

## TAGUNGSBERICHT ZUM 1. SHOE-WORKSHOP (EXTRACTION OF HIERACHICAL STRUCTURE FOR MACHINE LEARNING OF NATURAL LANGUAGE)

27-28.2.92, ITK, Tilburg

Ziel des Workshops war es, einen Kooperationsrahmen für die in diesem Bereich tätigen Wissenschaftlerinnen und Wissenschaftler zu schaffen und einen Überblick über den aktuellen Forschungsstand zu vermitteln. Der Teilnehmerkreis (23 Personen) stammte aus Belgien, der Bundesrepublik, Großbritannien und den Niederlanden.

*A. van den Bosch* und *W. Daelemans* (ITK, Tilburg) untersuchten die Fähigkeit neuronaler Netze (Elman- und Jordan-Netze), Silbengrenzen in orthographischen und phonologischen Repräsentationen niederländischer Wortformen zu bestimmen. Getestet wurden sowohl unterschiedliche Codierungen der Eingaben als auch verschiedene Netzwerk-Architekturen. Zusammenfassend konnte festgestellt werden, daß die verwendeten Netze die Aufgabe zwar bewältigen konnten, jedoch nicht signifikant besser als symbolische Ansätze.

*G. Durieux* (UIA, Antwerpen) entwickelte ein Analogie-basiertes Verfahren, um einfachen Holländischen Wortformen (Monomorphemen bzw. Lehnwörter, die sich wie solche verhalten) den Hauptakzent zuzuweisen. Es basiert auf der Identifikation von Ähnlichkeiten bzw. Abweichungen zwischen einer Wortform und anderen im Lexikon gespeicherten Wortformen. Vorteile gegenüber traditionellen regelbasierten Ansätzen bietet dieses Verfahren

vor allem aufgrund seiner Fehlertoleranz und der der Fähigkeit partielle Übereinstimmungen zu nutzen.

*T.M. Ellison* (CCS, Edinburgh) beschrieb ein unüberwachtes Lernverfahren für Vokal-Harmonien. Zur Repräsentation wurden nicht-deterministische endliche Automaten mit zwei Zuständen verwendet, die nicht nur einfache Harmonien sondern auch Transparenz und Opazität darzustellen erlauben. Das vorgestellte System wurde erfolgreich für die Sprachen Türkisch, Kirgisisch, Ungarisch und Yoruba getestet. Es zeigte sich, daß das Verfahren auch ohne sprach-spezifisches Wissens korrekte Vokal-Harmonien lernen konnte.

*S. Finch* (CCS, Edinburgh) beschrieb einen statistischen Ansatz zum Lernen lexikalischer Kategorien auf der Basis ungetaggtter Korpora. Dazu wird jede Wortform durch einen Vektor beschrieben, der die Distribution der lokalen Kontexte beschreibt, in denen die betreffende Wortform vorkommt. Mit Hilfe einer Clusteranalyse werden diese dann zu Gruppen zusammengefaßt. Erfolgreich getestet wurde das Verfahren anhand größerer englischer Korpora.

*P. Flach* (ITK, Tilburg) gab eine Einführung in die Grundlagen des Konzeptlernens und der Induktiven Logikprogrammierung. Er beschränkte sich auf induktive Verfahren des Konzeptlernens; d.h. auf Verfahren, die auf der Grundlage positi-

ver und negativer Beispiele Klassifikationsregeln bilden und stellte das *Version Space* Modell als ein allgemeines Modell für das Konzeptlernen vor. Ziel der Induktiven Logikprogrammierung ist es Verfahren zu entwickeln, die den beispielgesteuerten Erwerb von Logikprogrammen ermöglichen.

*S. Gillis* (UvA, Antwerpen) skizzierte die verschiedenen Positionen, die in der Psycholinguistik zu folgendem, für die erste Phase des Spracherwerbs zentralen Fragen eingenommen werden:

- . Wie ist die beim Spracherwerb von Kindern gezeigte Kreativität adäquat zu charakterisieren?
- . Gibt es sprachspezifische Prädispositionen oder wird Spracherwerb durch allgemeine kognitive Mechanismen gesteuert?
- . Sind die biologische Grundlage des Spracherwerbs artspezifisch ?
- . Ist es korrekt Spracherwerb als einen in seiner Form und zeitlichen Realisierung universalen Prozeß zu beschreiben?

*B. Manderick* (AI-LAB, Brüssel) berichtete über die vielfältigen Anwendungsmöglichkeiten der durch adaptive biologische Systeme inspirierten Genetischen Algorithmen(GA). Mit Hilfe der beiden "genetischen Operatoren" der Mutation und Selektion bieten GAen häufig eine effiziente Lösung für viele Such- und Optimierungsprobleme (wie z.B. das Rundreiseproblem). Darüberhinaus wurden GAen in der Bildverarbeitung, der Mustererkennung, dem maschinellen Lernen sowie in den Wirtschafts- und Sozialwissenschaften erfolgreich eingesetzt.

*S. Naumann* und *J. Schrepp* (LDV /CL, Trier) beschrieben ein System zum Erwerb syntaktischen Wissens. Den Kern des Systems bildet ein inkrementeller Lernalgorithmus,

der ausgehend von einem Satzkorpus eine Folge reversibler Syntaxen generiert. In jedem Schritt wird aus dem Korpus eine kleine Menge von Sätzen ausgewählt und von einem speziellen Parser analysiert, der den von der aktuellen Grammatik nicht erfaßten Sätzen partielle Strukturbeschreibungen zuordnet. Auf der Grundlage dieser Strukturbeschreibungen werden Hypothesen zur Erweiterung der Syntax formuliert, die schließlich zur Generierung der neuen Syntax führen.

*D.M. W. Powers* (ITK, Tilburg) berichtete von einer Reihe von Experimenten zum Erwerb lexikalischen und syntaktischen Wissens. Ziel dieser Experimente war das Erlernen von Wortklassen und syntaktischen Regeln. Verwendet wurden in allen Fällen statistische bzw. konnektionistische Verfahren. Bei den Wortklassenexperimenten zeigte sich, daß geschlossene Wortklassen vor den offenen Klassen gelernt wurden. Zu diesem Ergebnis wurde auf von einem weiteren Experiment (Klassifikation von Graphemen/Phonemen) bestätigt.

*J. C. Scholtes* (UvA, Amsterdam) stellte einen Ansatz zum daten-orientierten Parsen mit Hilfe eines Kohonen-Netzes (feature map) vor. Der Hauptunterschied zu anderen korpus-basierten Verfahren ist die Verwendung von statistischer als auch syntaktischer und semantischer Information während der Trainingsphase. Ein so angereichertes Korpus wird in einem Kohonen-Netz "gespeichert", das dann zum Parsen verwendet wird.

*J. Schrepp* (LDV /CL, Trier) gab einen kurzen Überblick über Ziele und Methoden der quantitativen und synergetischen Linguistik. Besonders hervorgehoben wurden die oft vernachlässigten Anwendungsmöglichkeiten statistischer Information in praktischen Systemen.

S.N. & J.S.