

# Korpusbasierte neuronale Netze zur Simulation von Sprach- und Gedächtnisprozessen

Reginald Ferber

GMD — IPSI, Dolivostr. 15, 64293 Darmstadt, Tel: +6151-869 847,  
Fax: +6151-869 818, e-mail: ferber@darmstadt.gmd.de

## 1 Einordnung der Untersuchungen

Während die "Fähigkeit zu Lernen" bei neuronalen Netzen vielfach als Vorteil des Paradigmas gewürdigt wurde, wird die Konstruktion der Repräsentation des zu untersuchenden Materials häufig eher als "lästige Notwendigkeit" aufgefaßt. Dabei zeigt sich schon bei Minsky und Papert (1988), daß die "Nichtlernbarkeit" von XOR durch das Perzeptron natürlich auch als unzureichende Repräsentation des Problems aufgefaßt werden kann. Die Wahl einer geeigneten Repräsentation ist also eine notwendige Voraussetzung zur Lösung eines Lernproblems, sie kann aber auch schon wesentliche Teile der Lösung vorwegnehmen. In vielen Untersuchungen bleibt unklar, inwieweit Ergebnisse aufgrund der gewählten Repräsentation erzielt wurden, und inwieweit tatsächlich der neuronale Lernmechanismus für die Lösung verantwortlich ist (Seidenberg & McClelland 1989). Das gilt insbesondere dann, wenn die Trainings- und Testbeispiele "von Hand" in die entsprechende Repräsentation überführt werden. Ein solches Vorgehen mag für Untersuchungen gerechtfertigt sein, die die Entwicklung eines funktionstüchtigen adaptiven Systems zum Ziel haben, es erscheint aber in den Fällen problematisch, in denen neuronale Netze als (psychologische) Modelle zur Erklärung von Lernvorgängen verwendet werden sollen.

Teilweise wurde versucht, "reine" Lernsimulationen zu konstruieren, indem 0-1-Folgen als Material verwendet wurden, die abschnittsweise als ein Wort, seine Phonologie, seine Bedeutung und sein Kontext gedeutet wurden. Die einzelnen Abschnitte wurden mit Zufallsmustern gefüllt, bzw. gemäß angenommener Ähnlichkeit von Wörtern und Kontexten kombiniert oder abgeändert. (Kawamoto 1993, Rödel 1994). Diese Versuche lassen aber wesentliche Eigenschaften der Sprache außer acht bzw. ersetzen sie durch wohldefinierte Abstandmaße, die in der Sprache eben gerade nicht so einfach zu definieren bzw. isolieren sind.

Aus dem Problem der geeigneten Repräsentation leitet sich ein weiteres Problem vieler Simulationsuntersuchungen aus dem Bereich der Sprachverarbeitung her: in der Trainingsphase werden relativ wenige Trainingsbeispiele in immer der selben Repräsentation dargeboten und allenfalls in Reihenfolge und Häufigkeit variiert, während natürliche Sprache eine ungeheure Variabilität der Formen aufweist. Diese Art des Trainings ist zum einen notwendig, um den Repräsentationsraum klein zu halten, zum anderen, weil anderenfalls inhaltlich ähnliche Sätze oder Wörter auch ähnlich repräsentiert werden müssten. Die Bestimmung solcher Ähnlichkeitsmaße ist aber bis jetzt ein ungelöstes Problem.

Bei den drei Simulationen, die im folgenden vorgestellt werden, wurden die Gewichte der Netze aus großen Textkorpora gewonnen. Das heißt, das "Trainingsmaterial", daß dem Lernen zugrunde lag, waren natürlichsprachliche Texte und nicht speziell konstruierte Beispielsätze in speziellen Repräsentationen. Um die vorhandenen Korpora verwenden zu können und eine genügend große statistische Basis zu haben, mußten die Modelle ansonsten so einfach wie möglich gehalten werden. Es zeigt sich aber, daß auch mit diesen einfachen Modellen bereits brauchbare Ergebnisse erzielt werden können.

## 2 Korpusbasierte assoziative Modelle

Bei den beiden zuerst geschilderten Simulationen wurden einfache lineare Systeme verwendet, die als lokalistische spreading activation Modelle aufgefaßt werden können. Die Gewichte der Modelle wurden aus Kookkurenzdaten von Wörtern in großen Textkorpora berechnet.

### 2.1 Simulation des Assoziationsversuchs (Wettler, Rapp & Ferber, 1993)

1957 / 58 wurden von Russell & Meseck (1959) deutsche Assoziationsnormen erhoben. D.h. zu 100 Stimuluswörtern wurden die spontanen Antworten von 331 Versuchspersonen aufgezeichnet, die häufigste Antwort zu jedem Stimuluswort bestimmt und als Primärantwort bezeichnet.

Um diese Daten zu simulieren, wurde ein deutschsprachiger Korpus aus Zeitungsartikeln und psychologischen Abstracts mit ca. 21 Millionen Wörtern zusammengestellt. Aus den 2 012 Wörtern aus dem Russell & Meseck Experiment und den weiteren 63 344 Wörtern, die mindestens 10 mal im Korpus vorkamen, wurde ein Vokabular mit insgesamt 65 356 Wörtern gebildet. Für alle Stimuluswörter wurde ermittelt, wie oft sie

im Korpus mit einem Abstand von 12 oder weniger Wörtern zu einem der Vokabularwörter vorkamen. Aus diesen Häufigkeitsdaten wurden assoziative Gewichte gemäß der Formel

$$a_{i,j} = \frac{p(i\&j)}{p(i)} * \frac{1}{p(j)^{0.68}} = \frac{1}{p(i)} * \frac{p(i\&j)}{p(j)^{0.68}} \quad (1)$$

berechnet, die aus Reliabilitätsgründen als

$$w_{i,j} = \begin{cases} \frac{\#(i\&j)}{\#(j)^{0.68}} & \text{if } \#(j) > 110 \\ \frac{\#(i\&j)}{110} & \text{if } \#(j) \leq 110 \end{cases} \quad (2)$$

implementiert wurde. (Dabei bezeichnet  $p(i\&j)$  die Wahrscheinlichkeit des gemeinsamen Auftretens der Wörter  $i$  und  $j$  und  $\#(i\&j)$  die Häufigkeit des gemeinsamen Auftretens.  $p(i)$  und  $\#(i)$  bezeichnen entsprechend die Auftretenswahrscheinlichkeit und -häufigkeit für ein einzelnes Wort  $i$ )

Zu jedem Stimuluswort wurde eine Rangreihe der übrigen 65 355 Wörter gemäß ihres assoziativen Gewichts gebildet. Das Wort mit dem höchsten Gewicht wurde als vorhergesagte assoziative Antwort bezeichnet. Die Ergebnisse der Simulation sind in Abbildung 1 dargestellt.

Abbildung 1: Ergebnisse der Simulation des Assoziationsversuchs

Experiment		Simulation	
Mittlere Anzahl Primärantworten pro Vp	22,5	Vorhergesagte Primärantworten	19
Mittlere Anzahl anderer Vp, die dieselbe Antwort gegeben haben	28,8	Mittlere Anzahl Vps, die die vorhergesagte Antwort gegeben haben	22,9
Mittlere Anzahl der Antworten, die von keiner anderen Vp gegeben wurden	19,8	Vorhergesagte Antworten, die von keiner Vp gegeben wurden	57

## 2.2 Simulation der Suchwortwahl von Datenbankrechercheuren (Ferber, Wettler & Rapp 1995)

Zu 94 schriftlichen Anfragen an einen Literaturinformationsservice wurden von professionellen Rechercheurinnen und Rechercheuren Literatursuchen in Datenbanken durchgeführt. Aus allen Wörtern der Anfragen und der Queries wurde ein Vokabular von 872 Termen erzeugt. Für alle Paare aus diesen Termen wurde die Anzahl der Dokumente unter den 246.889 Dokumenten der Datenbank PsycLIT ermittelt, in denen sie gemeinsam vorkamen. Mit diesen Kookurrenzdaten wurde gemäß der Formel

$$a_{i,j} = \frac{p(i\&j)}{p(i) \cdot p(j)} - 1 \quad (3)$$

implementiert durch

$$w_{i,j} = \frac{\#(i\&j) \cdot 246\,889}{\#(i) \cdot \#(j)} - 1 \quad (4)$$

ein vollständig vernetztes lokalistisches spreading activation Netz mit den 872 Termen als Knoten konstruiert. Ziel der Untersuchung war es, bei Eingabe aller Wörter einer schriftlichen Anfrage mit diesem Netz im Mittel vorherzusagen, welche Terme in der Query verwendet wurden. Dazu wurden die Knoten der Terme, die in der schriftlichen Anfrage vorkamen, gemäß ihrer Häufigkeit aktiviert und ein spreading activation Zyklus durchgeführt. Danach sollten die Terme, die für die Query gewählt wurden, am stärksten aktiviert sein. Es wurden mehrere Zyklen dieses Prozesses berechnet, wobei die Aktivierungen der Anfrageterme addiert wurde. Zur Auswertung wurden die Terme gemäß ihrer Aktivierung in eine Rangreihe gebracht und die mittleren Rangplätze von drei verschiedenen Klassen von Termen berechnet. (1) den Termen, die in der Anfrage und in der Query vorkamen (P&Q), die also von den Recherchierenden für die Query ausgewählt wurden, (2) den Termen, die in der Query, aber nicht in der Anfrage vorkamen (–P&Q), die also neu hinzugewählt wurden, und (3) den Termen, die zwar in der Anfrage, nicht aber in der Query vorkamen (P&–Q), die also von den Recherchierenden verworfen wurden. Gute Ergebnisse zeichnen sich durch hohe Aktivierung und damit niedrige Rangplätze für die ersten beiden Klassen und niedrige Aktivierung und damit hohe Rangplätze für die dritte Klasse aus.

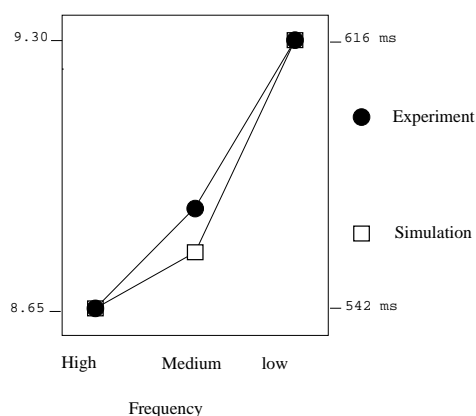
Für eine zusammenfassende Bewertung des Netzes wurden zunächst die 94 Rechercheprotokolle zufällig in zwei gleichgroße Teilsamples aufgeteilt. Mit dem ersten (dem Eichsample) wurden die Parameter des Netzes optimiert, mit dem zweiten (dem Testsample) wurde das Netz getestet. Bei einer Gültigkeit des Modells sollten die Ergebnisse für die beiden Samples ungefähr gleich sein. Die Ergebnisse der Simulation sind in Abbildung 2 dargestellt.

Abbildung 2 Mittlere Rangplätze der Termklassen bei der Suchwortwahl und ein Overlapmaß

Zykl	Eichsample:				Testsample:			
	P&Q	→P&Q	P&→Q	OVL	P&Q	→P&Q	P&→Q	OVL
1:	18.5	155.6	194.8	0.39	18.9	172.3	203.9	0.41
2:	24.5	156.0	216.3	0.38	22.9	180.4	227.9	0.40
3:	29.1	158.6	227.8	0.37	25.7	188.7	241.3	0.39
4:	32.2	161.2	235.0	0.37	27.7	194.7	249.9	0.38
5:	34.4	162.9	239.5	0.37	29.0	198.2	255.5	0.38

*OVL gibt den mittleren Anteil von der Simulation vorhergesagter Terme an den von den Recherchierenden verwendeten Termen an. Zwischen verschiedenen Recherchierenden fanden Saracevic & Kantor (1988) mit anderem Material einen mittleren Overlap von 0.27*

Abbildung 3: Simulation der Ergebnisse von Cosky (1976)



### 3 Simulation kognitiver Prozesse als Konvergenzprozesse

In zahlreichen experimentellen Untersuchungen sind Reaktionszeiten beim Erkennen von geschriebenen Wörtern zur Überprüfung von Modellen über die Organisation und Funktion des lexikalischen Gedächtnisses erhoben worden. Dabei sind sowohl Häufigkeitseffekte als auch Einflüsse des Kontexts, in dem ein Wort dargeboten wird, untersucht worden. Um solche experimentellen Daten zu simulieren, wurde ein einschichtiges rekurrentes "Brain State in a Box" Netz (Anderson, Silverstein, Ritz and Jones 1977) konstruiert und mit dem Brown Korpus trainiert. Die Knoten des Netzes repräsentierten die 1000 häufigsten Buchstabentupel bis zur Länge 5 im Brown Korpus, wobei nicht zwischen großen und kleinen Buchstaben unterschieden wurde. Zur Simulation wurden die Knoten der Zeichenketten, die in einem Wort vorkamen, aktiviert und dann solange Iterationen des Netzes durchgeführt, bis der Abstand zwischen zwei Aktivitätsmustern kleiner als 0.1 war. Die Anzahl der Iterationen wurde mit den Reaktionszeiten verglichen.

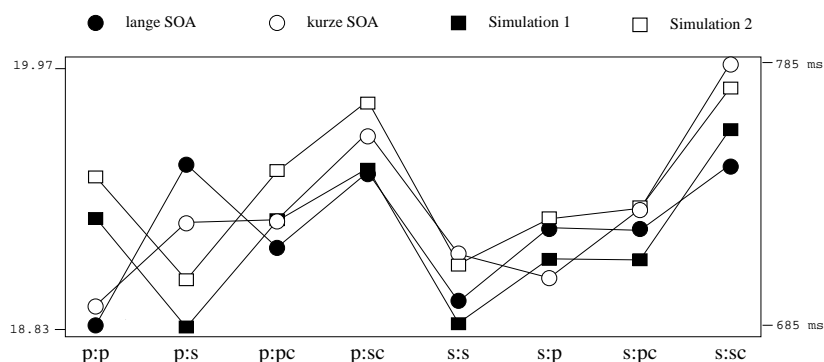
#### 3.1 Simulation von Worterkennungszeiten

Cosky (1976) fand Unterschiede in den Lesezeiten für häufige, mittelhäufige und seltene Wörter. Diese Unterschiede finden sich auch in der Simulation (vergl. Abbildung 3)

#### 3.2 Simulation von Kontexteinflüssen

Onifer und Swinney (1981) untersuchten den Einfluß mehrdeutiger Wörter mit unbalancierten Häufigkeiten der Bedeutungen, auf Worterkennungszeiten für Wörter, die mit einer der Bedeutungen verwandt waren oder nicht. Dabei hörten die Vp vor der Erkennungsaufgabe einen Satz, der das mehrdeutige Wort entweder in seiner häufigen (p=primären) oder in seiner seltenen (s=sekundären) Bedeutung enthielt. Anschließend mußten sie ein Wort erkennen, das zu einer der beiden Bedeutungen gehörte (p oder s) oder nicht (pc oder sc). Zur Simulation wurden zunächst die Knoten der Wörter aus dem Satz aktiviert und dieses Muster iteriert, anschließend wurden die Knoten des zu erkennenden Wortes aktiviert und die Anzahl der Iterationen bis zum Unterschreiten eines Abstands von 0.1 gezählt. Die Ergebnisse sind in Abbildung 4 dargestellt.

Abbildung 4: Zwei Simulationen der Ergebnisse von Onifer und Swinney (1981).



Bezeichnungen: *p* oder *s* vor dem Doppelpunkt: primäre oder sekundäre Bedeutung im gehörten Satz, nach dem Doppelpunkt: mit primärer oder sekundärer Bedeutung verwandtes Wort, *pc* und *sc* nach Häufigkeit gematchtes Kontrollwort in der Erkennungsaufgabe. Simulation 1 und Simulation 2 unterscheiden sich lediglich darin, daß in Simulation 1 die maximale Zahl der Primeschritte auf 15 beschränkt war, während in Simulation 2 lediglich bis zu 6 Primeschritte zugelassen waren. SOA bezeichnet das Zeitintervall zwischen dem auftreten des mehrdeutigen Wortes und der Erkennungsaufgabe.

## 4 Literatur

- Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* 84, 413-451.
- Cosky, M. J. (1976). The role of letter recognition in word recognition. *Memory and Cognition* 4(2), 207-214.
- Ferber, R., Wettler, M., & Rapp, R. An associative model of word selection in the generation of search queries. Accepted for publication by Journal of the American Society for Information Science (JASIS), 1995.
- Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language* 32, 474-516.
- Minsky, M. L., & Papert, S. A. (1988). *Perceptrons*. The MIT Press, Cambridge, Ma., (Expanded edition, first edition 1969).
- Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory and Cognition* 9(3), 225-236.
- Rödel, J. (1994). Neuronale Netzwerke als psychologische Modelle des episodischen Gedächtnisses in Listenlernparadigmen: Eine kritische Analyse konnektionistischer Methoden. Technical report, Universität Regensburg.
- Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III searchers, searches and overlap. *Journal of the American Society for Information Science* 3(39), 197-216.
- Seidenberg, M. S., & McClelland, J. L. (1989). Distributed, developmental model of word recognition and naming. *Psychological Review* No. 4, 523-568.
- Wettler, M., Rapp, R., & Ferber, R. (1993). Freie Assoziationen und Kontiguitäten von Wörtern in Texten. *Zeitschrift für Psychologie* 201, 99-108.

## 5 Stichworte

1. Lernen mit neuronalen Netzen
2. Simulation assoziativer und sprachlicher Prozesse
3. Korpusbasierte Methoden