Evaluation maschineller Übersetzungssysteme

Ein Projekt des Arbeitskreises „Maschinelle Übersetzung“ der GLDV

Uta Seewald

Seit Beginn des Jahres 1998 liegt der Schwerpunkt des Arbeitskreises „Maschinelle Übersetzung“ auf der Evaluierung maschineller Übersetzungssysteme. Das Evaluationsverfahren ist auf zwei Veranstaltungen des Arbeitskreises in Saarbrücken, am 30. Januar 1998 und am 8. Mai 1998, ausgearbeitet worden.

Auf der Veranstaltung am 30. Januar ging es zunächst um die Erarbeitung eines Kriterienkatalogs zur Durchführung der Evaluation. Um hierbei möglichst vielfältige Aspekte und Erfahrungen mit Evaluationsvorhaben zu berücksichtigen, berichteten die Teilnehmerinnen und Teilnehmer im ersten Teil der Veranstaltung über bereits durchgeführte Evaluationen oder einzelne Aspekte solcher Evaluationen. Der zweite Teil des Arbeitskreistreffens war schließlich der eigentlichen Erarbeitung des Kriterienkatalogs zur Evaluation von Übersetzungssystemen gewidmet. Der hier erarbeitete Kriterienkatalog sieht ein zweistufiges Evaluationsverfahren vor: die erste Stufe der Evaluation wird ausschließlich kommerziell vertriebene Übersetzungssysteme berücksichtigen. In der zweiten Stufe sollen dann in der Forschung im Einsatz bzw. in Entwicklung befindliche Systeme evaluiert werden, wobei insbesondere der Frage nachgegangen werden soll, was Forschungssysteme in ihrer Leistung gegenüber kommerziell vertriebenen Systemen auszeichnet. Um die in ihrem Aufbau, Umfang und hinsichtlich der Benutzerfreundlichkeit so unterschiedlichen Systeme, wie sie im Fall der kommerziellen und der Forschungssysteme vorliegen, vergleichen zu können, erschien es sinnvoll, das Schwergewicht der Evaluation auf linguistische Kriterien zu legen, da diese gleichermaßen an kommerziellen und an Forschungssystemen überprüft werden können.

Um sicherzustellen, daß die Evaluationsergebnisse in bezug auf die linguistische Abdeckung der Systeme tatsächlich vergleichbar sind, wurde ferner festgelegt, in der ersten Evaluationsphase ausschließlich Systeme mit dem Sprachpaar Englisch – Deutsch und zwar mit Englisch als Quell- und Deutsch als Zielsprache zu berücksichtigen: *T1 Professional* (Langenscheidt, GMS), *Personal Translator Plus ’98* (Linguatec/Rheinbaben & Busch, IBM), *Power Translator Pro* (Globalink), *SYSTRAN PROfessional für Windows* (Systran), *Transcend* (HEI-Soft,

Intergraph), *Logos* (Logos). Systeme der untersten Preisklasse wurden aufgrund ihrer in verschiedenen bereits durchgeführten Testläufen als unbefriedigend eingestuften Übersetzungsqualität von vornherein nicht einbezogen.

Aus einer Reihe von linguistisch für das betreffende Sprachpaar als relevant eingestuften Phänomenen wurden zunächst imperativische Strukturen, idiomatische Wendungen sowie die Erkennung von Komposita zu einer Recherche und Analyse an den für das Testkorpus vorgesehenen Textsorten ausgewählt. Ausschlaggebend für die Zusammenstellung des Testkorpus war die Frage, welche Textsorten aufgrund ihrer sprachlichen Struktur für eine maschinelle Übersetzung geeignet sind bzw. im professionellen Umfeld häufig mit maschineller Unterstützung übersetzt werden. Von den zunächst ausgewählten Textsorten bzw. Textsortenklassen, die sich aus Handbüchern, Instruktionstexten, Firmenjahresberichten, Handelskorrespondenz, Web-Seiten von Reiseanbietern sowie Web-Seiten aus dem Bereich des *Electronic Commerce* zusammensetzten, wurden auf dem Arbeitstreffen am 8. Mai auf der Basis der in der Zwischenzeit durchgeführten Textrecherchen und linguistischen Untersuchungen schließlich Instruktionstexte, d. h. Reparaturanweisungen aus der Automobilbranche und Softwareinstallationsanleitungen, sowie Web-Seiten aus den Bereichen Tourismus und *Electronic Commerce* ausgewählt.

Um die Vergleichbarkeit der in die Evaluation einbezogenen Systeme sowie eine möglichst hohe Transparenz in bezug auf die Testergebnisse zu gewährleisten, beschränkt sich die Evaluation zum einen ausschließlich auf linguistische Phänomene. Zum anderen ist die angestrebte Vergleichbarkeit und Transparenz auch der Grund, warum im Rahmen der ersten Evaluationsphase *Translation Memories* bzw. Satzarchive nicht einbezogen werden, selbst wenn diese bereits als Module einzelner Systeme angelegt sind. – Die Festlegung auf ein solches Verfahren führte schließlich dazu, idiomatische Wendungen aus dem Katalog der linguistischen Phänomene auszuklammern und stattdessen Konditionalsätze und syntaktische Koordinationen in die Untersuchung einzubeziehen. Die Evaluation der einzelnen linguistischen Phänomene wird jeweils von einem Evaluator bzw. einer Evaluatorin hauptverantwortlich durchgeführt, wobei alle Phänomene anhand von jeweils 300 Testsätzen an jedem der ausgewählten maschinellen Übersetzungssysteme überprüft werden.

Die Bewertung der grammatischen Korrektheit der maschinellen Übersetzung der ausgewählten linguistischen Phänomene soll anhand eines viergliedrigen Klassifikationsschemas erfolgen, das die Bewertungspunkte „Satz bzw. Syntagma vollständig korrekt“, „Satz bzw. Syntagma in bezug auf das zu überprüfende

linguistische Phänomen grammatisch korrekt“, „Satz bzw. Syntagma hinsichtlich des zu überprüfenden linguistischen Phänomens falsch“ und „ Satz bzw. Syntagma falsch; Fehlerursache nicht eindeutig entscheidbar“ umfaßt.

Die Ergebnisse der ersten Evaluationsphase, die auf den Arbeitskreistreffen im Januar und im Mai initiiert wurde, sollen im Rahmen eines Workshops auf der Konvens ’98 einem größeren Publikum vorgestellt und mit Anwendern maschineller Übersetzungssysteme, Vertretern aus dem Bereich der Forschung und dem industriellen Entwicklungsumfeld diskutiert werden.