Nichtmonotones Schließen und neuronale Netze

Reinhard Blutner, Humboldt-Universität Berlin

Kurzfassung

Der Beitrag faßt die Integration von symbolischen und subsymbolischen Ansätzen als ein Problem der *Vereinheitlichung* zweier anscheinend ganz unterschiedlicher Paradigmen auf. Die Zielstellung ist also nicht, zwei unterschiedliche Verarbeitungssysteme über eine geeignete Schnittstelle miteinander zu verkoppeln, sondern vielmehr, eine vereinheitlichende Darstellung zu finden, bei der symbolische und subsymbolische Verarbeitung als unterschiedliche *Perspektiven* einunddesselben kognitiven Vorgangs gedeutet werden können.

Die Grundidee wird am Beispiel von Hopfield-Netzen demonstriert. Zunächst wird gezeigt, daß Aktivierungszustände als Informationszustände aufgefaßt werden können, die durch eine Präzisierungsrelation strukturiert sind. Diese Relation stiftet unter gewissen Bedingungen einen deMorganschen Verband. Aktivierungszustände können demnach als Propositionen betrachtet werden. Weiter wird gezeigt, daß sich der Prozeß der Aktivierungsausbreitung (*asynchrones Updating)* als nichtmonotone Inferenz beschreiben läßt, wobei der zugrundeliegende Inferenzbegriff eine kumulative Logik begründet. Nimmt man eine *lokalistische Repräsentation* der Aktivierungszustände an (dabei wird eine transparente Zuordnung zwischen den Einheiten des Netzes und den atomaren Symbolen der Repräsentationssprache realisiert), dann läßt sich zeigen, daß es eine eineindeutige Zuordnung zwischen der Gewichtsmatrix von Hopfield-Netzen und einer (mit Gewichten versehenen) Datenbasis in Poole-Systemen gibt. Bezogen auf diese Korrespondenz bilden sich das asymptotische Verhalten der Aktivierungsausbreitung und Inferenzen in (gewichten) Poole-Systemen wechselseitig aufeinander ab.

Allgemeine Schlußfolgerungen: (i) Die demonstrierte Extraktionsmethode ermöglicht es einem Benutzer, die "Schlüsse" eines konnektionistischen Systems unmittelbar nachzuvollziehen. Sie werden als nichtmonotone Schlüsse in einer kumulativen Logik faßbar. (ii) Die Explosion des Verarbeitungsaufwands in traditionellen nichtmonotone Systemen bei großen Datenbasen läßt sich, jedenfalls unter gewissen Bedingungen, durch den Einsatz stochastischer Prozesse ("simulated annealing") vermeiden (implementativer Konnektionismus). (iii) Die vorgeschlagene Sichtweise liefert theoretische Aufschlüsse darüber, welche Arten von Logiken zur Beschreibung emergenter Eigenschaften neuronaler Netze dienen können.

1 Einleitung

Beim Vorliegen zwei einander widersprechender Theorien gibt es mindestens vier Möglichkeiten der Konfliktbewältigung: (ii) *Eine* Theorie setzt sich durch, indem sie die anderen Theorien eliminiert; (ii) *eine* Theorie setzt sich durch, indem sie gewisse Merkmale der anderen Theorie in sich aufnimmt; (iii) beide Theorien erkennen wechselseitig ihre Gültigkeit an, allerdings beschränkt auf gewisse Anwendungsdomänen, und sie versuchen gemeinsam zur Lösung komplexer Probleme beizutragen; (iv) beide Theorien werden im Rahmen einer neuen Theorie vereinheitlicht (Diese Vereinheitlichung kann in einer partiellen Reduktion bestehen: Reduktion der klassischen Thermodynamik im Rahmen der statistischen Physik. Oder sie setzt eine Neuschöpfung des theoretischen Paradigmas voraus, wobei sich die ehemals widersprechenden Theorien als unterschiedliche "Bilder" derselben Wirklichkeit erweisen: Auflösung des Welle-Teilchen-Dualismus im Rahmen der Quantenmechanik). Bezogen auf das Verhältnis zwischen Symbolverarbeitung und Konnektionismus nennen Smolensky u.a. (1992) diese vier Positionen (i) eliminativer Konnektionismus, (ii) implementativer Konnektionismus, (iii) hybrider Ansatz, (iv) integrativer Konnektionismus.

Ich möchte aus dreierlei Gründen die letztgenannte Position einnehmen. Erstens ermöglicht es diese Position, interessante Wechselbeziehungen zwischen Konnektionismus und algebraischer Semantik bzw. nichtmonotoner Logik sichtbar zu machen und für die Analyse

subsymbolischer Verarbeitungssysteme zu nutzen. Zweitens erhalten gewisse Systeme der nichtmonotone Logik dadurch eine besondere Fundierung und Auszeichnung gegenüber anderen Systemen. Und drittens eröffnet dieser Zugang die Möglichkeit, bei der Implementation (nichtmonotoner) symbolischer Inferenzsysteme auch auf Techniken subsymbolischer Systeme zurückzugreifen (z.B. stochastische Optimierungstechniken wie *simulated annealing*).

Der Plan für das weitere Vorgehen ist folgender. In Abschnitt 2 fasse ich Aktivierungszustände in Hopfield-Netzen als Informationszustände auf und definiere eine Präzisierungsrelation. Diese Relation erweist sich als eine Halbordnung, die einen deMorganschen Verband stiftet. Abschnitt 3 untersucht den Prozeß der Aktivierungsausbreitung (*asynchrones Updating)* im asymptotischen Grenzfall. Es wird gezeigt, daß sich dieser Prozeß durch einen nichtmonotonen, kumulativen Inferenzbegriff beschreiben läßt. Abschnitt 4 führt in kostenbasierte Poole-Systeme ein und beschreibt ihre Semantik mit Hilfe präferentieller Modelle. In Abschnitt 5 werden die Aktivierungszustände des Systems durch Ausdrücke einer elementaren Aussagenlogik repräsentieren. Auf dieser Grundlage wird der Zusammenhang zwischen der Kodierung von "Wissen" in der Verbindungsmatrix konnektionistischer Netze und der Repräsentation von Wissen in Poole-Systemen untersucht. Den Abschluß bildet Abschnitt 6 mit einer kurzen Betrachtung über den Wert derartiger Untersuchungen für die Verarbeitung natürlicher Sprachen.

2 Informationszustände in Hopfield-Netzen

Wir betrachten ein Hopfield-Netz mit dem Arbeitsbereich [-1,+1] für die einzelnen neuronalen Einheiten (+1: maximales Feuern; 0: Ruhezustand; -1: minimales Feuern).

Zustandsraum für n Neurone: $S = [-1, 1]^n$
Verbindungsmatrix: $-1 \le w_{ij} \le 1$, $w_{ij} = w_{ji}$, $w_{ii} = 0$
asynchrones Updating:

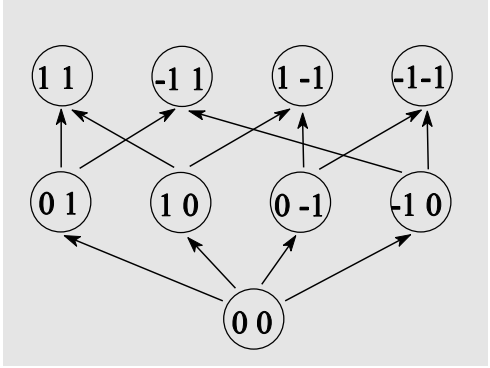
$$(1) \qquad s_i(t+1) = \begin{cases} \theta\,(\Sigma_j\, w_{ij} \cdot s_j(t), & \text{falls } i = \text{random}(1,n) \\ s_i\,, & \text{sonst} \end{cases}$$

Wenn wir voraussetzen, daß sich das System nur in Richtung höherer (positiver oder negativer) Aktivierung entwickelt, dann lassen sich Aktivierungszustände als Informationszustände deuten, die entsprechend ihrem Informationsgehalt geordnet sind:

Definition 1
$\langle S, \ge \rangle$ ist ein Poset von Informationszuständen gdw.
(i) $S = [-1, +1]^n$ (Zustandsraum)
(ii) $s,t \in S$: $s \ge t$ gdw. $s_i \ge t_i \ge 0$ oder $s_i \le t_i \le 0$, für alle $1 \le i \le n$.

Das Bestehen der Relation $s \ge t$ drückt aus, daß s informationell mindestens so reich (oder präzise) wie t ist. Als Element von Informationszuständen drückt der Wert 0 Unterspezifizierung aus, 1 und -1 drücken vollständige Spezifizierung aus. Informationszustände, deren Elemente nur 1 und -1 enthalten, heißen *total*.

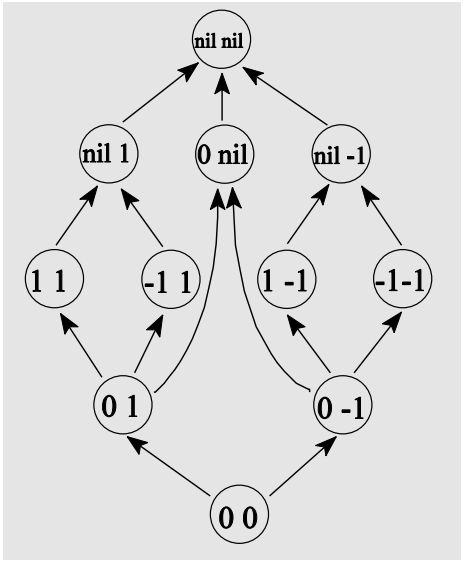
Beispiel: Poset von Informationszuständen für n=2.



Dieses Poset bildet allerdings noch keinen Verband. Jedoch ist es leicht möglich, das Poset zu einem Verband zu erweitern, wenn man "uneigentliche" Aktivierungen und "uneigentliche" oder "absurde" Informationszustände einführt. Ich schreibe "nil" für uneigentliche Aktivierungen.

Definition 2
$<S\cup\perp, \geq>$ ist ein erweitertes Poset von Informationszuständen gdw.
(i) $S = [-1, 1]^n$ (Zustandsraum)
(ii) $\perp = \{s: s_i = \text{nil für ein } 1\leq i\leq n\}$ (Menge der *absurden* Informationszustände)
(iii) $s,t \in S\cup\perp$: $s\geq t$ gdw. $s_i=\text{nil}$ oder $s_i\geq t_i \geq 0$ oder $s_i\leq t_i \leq 0$, für alle $1\leq i\leq n$.

Beobachtung 1
Das erweiterte Poset von Informationszuständen $<S\cup\perp, \geq>$ bildet einen deMorganschen Verband.



Dabei kann die Konjunktion $s\circ t = \sup\{s,t\}$ als *simultane Realisierung* der Zustände s und t verstanden werden, die Disjunktion $s\oplus t = \inf\{s,t\}$ als eine Art von *Generalisierung* und das Komplement s* reflektiert das Fehlen von Information.

Anmerkung: Diese Beobachtung ist die Verallgemeinerung eines Resultats von Balkenius & Gärdenfors (1991). Die Autoren untersuchen den binären Fall und finden eine Boolesche Algebra.

3 Asymptotische Updates von Informationszuständen

Sei f eine asynchrone (stochastische) Update-Funktion (vgl. den Ausdruck (1) für den Fall diskreter Zeit). Hopfield (1982) hat gezeigt, daß dann für das im vorigen Abschnitt charakterisierte System eine Ljapunov- oder Energiefunktion existiert, und zwar

(2) $E(s) = -\Sigma_{i>j}\, w_{ij}\cdot s_i\cdot s_j$.

Daraus ergibt sich, daß der Grenzwert $\lim_{n\to\infty} f^n(s)$ für jedes $s\in S$ existiert und einen stabilen Zustand (Resonanz) des Systems darstellt (lokales Energieminimum). Außerdem ist bekannt, daß im Falle gewisser stochastischer Update-Funktionen ("simulated annealing") der Zustand $\lim_{n\to\infty} f^n(s)$ ein *globales* Minimum der Energiefunktion darstellt (Cohen & Grossberg 1983).

3.1 Asymptotische *Updates* mit Klammerung

Im allgemeinen wird das *Updating* eines Informationszustands s einen Zustand f...f(s) ergeben, der nicht die durch s vermittelte Information einschließt (d.h., dieser Zustand ist nicht notwendig eine Präzisierung von s). Für das Folgende ist es jedoch wesentlich, das *Updating* eines Informationszustands s als Präzisierung zu fassen. Nach Balkenius & Gärdenfors (1991) läßt sich dies immer dadurch erreichen, daß die bereits im Ausgangszustand spezifizierten Werte (1 und -1) "festgeklammert" werden. Die geschieht technisch in folgender Weise:

Definition 3
Sei f eine asynchronische (stochastische) Update-Funktion (für ein Hopfield-System mit der Verbindungsmatrix w):
Update-Funktionen mit Klammerung:

$\underline{f}(s) = f(s) \circ s;$
$\underline{f^{n+1}}(s) = f(\underline{f^n}(s)) \circ s$

Asymptotische Update-Funktionen mit Klammerung

$ASUP_w(s) = \{t: t = \lim_{n \to \infty} \underline{f^n}(s)\}$

Der zufällige Charakter des *Updatings* bedingt, daß mehrere asymptotische *Updates* eines Ausgangszustands s auftreten können. Die Menge $ASUP_w$ (s) wird also im allgemeinen mehr als ein Element enthalten.

3.2 Energie-minimale Präzisierungen von Informationszuständen

Von einer etwas anderen Blickrichtung können wir diejenigen Präzisierungen eines Zustands s betrachten, die eine gewisse Kostenfunktion E minimieren:

Definition 4
Sei $<S, \geq>$ ein Poset von Informationszuständen, E eine reelle Funktion auf S. Die E-minimalen Präzisierungen von s sind wie folgt definiert:

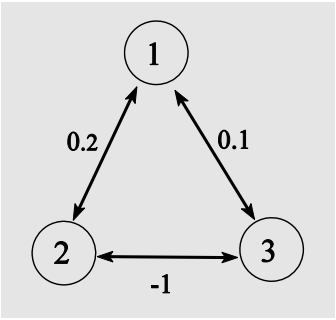
$\min_E(s) = \{t: t \geq s$ und es gibt kein $t' \geq s$, so daß $E(t') < E(t)\}$

Nehmen wir nun an, daß E die Ljapunov-Funktion (2) des Systems ist und betrachten wir stochastisches *Updating*, das ein "simulated annealing" realisiert. Dann wird, auch im Falle von Klammerung, der Zustand $\lim_{n \to \infty} \underline{f^n}(s)$ immer ein *globales* Minimum der Funktion E darstellen. Wir gelangen damit zu folgender

Beobachtung 2
Asymptotische *Updates* und E-minimale Präzisierungen eines Zustands s stimmen überein, d.h.:
$ASUP_w(s) = \min_E(s)$

3.3 Elementares Beispiel

$$w = \begin{pmatrix} 0 & 0.2 & 0.1 \\ 0.2 & 0 & -1 \\ 0.1 & -1 & 0 \end{pmatrix}$$



Ich betrachte den Anfangszustand <1 0 0> und seine möglichen Präzisierungen:

$\{t: t \geq <1\ 0\ 0>\}$	=		E
		{ <1 0 0>	0
		<1 0 1>	-0.1
		<1 0-1>	0.1
		<1 1 0>	-0.2
		<1-1 0>	0.2
Daraus ergibt sich:		<1 1 1>	0.7
$\min_E(<1\ 0\ 0>) = \{<1\ 1\text{-}1>\}$		<1 1-1>	-1.1 ☜
		<1-1 1>	-0.9
		<1-1-1> }	1.3

3.4 Aktivierungsausbreitung als nichtmonotone Inferenz

Unsere Auffassung von Informationszuständen als propositionalen Objekten ermöglicht es, die Präzisierungsrelation $\geq$ unmittelbar als einen strikten Inferenzbegriff zu deuten. Jedenfalls erfüllt die Relation $\geq$ die tarskischen Forderungen für einen (strikten) Inferenzbegriff: Reflexivität, Schnitt und Monotonie.

Balkenius & Gaerdenfors (1991) haben verdeutlicht, daß es möglich ist, die asymptotische Aktivierungsausbreitung in neuronalen Netzen als nichtmonotone (schwache) Inferenz zu beschreiben. Sei also $<S \cup \perp, \geq>$ ein erweitertes Poset von Informationszuständen für ein System mit n Einheiten. Sei w die Verbindungsmatrix und E die Energiefunktion. Dann läßt sich eine inferentielle Relation $\mid\sim$ zwischen Informationszuständen definieren, und zwar auf zweierlei Weise:

Definition 5 SCHWACHER INFERENZBEGRIFF (SIB)
(i) $s \mid\sim_w t$ gdw. $s' \geq t$ für jedes $s' \in ASUP_w(s)$ (SIB basierend auf asympt. Updates)
(ii) $s \mid\sim_E t$ gdw. $s' \geq t$ für jedes $s' \in \min_E(s)$ (SIB basierend auf E-minimalen Präzis.)

Als unmittelbare Folgerung von Beobachtung 2 ergibt sich, daß die beiden Varianten den gleichen Inferenzbegriff definieren; es gilt also: $s \mid\sim_w t$ gdw. $s \mid\sim_E t$. Außerdem ist es nicht schwer, die folgenden Beobachtungen zu beweisen:

Beobachtungen 3

(a)	$s \geq t$, dann $s \mid\sim_w t$	(SUPRAKLASSISCH)
(b)	$s \mid\sim_w s$	(REFLEXIVITÄT)
(c)	wenn $s \mid\sim_w t$ und $s \circ t \mid\sim_w u$, dann $s \mid\sim_w u$	(SCHNITT)
(d)	wenn $s \mid\sim_w t$ und $s \mid\sim_w u$, dann $s \circ t \mid\sim_w u$	(SCHWACHE MONOTONIE)

Dabei besagt (a), daß alle strikten Schlußfolgerungen auch als nichtmonotone Schlußfolgerungen durchgehen. Die Bedingungen (b-d) stellen auf den Punkt die generellen Forderungen dar, die Gabbay, Makinson, Gärdenfors, Kraus, Lehmann, Magidor (und viele andere) als konstitutiv für nichtmonotone Inferenzsysteme ansehen (*kumulative* Logiken).

Zur Illustration des schwachen Folgerungsbegriffs betrachte ich das in Abschnitt 4.1 gegebene Beispiel (mit der dort definierten Verbindunsgmatrix w).

$<1\ 0\ 0> \mid\sim_w <0\ 1\ 0>$	wegen $\min_E[<1\ 0\ 0>] = \{<1\ 1\ -1>\}$
$<1\ 0\ 0> \mid\sim_w <0\ 0\ -1>$	"
$<1\ 0\ 0> \circ <0\ 1\ 0> \mid\sim_w <0\ 0\ -1>$	wegen SCHWACHER MONOTONIE
$<1\ 0\ 0> \circ <0\ 0\ 1> \mid\sim_w <0\ -1\ 0>$	wegen $\min_E[<1\ 0\ 1>] = \{<1\ -1\ 1>\}$

4 Kostenbasierte Poole-Systeme

In konnektionistischen Systemen ist Wissen kodiert in Form der Verbindungsmatrix oder der Energiefunktion. In symbolischen System bedient man sich üblicherweise (Default-) logisch basierter Datenbasen zur Wissensrepräsentation. Ein prominentes Beispiel für derartige Systeme bilden die von Poole (z.B. Poole 1988, 1994) ausgearbeiteten Systeme. Ich führe zunächst eine neue Variante derartiger Systeme ein, die ich *gewichtete Poole-Systeme* nenne. Außerdem entwickle ich eine Präferenzsemantik, die sich in Abschnitt 5 als das entscheidende Bindeglied erweisen wird zur Darstellung des Zusammenhangs zwischen diesen Systemen und Hopfield-Netzen. Hier und im folgenden betrachte ich die Sprache L_{At} der elementaren Aussagenlogik über einem Alphabet $At = \{p_1, ..., p_N\}$ (von Elementarsymbolen).

4.1 Grundbegriffe (vgl. Poole 1988, 1994, Brewka 1991)

Definition 6

Ein Tripel $<At, \Delta, g>$ heißt ein gewichtetes Poole-System gdw.

(i) At ist eine nichtleere Menge (atomare Symbole)

(ii) Δ ist eine Menge konsistenter Formeln der Sprache L_{At} (mögliche Hypothesen)

(iii) $g: \Delta \mapsto [0,1]$ (Gewichtsfunktion)

Sei nun $T = <At, \Delta, g>$ ein gewichtetes Poole-System und sei α eine konsistente Formel (zur Repräsentation von Fakten). Dann führe ich folgende Begriffe ein:

Definition 7

(A) *Ein Szenario von α in T* ist eine Teilmenge Δ' von Δ, wobei $\Delta' \cup \{\alpha\}$ konsistent ist.

(B) *Das Gewicht eines Szenarios* Δ' ist

$$G(\Delta') = \Sigma_{\alpha' \in \Delta'} g(\alpha') - \Sigma_{\alpha' \in (\Delta - \Delta')} g(\alpha')$$

(C) *Ein maximales Szenario von α in T* ist ein ist ein Szenario, dessen Gewicht von keinem anderen Szenario (von α in T) übertroffen wird.

Damit sind wir vorbereitet, die folgende (skeptische) kumulative Ableitungsrelation zu definieren:

Definition 8
$\alpha \mid\!\sim_T \beta$ gdw. β eine (strikte) Schlußfolgerung jedes maximalen Szenarios von α in T ist.

4.2 Ein elementares Beispiel

$At = \{p_1, p_2, p_3\}$
$\Delta = \{p_1 \leftarrow_{0.2} p_2,\ p_1 \leftarrow_{0.1} p_3,\ p_2 \leftarrow_{1.0} \sim p_3\}$ (die Gewichte sind als Indizes notiert)

einige (relevante) Szenarios von p_1:

	G
{}	-1.3
$\{p_1 \leftarrow p_2\}$	-0.9
$\{p_1 \leftarrow p_2,\ p_1 \leftarrow p_3\}$	-0.7
$\{p_1 \leftarrow p_2,\ p_2 \leftarrow \sim p_3\}$	1.1 ☜
$\{p_1 \leftarrow p_3,\ p_2 \leftarrow \sim p_3\}$	0.9

Damit gilt: $p_1 \mid\!\sim_T p_2$, $p_1 \mid\!\sim_T \neg p_3$

4.3 Die Semantik gewichteter Poole-Systeme

Sei $T = \langle At, \Delta, g\rangle$ ein gewichtetes Poole-System mit $At = \{p_1, ..., p_N\}$. Eine (totale) Interpretationsfunktion für L_{At} ist eine Funktion ν von At in {-1, 1}. Für die Bewertung $[\![\alpha]\!]_\nu$ einer Formel α von L_{At} relativ zur Interpretation ν gelten die üblichen Klauseln:
$[\![\alpha \wedge \beta]\!]_\nu = \min([\![\alpha]\!]_\nu, [\![\beta]\!]_\nu)$, $[\![\alpha \vee \beta]\!]_\nu = \max([\![\alpha]\!]_\nu, [\![\beta]\!]_\nu)$, $[\![\sim\alpha]\!]_\nu = -[\![\alpha]\!]_\nu$. Für Interpretationsfunktionen ν läßt sich ein Maß angeben, daß ausdrückt wie stark die gewählte Interpretation mit dem Hypothesensystem Δ in Konflikt gerät:

(3) $\mathscr{E}(\nu) = -\Sigma_{\alpha \in \Delta}\, g(\alpha) \cdot [\![\alpha]\!]_\nu$

Aus Gründen, die im nächsten Abschnitt offensichtlich werden, nenne ich dieses Maß die *Energie* der Interpretation. Sei nun α eine (konsistente) Formel von L_{At}. Die Begriffe *Modell* und *präferentes Modell* sind wie folgt definiert:

Definition 9
(A) Eine Interpretation ν heißt ein *Modell* von α gdw. $[\![\alpha]\!]_\nu = 1$.
(B) Ein *präferentes Modell* von α ist ein Modell, dessen Energie von keinem anderen Modell von α unterboten wird.

Damit sind alle Bausteine zusammengetragen, die erforderlich sind, um einen semantischen (kumulativen) Folgerungsbegriff zu definieren:

Definition 10
$\alpha \mid\!\sim_T \beta$ gdw. jedes präferente Modell von α ein Modell von β ist.

Es läßt sich zeigen, daß dieser Folgerungsbegriff eine korrekte und vollständige Charakterisierung der in Definition 8 gefaßten kumulative Ableitungsrelation liefert:

Beobachtung 4
Für beliebige Formeln α und β von L_{At} gilt: $\alpha \mid\!\sim_T \beta$ gdw. $\alpha \mid\!\sim\!\!\!=_T \beta$.

Der Beweis folgt der von Poole (1994) vermittelten Grundidee (für Einzelheiten vgl. http://www2.rz.hu-berlin.de/asg/blutner/psylogic.ps).

5. Integration kostenbasierter Poole-Systeme und Hopfield-Netze

Um symbolische und subsymbolische Verarbeitung als unterschiedliche *Perspektiven* einunddesselben (kognitiven) Vorgangs zu deuten, ist es zunächst erforderlich, Symbolstrukturen und Aktivierungszustände aufeinander zu beziehen. In Abschnitt 5.1 wird dies mit Hilfe von Standardtechniken der algebraischen Semantik demonstriert. Abschnitt 5.2 dehnt den Bezug auf die beiden Inferenzbegriffe aus.

5.1 Symbolische Repräsentation von Informationszuständen

Ich betrachte ein Hopfield-Netz mit n Einheiten und benutze die aussagenlogische Sprache L_{At}, um (einige/alle) Zustände des Netzes zu *repräsentieren*. Mit anderen Worten: Ich sehe die Sprache L_{At} als ein *symbolisches Mittel* an, um über die Aktivierungszustände des Systems zu sprechen. Üblichen Vorstellungen der algebraischen Semantik folgend, läßt sich diese Idee wie folgt ausdrücken: Die nicht-logischen Symbole der Sprache werden durch die Elemente der entsprechenden algebraischen Struktur (deMorganscher Verband) interpretiert–das sind in unserem Falle die Aktivierungszustände; die logischen Symbole der Sprache entsprechen gewissen Operationen in der Struktur ($\wedge \mapsto \circ$, $\vee \mapsto \oplus$; die innere Negation ~ entspricht der Operation $(-s)_i = -s_i$, sie konvertiert positive in negative Information und *vice versa*; die Komplementoperation * –sie reflektiert das *Fehlen* von Information– hat bisher noch keine Entsprechung im logischen Vokabular von L_{At} gefunden; führe also ein (äußeres) Negationssymbols $\neg$ ein). Die auf diese Weise konstituierten algebraischen Modelle der Sprache L_{At} nenne ich Hopfield-Modelle:

Definition 11
Sei $\langle S \cup \perp, \geq \rangle$ das erweiterte Poset von Informationszuständen für ein System mit n Einheiten.

(i) $\langle S \cup \perp, \geq, \uparrow\downarrow \rangle$ ist ein *Hopfield-Modell* (für L_{At}) gdw. $\uparrow\downarrow$ eine Funktion ist, die jedem atomaren Symbol ein bestimmtes Element von $S \cup \perp$ zuordnet und darüber hinaus die folgenden Bedingungen erfüllt:
$\uparrow \alpha \wedge \beta \downarrow = \uparrow \alpha \downarrow \circ \uparrow \beta \downarrow$, $\quad \uparrow \alpha \vee \beta \downarrow = \uparrow \alpha \downarrow \oplus \uparrow \beta \downarrow$
$\uparrow \neg \alpha \downarrow = \uparrow \alpha \downarrow^*$, $\quad \uparrow \sim \alpha \downarrow = -\uparrow \alpha \downarrow$

(ii) $\langle S \cup \perp, \geq, \uparrow\downarrow \rangle$ ist ein lokalistische Hopfield-Modell (für L_{At}) gdw. $\langle S \cup \perp, \geq, \uparrow\downarrow \rangle$ ein Hopfield-Modell ist und $\uparrow\downarrow$ folgende Zuordnungen realisiert:
$\uparrow p_1 \downarrow = \langle 1\ 0\ \ldots\ 0 \rangle$, $\uparrow p_2 \downarrow = \langle 0\ 1\ \ldots\ 0 \rangle$, ... , $\uparrow p_n \downarrow = \langle 0\ 0\ \ldots\ 1 \rangle$.

(lokalistische Modelle realisieren eine transparente Zuordnung zwischen den Einheiten des Netzes und den atomaren Symbolen der Repräsentationssprache).

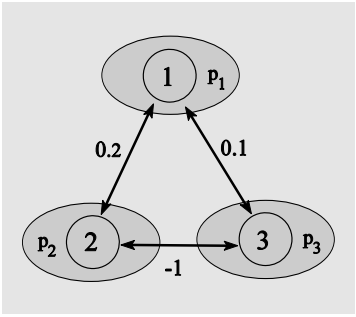
Ich will sagen, ein Informationszustand s wird durch eine Formel α vom L_{At} (relativ zu M) *repräsentiert*, falls $\uparrow \alpha \downarrow = s$ ist.

Als elementares Beispiel betrachte ich ein lokalistisches Modell für das folgende Netz:

$\uparrow p_1 \downarrow = <1\ 0\ 0>$

$\uparrow p_2 \downarrow = <0\ 1\ 0>$

$\uparrow p_3 \downarrow = <0\ 0\ 1>$



Es ist unmittelbar einsichtig, daß gilt:

p_1	repräsentiert <1 0 0>	p_2	repräsentiert <0 1 0>
p_3	repräsentiert <0 0 1>	$p_1 \wedge p_2$	repräsentiert <1 1 0>
$\sim p_1$	repräsentiert <-1 0 0>	$p_1 \wedge p_2 \wedge \sim p_3$	repräsentiert <1 1 -1>

Ich nenne einen Zustand $s \in S$ *symbolisch* (relativ zu einem Hopfield-Modell M) gdw. der Zustand s durch eine Formel α in L_{At} repräsentiert wird. Es ist leicht zu sehen, daß bezogen auf ein lokalistisches Modell jeder Zustand symbolisch ist.

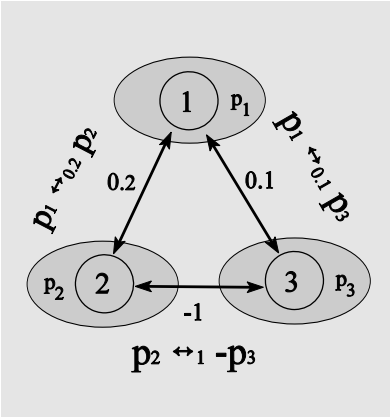
5.2 Übersetzung von Hopfield-Netzen in gewichtete Poole-Systeme

Für die in Abschnitt 3 untersuchten Hopfield-Netze gilt, daß für jeden (partiellen) Informationszustand u eine totale Präzisierung $t \geq u$ existiert mit $E(t) \leq E(u)$. Daraus folgt, daß die Klasse der asymptotischen *Updates* eines Aktivierungszustands s, $ASUP_w(s)$, mit jedem partiellen Informationszustand u auch eine seiner totalen Präzisierungen t enthält. Unter speziellen Bedingungen an die Netzarchitektur (keine "isolierten" Knoten etc.) läßt sich zeigen, daß $ASUP_w(s)$ nur totale Informationszustände enthält. Das besagt, daß sich unter diesen Bedingungen jeder Aktivierungszustand asymptotisch immer in total präzisierte Zustände entwickelt. (Das folgt daraus, daß für jeden (partiellen) Informationszustand eine totale Präzisierung mit *geringerer* Energie existiert)

Die Beschränkung auf Systeme, deren Dynamik asymptotisch immer zu totalen Informationszuständen führt, ermöglicht es, die nichtmonotone Inferenzrelation zwischen Aktivierungszuständen in Hopfield-Netzen (Definition 5) unmittelbar auf die Ableitungsrelation zwischen Formeln der Sprache L_{At} für gewichtete Poole-Systeme zu beziehen, jedenfalls solange eine lokalistische Betrachtungsweise von Hopfield-Netzen gewählt wird. Für den allgemeinen dynamischen Fall gelingt die Korrespondenz erst dann, wenn ein partieller Modellbegriff beim Aufbau der Poole-Systeme berücksichtigt wird. Das ist jedoch eine Aufgabe, die den gegenwärtigen Rahmen der Untersuchung sprengen würde.

Sei nun ein Hopfield-System mit n Einheiten und der Verbindungsmatrix w gegeben. Weiter sei $At = \{p_1, ..., p_N\}$ die Menge der atomarer Symbole der Sprache L_{At}. Ich betrachte die folgenden Formeln von L_{At}:

$\alpha_{ij} = [p_i \leftrightarrow \text{sign}(w_{ij})\ p_j\ ]$, für $1 \le i < j \le n$



Für jede Verbindungsmatrix w kann damit das folgende gewichtete Poole-System T_w = <At, Δ_w, g_w> konstruiert werden:

$\Delta_w = \{\alpha_{ij}: 1 \le i < j \le n\}$
$g_w(\alpha_{ij}) = |w_{ij}|$

Es wird sich erweisen, daß diese "Übersetzung" von Verbindungsmatrizen in gewichtete Poole-Systeme zu einer strikten Korrespondenz zwischen den beiden Inferenzbegriffen führt, eine lokalistische Betrachtung von Hopfield-Netzen vorausgesetzt.

Zunächst ist es erforderlich, auf die wechselseitige Entsprechung zwischen totalen Informationszuständen und totalen Interpretationsfunktionen (vgl. Abschnitt 4.3) hinzuweisen. Jedem totalen Informationszustand t entspricht nämlich genau eine totale Interpretationsfunktion ν/t mit $\nu/t\ (p_i) = t_i$ (und umgekehrt). Die folgenden Beobachtungen sind dann leicht zu bestätigen:

Beobachtungen 5

(a) $[\![p_i]\!]_{\nu/t} = t_i$

(b) $[\![\sim\alpha]\!]_{\nu/t} = -[\![\alpha]\!]_{\nu/t}$

(c) $[\![\alpha \leftrightarrow \beta]\!]_{\nu/t} = [\![\alpha]\!]_{\nu/t} \cdot [\![\beta]\!]_{\nu/t}$

(d) $t \ge \alpha$ gdw. $[\![\alpha]\!]_{\nu/t} = 1$, falls die Formel α eine Konjunktion von Literalen (Atome oder deren innere Negation) ist

(e) $\mathscr{E}(\nu/t) = E(t)$ (d.h. $\Sigma_{\alpha \in \Delta}\ g(\alpha) \cdot [\![\alpha]\!]_{\nu/t} = \Sigma_{i>j}\ w_{ij} \cdot t_i \cdot t_j$)
wobei E die Energiefunktion eines Hopfield-Netzes mit der Verbindungsmatrix w ist und $\mathscr{E}$ die Energiefunktion des gewichteten Poole-Systems T_w.

Wenn die Formeln α und β Konjunktionen von Literalen sind, dann ergibt aus den Beobachtungen 5 sich daraus und aus den Definitionen 5 und 10 das folgende

Theorem
$\uparrow\alpha\downarrow\ \sim_w\ \uparrow\beta\downarrow$ gdw. $\alpha\ \mathrel{|\!\sim}_T\ \beta$ (gdw. $\alpha\ \mathrel{|\!\sim}\!\!\!\!-_T\ \beta$).

Damit ist der Zusammenhang zwischen der (asymptotischen) Aktivierungsausbreitung in Hopfield-Netzen einerseits und nichtmonotonen Schlüssen in gewichteten Poole-Systemen andererseits ausgedrückt. Dieser Zusammenhang ist aus zweierlei allgemeinen Gründen nützlich: (i) Dank diesem Zusammenhang kann eine Schwäche konnektionistischer Systeme überwunden werden, die in der weitgehenden "Undurchschaubarkeit" der Verarbeitungseffekte derartiger Systeme besteht. Die demonstrierte Übersetzungsmethode ermöglicht es einem Benutzer, die "Schlüsse" eines konnektionistischen Systems unmittelbar nachzuvollziehen. Sie werden als symbolische, nichtmonotone Schlüsse in einer kumulativen Logik faßbar. (ii) Die Explosion des Verarbeitungsaufwands in traditionellen nichtmonotonen Systemen bei großen Datenbasen läßt sich möglicherweise durch den Einsatz stochastischer Prozesse ("simulated annealing")

vermeiden (implementativer Konnektionismus).

Darüber hinaus liefert die vorgeschlagene Sichtweise theoretische Aufschlüsse darüber, welche Arten von Logiken zur Beschreibung emergenter Eigenschaften neuronaler Netze dienen können und gibt derartigen Logiken eine besondere *theoretische Rechtfertigung*.

6. Potentielle Anwendungen für die Verarbeitung natürlicher Sprachen

Geht es um die Belange der natürlichen Sprachverarbeitung, dann scheint eine unmittelbare Anwendung der hier dargestellten Überlegungen zur Integration symbolischer und konnektionistischer Ansätze auf den ersten Blick nahezu völlig ausgeschlossen. Dazu sind die gegenwärtigen Beschränkungen zu drastisch und die Vereinfachungen zu grob. Das betrifft insbesondere die vollständige Ausklammerung der Bindungsproblematik, die Beschränkung auf eine lokalistische Betrachtung von Hopfield-Netzen und die weitgehende Ignoranz gegenüber der asymptotischen Einstellung von *partiellen* Informationszuständen. Wenn die vorliegende Studie von einem gewissen Wert ist, dann besteht dieser offenbar in erster Linie darin, eine Methode zu exemplifizieren, die darauf abzielt, eine komplexe Wirklichkeit durch die Betrachtung aus unterschiedlichen Perspektiven besser zu verstehen. Die Herstellung eines systematischen Bezugs zwischen den unterschiedlichen Perspektiven bildet den Kern eines derartigen Forschungsprogramms und unterscheidet es beispielsweise von einem hybriden Ansatz (z.B. Wermter & Lehnert 1989; für eine markante Kritik vgl. Cottrell 1989). In unseren Falle führt die integrative Methode auf interessante Wechselbeziehungen zwischen konnektionistischer Theorienbildung und einigen neueren Entwicklungen innerhalb der modelltheoretischen Semantik (Präferenzsemantik, Defaultsemantik, Datensemantik, *Update*-Semantik etc.).

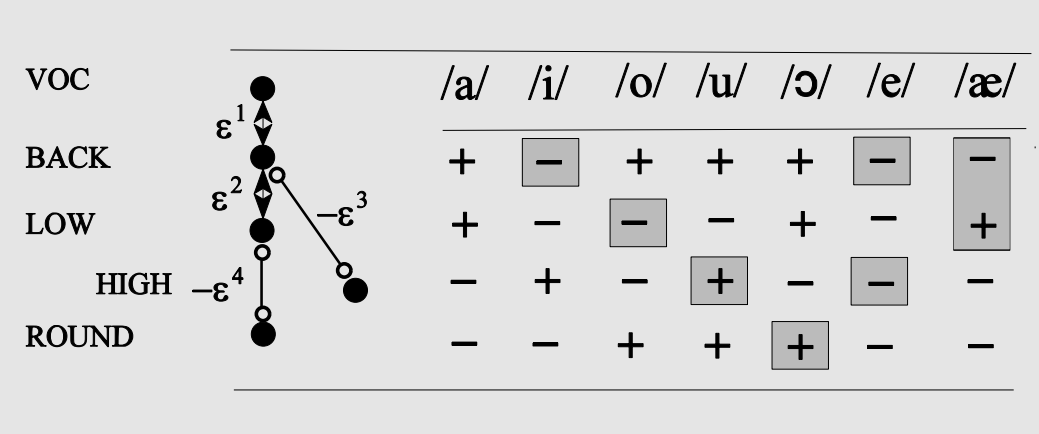
Dennoch erscheint es angemessen, auf einen Aspekt der *Semantik* und *Pragmatik* natürlicher Sprachen etwas genauer einzugehen, der in jüngster Zeit eine erhebliche Rolle in der natürlichen Sprachverarbeitung spielt. Ich meine den Aspekt der (semantischen) Unterspezifizierung und die Realisierung von Mechanismen zur "plausiblen" Vervollständigung unterspezifizierter semantischer Repräsentationen (vgl. van Deemter & Peters 1996; van der Sandt, Blutner & Bierwisch 1997).

Es ist zweckmäßig, die Idee der Unterspezifizierung, der wissenschaftsgeschichtlichen Entwicklung folgend, zunächst kurz an einem elementaren Beispiel aus der (intrasegmentalen) Phonologie zu erläutern. Ich betrachte das nebenstehende Segmentsystem und will eine möglichst ökonomische Beschreibung dieses Fragments durch die Merkmale BACK, LOW, HIGH, ROUND geben. Diese Merkmale können als die Atome einer elementaren Sprache L_{At} aufgefaßt werden (wobei folgende strikten Restriktionen vorausgesetzt werden:
LOW → ~HIGH; ROUND → BACK).

−back	+back	
/i/	/u/	+high
/e/	/o/	−high/−low
/æ/	/ɔ/ /a/	+low

Das generische Wissen eines phonologischen Agenten hinsichtlich dieses Fragments kann durch ein Hopfield-Netz repräsentiert werden. Dabei werden exponentielle Gewichte mit der Basis $0 < \varepsilon \leq 0.5$ benutzt werden (eine Annahme, die auch die Optimalitätstheorie voraussetzt). Die folgende Grafik zeigt, daß nur ein geringer Teil der Merkmalsausprägungen explizit gegeben sein

muß (im Bild dunkel markiert), um eine vollständige Spezifizierung zu erreichen.



Der gleiche Effekte läßt sich durch die Zuordnung des folgenden Poole-Systems erreichen:

VOC $\leftrightarrow_{\varepsilon^1}$ BACK; BACK $\leftrightarrow_{\varepsilon^2}$ LOW
LOW $\leftrightarrow_{\varepsilon^4}$ ~ROUND; BACK $\leftrightarrow_{\varepsilon^3}$ ~HIGH

Man beachte, daß diese Defaults unmittelbar als Ausdrücke der Markiertheitstheorie von Keane (1975) verstanden werden können. Die Auffüllung unterspezifizierter Merkmalsmatrizen kann man so aus zwei Perspektiven sehen: der gewohnten markiertheits-theoretischen Perspektive und einer netzlinguistischen Perspektive. Beide Perspektiven haben offensichtlich ihre Vorzüge.

Motiviert durch das *Puzzle der kombinatorischen Explosion* haben gegenwärtige Untersuchungen zur semantischen und konzeptuellen Interpretation die Idee der Unterspezifizierung aufgegriffen mit dem Ziel, möglichtst alle "nichtmonotonen" Operationen zu eliminieren, insofern diese zu Informationsverlusten bzw. zur Zerstörung einmal erzeugter semantischer Repräsentationen führen (z.B. van Deemter & Peters 1996; van der Sandt, Blutner & Bierwisch 1997). Die dabei benutzten symbolischen Techniken zur "plausiblen" Vervollständigung unterspezifizierter semantischer Repräsentationen basieren im wesentlichen entweder auf Abduktion (z.B. Hobbs et al. 1993; Blutner, Leßmöllmann, & van der Sandt 1995) oder auf bestimmten nichtmonotonen Logiken (z.B. Lascarides & Asher 1993 und verschiedene Papiere in van Deemer & Peters 1995). Es liegt in der Natur dieser symbolischen Mechanismen, daß sie wegen ihrer eigenen Komplexitätsproblematik den angestrebten Effekt teilweise oder vollständig wieder beseitigen. Die Einbeziehung konnektionistisch motivierter Verarbeitungsmechanismen (auch im Sinne des implementativen Konnektionismus, z.B. Derthick 1990) kann hier nützlich sein. Für die Klärung der dabei zahlreich auftretenden theoretischen Fragen erscheint ein integrativer Ansatz besonders vielversprechend.

Literatur

Balkenius, C. & Gärdenfors, P. (1991): "Nonmonotonic inferences in neural networks". In J.A. Allen, R. Fikes, & E. Sandewall (eds), *Principles of knowledge representation and reasoning*. San Mateo, CA: Morgan Kaufmann.

Blutner, R., Leßmöllmann, A., & van der Sandt, R. (1996): "Conversational implicature and lexical pragmatics". In: *Proceedings of the AAAI Spring Symposium on Conversational Implicature*. Stanford, 1-9.

Brewka, G. (1991): *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambidge University Press.

Cohen, M.A. & Grossberg, S. (1983): "Absolute stability of global pattern formation and parallel memory starage by competive neural networks". *IEEE Transactions on Systems, Man , and Cybernetics*, SMC-13, 815-826.

Cottrell, G.W. (1989): "Research Note: A hybrid model of the intentional behavior of the dog". *Connection Science*, 3, 341-342.

Derthick, M. (1990): "Mundane reasoning by setting on a plausible model". *Artificial Intelligence*, 46, 107-157.

Hobbs

Hopfield, J.J. (1982): "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences* 79, 2554-2558.

Kean, M.L. (1975): *The theory of markedness in generative grammar*. Ph.D. thesis, MIT, Cambridge, Mass.

Lascarides, A. & Asher, N. (1993): "Temporal interpretation, discourse relation, and common sense entailment". *Linguistics and Philosophy*, 16, 437-494.

Poole, D. (1988): "A logical framework for default reasoning". *Artificial Intelligence*, 36, 27-47.

Poole, D. (1996): "Who chooses the assumptions?". In P. O'Rorke (Ed.), *Abductive Reasoning*. Cambridge: MIT Press.

Shoham Y. (1986): *Reasoning about change: Time and causation from the standpoint of artificial intelligence.* Ph. D. Thesis, Yale University.

Smolensky, P. Legendre, G. , & Miyata, Y (1992): *Principles for an integrated connectionist/symbolic theory of higher order cognition.* Unpublished Paper, University of Colorado.

van Deemter, K. & Peters, S. (1996): *Semantic Ambiguity and Underspecification*. Stanford, California: CSLI Publications.

van der Sandt, R., Blutner, R. & Bierwisch, M. (1997): *From underspecification to interpretation.*

Wermter, S. & Lehnert, W.G. (1989): "A hybrid symbolic/connectionist model for noun phrase understanding". *Connection Science*, 3, 255-272.