Ausblick

Uta Seewald-Heeg (Hochschule Anhalt)
Rita Nübel (IAI Saarbrücken)

1 Evaluierung von Translation-Memory-Systemen

Eine Bewertung der Leistungsfähigkeit von Translation Memories bzw. Satzarchiv-Modulen vollautomatischer Übersetzungssysteme stellt unterschiedliche Anforderungen an das Evaluierungsszenario, das Datenmaterial und die Durchführung der Evaluierung. Voraussetzungen hierzu wurden in der abschließenden Diskussion des vom Arbeitskreis „Maschinelle Übersetzung“ veranstalteten Workshops auf der Basis der Beiträge formuliert, die in diesem Band zusammengestellt wurden, um einzelne Schritte des Evaluierungsvorhabens festzulegen.

2 Evaluationsdesign

Ausgangspunkt der Diskussion um Qualitätsanforderungen und mögliche Testprozeduren bildete das von der EAGLES-Arbeitsgruppe „Assessment and evaluation“ [EAGLES96] entworfene Standardverfahren zur Evaluierung von NLP-Software, das sich an der ISO-Norm 9126 orientiert und im wesentlichen als dreistufiges Verfahren konzipiert ist.[1] Die dort formulierten Qualitätskriterien für natürlichsprachige Systeme umfassen Funktionalität (*functionality*), Zuverlässigkeit (*reliability*), Brauchbarkeit (*usability*), Leistungsfähigkeit (*efficiency*), Wartbarkeit (*maintainability*) sowie Portierbarkeit (*portability*) des jeweils untersuchten Softwareproduktes.

In einem eigens für Translation-Memory-Systeme entworfenen Evaluationsdesign wird darüber hinaus zwischen sogenannten *on-line*-Funktionen und *off-line*-Funktionen von TMs unterschieden. Zu den *off-line*-Funktionen werden die Art der Textanalyse, die Möglichkeiten des Imports und Exports von Text in die TM-Datenbank, die Segmentierung der Texteinheiten bei der Übersetzung und beim Alignment, d.h. dem Import bereits übersetzer Texte und deren quellsprachigen Dokumenten in synchronisierter Form, sowie die für das Alignment selbst zur Verfügung stehenden Funktionen gezählt. Als *on-line*-Funktionen werden demgegenüber Parameter wie die maximale Größe des TM, die Retrieval-Geschwindigkeit bei der Suche im Übersetzungsarchiv sowie die Trefferquote beim Abgleich eines zu übersetzenden Satzes mit den Daten des TM bezeichnet. Dabei geht in

die Bewertung der einzelnen Funktionen stets die Zahl der jeweils benötigen Arbeitsschritte ein.

Neben Qualitätsanforderungen an Translation-Memory-Systeme wurden auch Metriken für die Tests festgelegt sowie Anforderungen an adäquates Testmaterial und organisatorische Fragen zur konkreten Durchführung der Evaluierung erörtert.

2.1 Qualitätskriterien für Translation-Memory-Systeme

Als wesentliche Gründe für den professionellen Einsatz von Translation Memories werden

- bessere Übersetzungsqualität durch konsistente Übersetzungen
- Zeiteinsparung durch die Reduktion von Mehrfachübersetzungen
- Wiederverwendbarkeit von Daten

genannt. Hinsichtlich der Wiederverwendbarkeit bereits übersetzter Texte oder Textfragmente spielt vor allem die Retrievalfunktion der Translation Memories eine entscheidende Rolle. Die für die Evaluierung zentralen Qualitätsmerkmale betreffen daher in erster Linie die *on-line*-Funktionen des EAGLES-Evaluationsdesigns.[2]

Unter den *on-line*-Funktionen soll zunächst die Bewertung der Retrievalleistung im Vordergrund der Evaluierungsaktivitäten stehen. Besonderes Augenmerk soll auf die Präzision der Treffer gelegt werden und die Verlässlichkeit der von den Systemen für Retrievals bei modifizierten Eingaben kalkulierten Trefferraten, die mittels *Fuzzy-Match*-Algorithmen ermittelt werden. Wie der Beitrag von Seewald-Heeg und Nübel illustriert, sind beispielsweise die Match-Werte, die von den Satzarchivmodulen der untersuchten maschinellen Übersetzungssysteme berechnet werden, zum einen unplausibel, wenn man die Angaben mit den tatsächlich gelieferten Retrievals vergleicht; zum anderen weichen die Angaben beim Vergleich der beiden Systeme erheblich voneinander ab, was gegen Aussagen über allgemeingültige Schwellenwerte für brauchbare bzw. nicht brauchbare Kandidaten für Übersetzungen spricht. Wenn ein Überseter beispielsweise den häufig als Schwellenwert für brauchbare Retrievals angesehenen Match-Wert von 70% festlegt, ist für ihn entscheidend, ob die beim Retrieval ermittelten Kandidaten oberhalb dieser Grenze tatsächlich ohne unverhältnismäßig hohen zusätzlichen Arbeitssaufwand (z.B. umfangreiche Posteditionsaktionen) für die Er-

stellung einer Übersetzung verwendet werden können, bzw. ob die ermittelten Kandidaten, die unterhalb dieser Grenze liegen, tatsächlich nicht oder nur mit erhöhtem Arbeitsaufwand in die Übersetzung integriert werden können.

Es ist beabsichtigt, neben der Retrievalleistung auch die Alignment-Funktionen der Systeme in die Evaluierung einzubeziehen. Darüber hinaus sollen auch die von den verschiedenen Translation-Memory-Systemen verarbeitbaren Dateiformate berücksichtigt werden, da sie bei der Entscheidung für ein bestimmtes System für zahlreiche Anwender von zentraler Bedeutung sind.

2.2 Evaluierungsmethodologie

Es wird davon ausgegangen, dass bei der Evaluierung keine vorab bereits angelegten TM-Datenbanken verwendet werden. Die Translation Memories werden erst nach und nach aufgebaut, so dass sich der Umfang des Referenzmaterials im Laufe der Evaluierung schrittweise vergrößert.

Für die Evaluierung der *Fuzzy-Match*-Funktion soll zunächst eine Klassifikation möglicher im Text auftretender Modifikationstypen erstellt werden, die in den Testdaten entsprechend reflektiert sein muss. Die Bewertung der Erkennungsleistung der verschiedenen Modifikationen beim Retrieval soll auf einer qualitativen Klassifikation der Retrievalergebnisse erfolgen, die anschließend quantitativ ausgewertet wird. Hierzu muss zunächst ein Bewertungsschema konzipiert werden, das das Maß der Ähnlichkeit zwischen Testsatz und Referenzsatz berücksichtigt.

2.3 Systeme

Die Auswahl der zu evaluierenden Systeme hängt u.a. von ihrer Verfügbarkeit an den einzelnen Evaluierungsstandorten ab, die *nicht* die Standorte der industriellen Anbieter sein werden, um größtmögliche Objektivität und identische Bedingungen für die Durchführung des Evaluierungsvorhabens zu garantieren. Neben den Translation Memories der Firmen Trados und Star, die bereits auf dem Workshop vorgestellt wurden, sollen möglichst viele der auf dem Markt verfügbaren Translation-Memory-Systeme in die Evaluierung einbezogen werden.

3 Durchführung der Evaluierung und Auswertung der Ergebnisse

Im Anschluss an die Zusammenstellung des oben beschriebenen Testkorpus sollen Teile des Korpus mittels der Alignment-Funktionen in die zu evaluierenden Systeme eingelesen werden. In den Testdurchläufen sollen die *Fuzzy-Match*-Funktionen dann jeweils mit den modifizierten Daten bzw. aktualisierten Textversionen getestet und die Retrievalergebnisse entsprechend dem oben beschriebenen Klassifikationsschema bewertet werden.

Es ist geplant, die Ergebnisse der Evaluation abschließend zusammenfassend zu dokumentieren und der Öffentlichkeit zu präsentieren.

Literatur

[**EAGLES96**] EAGLES Evaluation of Natural Language Processing Systems. Final Report. EAGLES Document EAG-EWG-PR.2, Oktober 1996.

ANMERKUNGEN

[1] Vgl. den URL http://www.issco.unige.ch/projects/ewg96/ewg96.html.

[2] Siehe auch das Protokoll der Sitzung des Arbeitskreises „Maschinelle Übersetzung“ vom 19.2.1999, das auf dem URL
http://www.heeg.de/~uta/AK-Protokoll-TM-1.html
publiziert ist.