten. Üblicherweise werden dazu Bigram- oder Trigram-Modelle verwendet. Zum Training eines LM benötigt man sehr viel mehr Wörter als für das Training der akustischen Modelle. Daher wird meistens auf sehr große Text-Korpora zurückgegriffen, meistens Zeitungstexte.

  Die veruracht folgende mögliche Fehlerquellen:

  - Das Textmaterial stammt aus einer anderen Domäne
  - Das Textmaterial ist geschriebene Sprache (= gelesene Sprache) und eignet sich nur bedingt als Trainingsmaterial für die Spracherkennung.

Abhilfe: bei der Aufnahme der Sprachdaten so realistische Bedingungen wie möglich schaffen.

4.3.3 Veränderung von mehr als einem 'Parameter' im Baseline-System

In diesem Falle ist nicht sicher feststellbar, was die möglicherweise signifikante Änderung der Ergebnisse bewirkt hat.
Abhilfe: Immer nur einen Parameter verändern.

4.3.4 Verwendung von zu wenig Daten im Test

Ein zu kleiner Testkorpus führt zu nicht sicheren Ergebnissen (vgl. 'Chi-Quadrat-Test').
Statistisch heißt das, dass im Experiment die Anzahl der beobachteten Ereignisse zu klein ist, und damit das Konfidenzintervall von gemessenen Durchschnittswerte viel zu groß wird.
Beispiel:
Ein Baseline-System habe eine Worterkennungsrate (WA) von 75,6%
Nach einer Modifikation des Systems messen wir eine WA von 77,7% und damit eine scheinbare Verbesserung.
Das beobachtete Ereignis hier ist 'Wort erkannt / nicht erkannt'.
Der Test-Korpus enthält 1210 Wörter.
Nach dem Chi-Quadrat-Test ist die Änderung von 75,6% auf 77,7% bei 1210 Tests mit 0,0849 Wahrscheinlichkeit zufällig ($p = 0.0849$) und somit nach allgemeinem wissenschaftlichen Konsens nicht mehr signifikant.
Das bedeut aber nicht, dass das Experiment nicht erfolgreich sein *könnte*; lediglich die Anzahl der Tests (Wörter) ist zu gering, um es nachweisen zu können.
Abhilfe: Schon beim Entwurf des Experiments die Signifikanz-Intervalle abschätzen und geg.falls den Testkorpus vergrößern.

4.3.5 Zu oberflächliches Studium der Forschungsliteratur

In diesem Forschungsfeld gibt es hauptsächlich Veröffentlichungen auf den aktuellen grossen Konferenzen (z.B. 'Interspeech' oder 'ICASSP'). Es empfiehlt sich auf jeden Fall, vor einer Veröffentlichung diese Kongressbände anzuschauen. Noch besser ist es, jemanden zu konsultieren, der schon seit Jahren in diesem Gebiet forscht.

4.3.6 Fehler verursacht durch Softwarefehler

Das ist nie ganz auszuschließen und kommt - leider - immer wieder vor. Es ist (leider) logisch unmöglich zu beweisen, dass eine Software fehlerfrei ist. Die einzige Abhilfe ist dauernde Kontrolle und kritische Abschätzung, ob gemessene Werte realistisch sein können.

4.3.7 Überadaption

Unter Überadaption versteht man den Effekt, der immer dann eintritt, wenn ein Modell zwar eine bestimmte Menge von (Trainings-)Sprachmaterial perfekt nachbildet, dieses aber nicht repräsentativ genug für die eigentliche Anwendung ist.

Wir unterscheiden zwei Fälle:

1. Überadaption an das Trainingsmaterial

   Theoretisch sollte das Trainingsmaterial unendlich groß sein. Da dies aus praktischen Gründen nicht möglich ist, aber andererseits immer schnellere und specherintensivere Computer sehr große Modellierungen erlauben, kann es leicht passieren, dass die Anzahl der statistischen Parameter eines Systems in dieselbe Größenordnung kommt wie die Anzahl der Trainingssamples.

   Beipiel:

   HMM Spracherkenner mit 1000000 Modell-Parametern (Mittelwerte und Kovarianzen)
   Trainingsmaterial: 100 Sätze (= ca. 300 Sekunden)

   300 sec entspricht 30000 Merkmalsvektoren der Dimension 40 = 1200000 Werte

   In diesem Fall wird (bei angenommener Gleichverteilung) 1 Modellparameter aus 1,2 Werten des Trainingsmaterials geschätzt. Dies ist viel zu wenig für eine gesicherte statistische Abschätzung (und Generalisierung).

   In so einem Extremfall (sehr viele Modellparameter und sehr wenige Daten) wird das Modell für jede Beobachtung eine optimale Modellierung bereitstellen. Die resultierenden Modelle sind extrem 'scharf', d.h. liefern extrem gute statistische Voraussagen ($p(s|W)$) für Daten aus der Trainingsstichprobe, aber schlechte Voraussagen für alles andere.
   Die Folge: die Modelle sind nicht robust oder sie generalisieren nicht mehr.

   Besonders gefährlich ist diese Form der Überadaption bei der Verwendung von sehr großen neuronalen Netzen (weil diese sehr viele Parameter enthalten).

   Lösung:

   - die Anzahl der Parameter des Modells so wählen, dass sie um mindestens zwei bis drei Größenordnungen unter der Anzahl der Werte im Trainingsmaterials liegen.
   - während des Trainings regelmäßig anhand einer disjunkten Stichprobe (development set) testen, welche Leistung der Erkenner hat. Nimmt diese Leistung ab oder stagniert, das Training abzubrechen, bevor Überadaption eintritt (die Leistung nimmt mit weiterem Training wieder ab!).

2. Überadaption an das Testmaterial

   Die Performanz eines Erkenners wird nicht nur durch die Trainingsdaten beeinflusst, sondern auch durch eine Vielzahl von sog. globalen Parametern. Solche Parameter sind z.B.

   - Gewichtung zwischen Akustik und Sprachmodell
   - Wortende-Strafen (word end penalty)
   - Minimal-Wahrscheinlichkeiten für nie beobachtete Ereignisse

   Diese globalen Parameter können nicht durch ein Trainingsverfahren gelernt werden, sondern müssen anhand einer unabhängigen Stichprobe von Hand optimiert werden. Daraus folgt jedoch wieder, dass der Erkenner an das Testmaterial adaptiert wird, d.h. das seine globalen Parameter (ev. sogar Neuentwicklungen in der Topologie oder im Algorithmus selber) zwar optimal für das Testmaterial funktionieren, aber vielleicht nicht auf andere Sprache.

   Lösung:

   Verwendung eines Entwicklungs-Korpus (development test set) anstelle des Testmaterials. Erst ganz zum Schluss, wenn das System garantiert nicht mehr verändert wird, kann der Erkenner mit dem eigentlichen Test-Korpus evaluiert werden.

   (Wenn kein Entwicklungskorpus zur Verfügung steht, werden globale Parameter oft durch 'Test' auf dem Trainingskorpus optimiert. Das ist jedoch suboptimal, weil i.A. der Test auf das Trainingsmaterial so gut funktioniert (100%), dass eine heuristische Optimierung gar nicht mehr möglich ist.)

Abschließend sollte man noch bemerken, dass jeder Wissenschaftler mit den rudimentären Gesetzen der Wissenschaftstheorie vertraut sein sollte (B. Russel). Als sehr gut lesbare Einführung empfehle ich das Büchlein 'Das Sockenfressende Monster in der Waschmaschine' von Christoph Bördlein (Alibri Verlag).