Evaluierungsworkshop*

John Nerbonne

Ein Workshop zur Evaluierung in natürlichsprachlichen Systemen und in der Spracherkennung wurde vom 30. April bis 2. Mai 1992 in Edinburgh abgehalten. Bemerkenswert daran war nicht nur die Tatsache, daß viele Gruppen die Wichtigkeit von Evaluierung erkannt haben, sondern auch daß inzwischen verschiedene Aktivitäten in dieser Richtung unternommen werden. Im allgemeinen wächst die Erkenntnis, daß die systematische Evaluierung sowohl auf wissenschaftlicher Ebene als auch anwendungsbezogen der Sprachtechnologie Vorschub leisten kann.

Wichtige Daten

Titel:
The Strategic Role of Evaluation in Natural Language Processing and Speech Technology

Leitung:
Henry Thompson, Edinburgh

Träger:
ESPRIT Working Group on Dialogue and Discourse DANDI European Network of Excellence in Language and Speech ELKSNET Human Communication Research Centre, University of Edinburgh

Teilnehmer:
Madeleine Bates, BBN, Cambridge
Robert Dale, University of Edinburgh
Björn Gambäck, Swedisches Institut für Informatik
Dafydd Gibbon, Universität Bielefeld
James Hieronymus, University of Edinburgh
Lee Humphreys, University of Essex
Stephen Isard, University of Edinburgh
Karen Sparck Jones, University of Cambridge
Maghi King, ISSCO, Genf
Ewan Klein, University of Edinburgh
Marc Moens, University of Edinburgh
John Nerbonne, DFKI, Saarbrücken
Klaus Netter, DFKI, Saarbrücken

Das Protokoll liegt als Publikation des Human Communication Research Centres vor: Record of DANDI/HCRC/ELSNET Evaluation Workshop, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW.

Kommentar

Der Workshop hatte zum Ziel, Experten aus verschiedenen Sprachverarbeitungsbereichen (Spracherkennung, Sprachverstehen, Übersetzung, Generierung, Sprachsynthese, Information Retrieval, und Message Understanding), die an Evaluierungsproblemen gearbeitet hatten, zusammenzubringen, um den Austausch von Methoden, Erfahrungen, und Einsichten ihrer Arbeiten zu ermöglichen. Der Workshop setzte die Arbeit aus dem früheren DARPA Workshop (Palmer and Finin 1990) und dem 1991 ACL Workshop in Berkeley (Neal and Walter 1991) fort.

Eines der wichtigsten Ergebnisse der Vorträge und der Diskussion war eine nicht nur terminologische Unterscheidung, deren Fehlen in der Praxis oftmals zu Mißverständnissen führt:

Diagnostik dient zur Feststellung von Fehlern und der Fragmentabdeckung eines Moduls, **Assessment** klärt, wie gut ein Modul seine Aufgabe erfüllt, und **Evaluierung** bewertet, inwiefern ein Modul für eine vorgesehene Anwendung tauglich ist

Übereinstimmung herrschte darin, daß diese verschiedenen Aspekte sauber getrennt werden sollten und daß sehr häufig als Evaluierung betrachtet wird, was im Grunde eher als Assessment oder Diagnostik aufzufassen wäre – wobei der Wert der anderen Aktivitäten nicht geringer geschätzt wird. Der Sprachingenieur braucht Diagnostikwerkzeuge, der Systementwickler Assessment und der Kunde Evaluierung. Um ein paralleles Beispiel aus einem nicht-linguistischen Bereich zu nennen: Man verwendet Dichtemessungsgeräte, um Autobatteriefehler zu diagnostizieren, aber auch Voltmeter, um Assessment vorzunehmen, und schließlich Ausfallhäufigkeitsstatistiken aus der Verbraucherzentrale, um das Verhalten im strapazierenden Motorraum zu evaluieren. Aus dieser Perspektive ist klar, daß „User-Satisfaction" in erster Linie Sache der Evaluierung ist, die keineswegs alles andere ersetzen kann. Andererseits kann eine reine Fehlerdiagnostik anhand eines Test-Korpus mitunter nur wenig Aufschluß über die Anwendungstauglichkeit eines Moduls oder gar Systems liefern.

Es ist offensichtlich, daß verschiedene Anwendungsbereiche (man denke hier an Sprachsynthese verglichen mit Informationretrieval) auch ganz unterschiedliche Methoden zur Evaluierung verlangen, gleichzeitig können sie sich aber gleicher beziehungsweise ähnlicher Assessment- und Diagnostikarbeiten in

* Diese Arbeit wird durch Forschungsauftrag ITW 9002 0 des BMFT an das DFKI DISCO Projekt unterstützt.

dem Maße bedienen, in dem sie gleiche beziehungsweise ähnliche Module einsetzen. Es spricht zum Beispiel nichts dagegen, daß die morphologische Analyse von den zwei Systemen verglichen werden (Assessment), beziehungsweise daß gleiche Diagnostikwerkzeuge verwendet werden.

Eine Arbeit zum Deutschen

Als einzige Arbeit speziell zum Deutschen dürfte der Vortrag Klaus Netters über DiTo, ein Diagnostikwerkzeug für deutsche Syntax, von Interesse sein (vgl. auch Klein et al. 1992). DiTo will die Fehlerdiagnose bei Syntaxkomponenten natürlichsprachlicher Systeme unterstützen. Die Fehlerdiagnose hilft bei Debugging, bei dem internen Einschätzen des Fortschritts, bei der Entwicklung und bei der Systemwartung, wo ein systematisches Monitoring eingesetzt werden kann, um unvorhergesehene Rückfälle bei neuen Versionen abzufangen. Zentraler Bestandteil von DiTo ist ein systematisch konstruierter Datenkatalog, der langfristig die wesentlichen Bereiche deutscher Syntax anhand von Beispieldaten abdecken soll, wobei bisher unter anderem die Bereiche Verbrektion und Satzkoordination behandelt werden. Das Korpus enthält sowohl wohlgeformte als auch nicht-wohlgeformte Wortketten, wobei eine grob-gerasterte Konstitutentenstruktur (Vorkommen von NP, PP) bei den wohlformierten Ausdrücken annotiert ist.

Die Beispieldaten und die syntaktischen Annotationen sind in einer relationalen Datenbank organisiert, um den Zugriff auf die Daten zu erleichtern, die Einträge konsistent zu halten und gute Voraussetzungen für die Erweiterung um neue Syntaxbereiche bereitzustellen. Damit das Werkzeug möglichst breite Benutzung genießt, wurde es ausschließlich in Public Domain Software – awk, LEX und YACC in UNIX – implementiert.

Im Augenblick versuchen Netter und seine Kollegin, Judith Klein, andere Gruppen dafür zu gewinnen, neue Syntaxbereiche in Zusammenarbeit mit ihnen aufzuarbeiten, um als Gegenleistung die gesamte Datenbank und dazugehörige Werkzeuge (im wesentlichen Konsistenzprüfungsprogramme und andere Serviceprogramme) zu erhalten. Es besteht bereits eine Kooperation mit dem Institut für angewandte Informationswissenschaft, Saarbrücken, in deren Rahmen Daten zu Funktionsverbgefüge erarbeitet werden, ein Zusammenarbeiten mit dem Institut für Computerlinguistik an der Universität Koblenz ist vorgesehen.

Höhepunkte des Workshops

Unter vielen sehr interessanten Berichten und Diskussionen auf dem Workshop verdienten besondere Aufmerksamkeit die Beiträge zu Information Retrieval und zu sprachverstehenden Systemen (Systeme, die gesprochene Sprache verstehen), bei denen sich Evaluierung bereits als äußerst wertvoll erwiesen hat.

• Karen Sparck Jones berichtete von jahrelangen Erfahrungen (unter anderem von eigenen Entwicklungen) in Evaluierung im Bereich der Informationretrieval. Inzwischen gehören diese Methoden, die im wesentlichen auf den Prozentsatz der korrekt gefundenen Information (Recall) und Prozentsatz der gefundenen Informationen, die auch einschlägig und relevant sind (Präzision), abzielen, zur normalen Arbeit und führen zu einer besseren Bewertung der tatsächlichen Fortschritte in der Technologie. Sparck Jones warnte vor dem zu naiven Gebrauch von Evaluierungsdaten und erinnerte daran, daß man begrifflich zwischen folgenden Aspekten der Evaluierung streng trennen sollte:

Kriterium
Welche abstrakte Eigenschaft des Systems ist das Ziel der Evaluierung? Wie wichtig ist diese Eigenschaft? Um das oben erwähnte Beispiel der Syntax aufzugreifen, dürfte die PRÄZISION ein interessantes Kriterium sein: definiert eine konkrete Grammatik genau die wohlformierten Sätze (in Hinblick auf eine Anwendungsart beziehungsweise konkrete Anwendung)?

Maß
Welches konkrete Verhalten des Systems spiegelt das Kriterium wider? Gibt es Verzerrungen, so daß die abstrakte Eigenschaft nur teilweise oder lückenhaft zum Vorschein kommt? Beispiel: Prozentsatz der korrekt ermittelten Sätze aus einem vorgegebenen Korpus, beziehungsweise die sich komplementär verhaltende Fehlerquote. Falls das Korpus nicht-wohlgeformte Beispiele enthält, ist es interessant, sowohl den Prozentsatz der wohlgeformten Ausdrücke zu ermittlen, die als solche erkannt werden, als auch den Prozentsatz der nicht-wohlgeformten, die korrekt als nicht-wohlgeformt abgelehnt werden (und den sich hierzu komplementären Prozentsatz der falschen Positiva – nicht-wohlgeformte Sätze, die ein System als doch wohlgeformt akzeptiert). Hier treten auch Fragen auf, die zum Beispiel die Repräsentivität des Korpus und die Zuverlässigkeit der darin enthaltenen Daten betreffen.

Methode
Wie wird das Maß ermittelt? Welche Zusatzannahmen sind nötig, um zu rechtfertigen, daß man ein echtes Bild erhält? Beispiel: Um eine Syntaxkomponente einem Assessment zu unterziehen, ist es naheliegend, einen Parser einzusetzen, der relativ direkt (und hoffentlich zuverlässig) Systemvorhersagen hervorrufen kann. Das Resultat kann jedoch dadurch beeinträchtig werden, daß normalerweise auch eine Morphologie nötig ist, die ihrerseits fehlerträchtiger sein kann, oder dadurch, daß eine Syntax oft viele Informationen enthält, die nur der semantischen Auswertung dienen, so daß das Systemverhalten dadurch verzerrt wird.

• Beeindruckend sind auch die Ergebnisse der Evaluierung in dem DARPA SLS Projekt, wie aus den Berichten von James Hieronymus und Madeleine Bates zu erfahren war. Anfangs arbeiteten fünf verschiedene Gruppen an dem gleichen Problem (Verstehen spontan gesprochener Sprache, um Auskunft über Flugpläne zu erhalten) mit den gleichen Evaluierungsmethoden. Dabei traten solch unerwartet rapide Verbesserungen in der Erkennung ein, daß sich inzwischen auch industrielle Partner (Bell Labs, Dragon Systems) den Evaluierungszyklen unterziehen (nur um daran teilzunehmen). Manche neuen Methoden werden demnächst in Produkten eingesetzt.

Die Vortragenden führen diese Erfolge weniger auf das Zuckerbrot und Peitsche von DARPA (die sowieso für die freiwilligen Industriepartner uninteressant sind), sondern auf die Tatsache, daß die Einigung auf eine Evaluierungsmethode und -aufgabe erstmals detaillierte technische Vergleiche zwischen den Systemen ermöglicht. So werden erfolgreiche beziehungsweise erfolglose Methoden viel schneller gemeinsam anerkannt und verwendet beziehungsweise fallengelassen.

KI

References

Klein, J., K. Diagne, L. Dickmann, J. Nerbonne, and K. Netter. 1992. DiTo – Ein Diagnostikwerkzeug für deutsche Syntax. In Tagungsband KONVENS '92

Neal, J., and S. Walter (ed.). 1991. Proceedings of the ACL Workshop on the Evaluation of NLP Systems. Rome Laboratory. RL-TR-91-362.

Palmer, M., and T. Finin. 1990. Workshop on the Evaluation of Natural Language Processing Systems. Computational Linguistics 16(3):175-181.

John Nerbonne, DFKI, Stuhlsatzenhausweg 3, 6000 Saarbrücken 11, e-mail:nerbonne@dfki.uni-sb.de