Autorenunterstützung für die Maschinelle Übersetzung

Melanie Siegel[1]

Zusammenfassung

Der Übersetzungsprozess der Technischen Dokumentation wird zunehmend mit Maschineller Übersetzung (MÜ) unterstützt. Wir blicken zunächst auf die Ausgangstexte und erstellen automatisch prüfbare Regeln, mit denen diese Texte so editiert werden können, dass sie optimale Ergebnisse in der MÜ liefern. Diese Regeln basieren auf Forschungsergebnissen zur Übersetzbarkeit, auf Forschungsergebnissen zu Translation Mismatches in der MÜ und auf Experimenten.

Einleitung

Mit der Internationalisierung des Markts für Technologien und Technologieprodukte steigt die Nachfrage nach Übersetzungen der Technischen Dokumentation. Vor allem in der Europäischen Union steigt das Bewusstsein, dass es nicht ausreicht, englischsprachige Dokumentation zu liefern, sondern dass Dokumentation in die Muttersprache der Kunden übersetzt werden muss.

Diese Übersetzungen müssen schnell verfügbar, aktualisierbar, in mehreren Sprachen gleichzeitig verfügbar und von hoher Qualität sein.

Gleichzeitig gibt seit einigen Jahren erhebliche technologische Fortschritte in der Maschinellen Übersetzung: Es gibt regelbasierte und statistische Systeme, aber auch hybride Übersetzungsverfahren. Diese Situation hat dazu geführt, dass Firmen mehr und mehr versuchen, ihre Übersetzungsanstrengungen mit MÜ zu unterstützen. Dabei treten allerdings eine Reihe von Problemen auf:

1. Die Nutzer kennen die Möglichkeiten und Grenzen der MÜ nicht gut genug. Sie werden in ihren Erwartungen enttäuscht.
2. Um die Systeme zu testen, werden völlig ungeeignete Texte übersetzt, wie z. B. Prosa.
3. Auch Technische Dokumentation, die an die MÜ geschickt wird, ist oft nicht von ausreichender Qualität, ebenso wenig wie Texte, die an humane Übersetzer geschickt werden. Allerdings können humane Übersetzer diesen Mangel an Qualität im Ausgangsdokument ausgleichen, während MÜ-Systeme dazu nicht in der Lage sind.
4. Statistische MÜ-Systeme müssen auf parallelen Daten trainiert werden. Oft werden dafür TMX-Dateien verwendet, die aus Translation Memory – Systemen herausgezogen werden. Da aber diese Daten oft unsauber sind und fehlerhafte und inkonsistente Übersetzungen enthalten, ist auch die Qualität der trainierten Übersetzung schlecht.

Wir haben uns mit der Frage beschäftigt, wie die Autoren Technischer Dokumentation darin unterstützt werden können, Dokumente für die MÜ optimal vorzubereiten, um auf diese Weise optimale Übersetzungsergebnisse zu bekommen.

Das Ziel der Untersuchungen ist, die Möglichkeiten und Grenzen der MÜ genauer zu spezifizieren, daraus Handlungsoptionen für Autoren abzuleiten und diese durch automatische Verfahren zu unterstützen. Dabei gehen wir in drei Schritten vor:

- Wir untersuchen die Schwierigkeiten, die ein humaner Übersetzer hat, darauf, ob sie auf MÜ-Systeme übertragbar sind.

[1] Siegel, Melanie (2011): Autorenunterstützung für die Maschinelle Übersetzung. In: Hedeland, H., Schmidt, T. und Wörner, K. (eds.): Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011.

- Wir experimentieren mit automatisch prüfbaren Regeln der Autorenunterstützung und übersetzen Texte vor und nach der Umformulierung mit MÜ.
- Wir ziehen Untersuchungen zu „Translation Mismatches“ in der MÜ heran, um Strukturen zu finden, die besonders schwer automatisch übersetzbar sind.

Schwierigkeiten von humanen Übersetzern – Schwierigkeiten von MÜ-Systemen

Heizmann (1994:5) erläutert den Übersetzungsprozess für humane Übersetzer: "*In our opinion, translation is basically a complex decision process. The translator has to base his or her decisions upon available information, which he or she can get from various sources.*" Diese Aussage ist auch auf den Übersetzungsprozess in der MÜ übertragbar und verdeutlicht schon, dass es notwendig ist, der Maschine möglichst wenige komplexe Entscheidungsprozesse aufzubürden.

Ausgehend davon, dass ein MÜ-System einem eher unprofessionellen Übersetzer ähnlich ist, dem die Texte für die Übersetzung so vorbereitet werden sollten, dass sie einfacher übersetzbar sind, ziehen wir Parallelen vom unprofessionellen Übersetzer zum MÜ-System. Der Ausgangstext für Übersetzer wie für ein MÜ-System muss so angepasst werden, dass die Probleme möglichst umgangen werden, die der unprofessionelle Übersetzer und das MÜ-System haben:

- Die Übersetzung einzelner Wörter, Phrasen und Sätze, ohne die Möglichkeit, größere Übersetzungseinheiten in Betracht zu ziehen, erfordert, dass satzübergreifende Bezüge vermieden werden müssen, wie z.B. Anaphern.
- Die Unmöglichkeit der Paraphrasierung erfordert einfache Satzstrukturen ohne Ambiguitäten. Wichtig ist es auch, metaphorische Sprache zu vermeiden, da diese oft nicht einfach übersetzt werden kann, sondern Paraphrasierung erfordert.
- Eine Übersetzung ohne Weltwissen führt dazu, dass Wörter mit unterschiedlichen Bedeutungen in verschiedenen Domänen (Homonyme) falsch übersetzt werden. Solche potentiell ambigen Wörter müssen vermieden werden.
- Da das Spektrum von Übersetzungsvarianten potentiell größer als bei professionellen Übersetzern ist, ist eine systematische Terminologiearbeit am Ausgangstext hilfreich, die Terminologievarianten im Ausgangstext schon mal eliminiert.
- Da die MÜ ebenso wie der unprofessionelle Übersetzer wenige Hilfsmittel hat, die Hintergrundwissen zum beschriebenen Sachverhalt geben, muss die Beschreibung möglichst klar und verständlich sein. Das erfordert einfache Satzstrukturen.

Relevanz von automatisch prüfbaren Regeln der Autorenunterstützung

In einem Experiment haben wir einige Dokumente der technischen Dokumentation mit dem MÜ-System Langenscheidt T1 übersetzen lassen. Danach haben wir die Dokumente mit einer großen Anzahl automatisch prüfbarer Regeln aus acrolinx IQ geprüft. Die Ergebnisse der Prüfungen haben wir umgesetzt, indem wir die Ausgangstexte umformuliert haben. Diese umformulierten Texte haben wir dann wieder mit Langenscheidt T1 automatisch übersetzt und die Übersetzungen miteinander verglichen. Das Ziel dieses Experiments ist es, herauszufinden, welche Regeln der Autorenunterstützung wichtige Effekte auch für die MÜ haben. Einige dieser Regeln haben wir im vorangegangenen Abschnitt Schwierigkeiten von humanen Übersetzern – Schwierigkeiten von MÜ-Systemen schon vorgestellt.

Aufgrund dieser Experimente haben wir ein Regelset zusammengestellt, das wir im nächsten Abschnitt vorstellen.

Erste Ergebnisse der Experimente

Rechtschreibung und Grammatik: Das Regelset für die deutschen Ausgangstexte enthält zunächst die Standard-Grammatik- und Rechtschreibregeln. Die Experimente haben klar gezeigt, dass ein MÜ-System keine sinnvollen Ergebnisse liefert, wenn der Eingabetext Rechtschreib- und Grammatikfehler enthält. Wenn ein Wort unbekannt ist, weil es falsch geschrieben ist, dann ist auch keine Übersetzung mit dem MÜ-System möglich. Allerdings führt nicht jeder Rechtschreibfehler auch zu Übersetzungsproblemen: Die Experimente haben gezeigt, dass das untersuchte MÜ-System tolerant zu alter und neuer deutscher Rechtschreibung ist – beide Varianten „muß“ und „muss“ wurden korrekt übersetzt.

Regeln zu Formatierung und Zeichensetzung: Der Gebrauch von Gedankenstrichen führt zu komplexen Sätzen im Deutschen, die Probleme bei der Übersetzung bereiten.

Regeln zum Satzbau: Beim Satzbau geht es zunächst darum, komplexe Satzstrukturen zu vermeiden. Oberstes Gebot ist hier, zu lange Sätze zu vermeiden. Komplexe Satzstrukturen entstehen durch die folgenden Konstruktionen, wie Einschübe, Hauptsatzkoordination, Trennung von Verben, eingeschachtelte Relativsätze, Schachtelsätze, Klammern, Häufung von Präpositionalphrasen, Beschreibung mehrerer Handlungen in einem Satz, umständliche Formulierungen und Bedingungssätze, die nicht mit „wenn“ eingeleitet sind. Ein anderes Problem für die MÜ sind ambige Strukturen, die durch Substantivkonstruktionen und elliptische Konstruktionen entstehen.

Regeln zur Wortwahl: Füllwörter und Floskeln sind deshalb schwierig für die MÜ, weil nicht paraphrasiert werden kann. Das MÜ-System versucht, diese Wörter zu übersetzen, obwohl ein professioneller Übersetzer sie weglassen oder umformulieren würde. Umgangssprache und bildhafte Sprache sind ebenfalls ein großes Problem. Pronomen sind dann schwierig zu übersetzen, wenn der Bezug außerhalb des Satzkontexts liegt und unklar ist. Bei der Verwendung von ambigen Wörtern kann das MÜ-System in vielen Fällen die Ambiguität nicht auflösen. Das passiert zum Beispiel bei der Verwendung von Fragewörtern in anderen Kontexten als einer Frage. Gerade ausdrucksschwache Verben mit ambigem Bedeutungsspektrum sind problematisch. Der Nominalstil, bei dem Verben nominalisiert werden, kann im Englischen zu komplexen und falschen Konstruktionen führen.

Anwendung der Regeln, Umformulierungen und Übersetzungen

Ein wichtiger Teil der Fragestellung war aber nun, ob die Anwendung der implementierten Regeln zur Autorenunterstützung tatsächlich eine Auswirkung auf die Ergebnisse der MÜ hat. Im oben beschriebenen Experiment haben wir die aufgestellten und implementierten Regeln zur Autorenunterstützung auf zwei Dokumente angewendet und die Texte nach den Empfehlungen der Regeln umformuliert. Anschließend haben wir untersucht, welche der Regeln am häufigsten auftraten und die meisten Effekte für die Qualität der MÜ-Ausgaben hatten. Hier muss jedoch angemerkt werden, dass dieses Experiment bisher nur mit zwei Dokumenten durchgeführt wurde, einer Anleitung zum Ausbau von Zündkerzen am Auto und einer Anleitung zur Installation einer Satellitenschüssel. Ein interessantes Ergebnis: In fast der Hälfte der Fälle konnte der Satz anhand von lexikalisch-basierten Regeln so verbessert werden, dass die Maschinelle Übersetzung gute Ergebnisse lieferte.

Untersuchungen zu Translation Mismatches und daraus resultierende Empfehlungen

Kameyama et al. (1991) verwendeten den Begriff "Translation Mismatches", um ein Schlüsselproblem der maschinellen Übersetzung zu beschreiben. Bei Translation Mismatches handelt

es sich um Information, die in der einen am Übersetzungsprozess beteiligten Sprache explizit nicht vorhanden ist, die aber in der anderen beteiligten Sprache gebraucht wird. Der Effekt ist, dass die Information in der einen Übersetzungsrichtung verloren geht und in der anderen hinzugefügt werden muss.

Das hat - wie Kameyama beschreibt - zwei wichtige Konsequenzen:

"First in translating a source language sentence, mismatches can force one to draw upon information not expressed in the sentence - information only inferrable from its context at best. Secondly, mismatches may necessitate making information explicit which is only implicit in the source sentence or its context." (S.194)

Translation Mismatches sind für die Übersetzung eine große Herausforderung, weil Wissen, das nicht direkt sprachlich kodiert ist, inferiert werden muss. Welche Translation Mismatches relevant sind, das hängt aber stark von der Information ab, die in den beteiligten Sprachen kodiert ist. Für das Sprachpaar Deutsch-Englisch konnten wir in den Experimenten die folgenden Translation Mismatches identifizieren:

- **Lexikalische Mismatches**. Die Bedeutung ambiger Wörter in der Ausgangssprache muss in der Zielsprache aufgelöst werden, wie z.B. bei „über" -> „about", „above".
- **Nominalkomposita**. Nach den Regeln der deutschen Rechtschreibung müssen Nominalkomposita entweder zusammen oder mit Bindestrich geschrieben werden. Wenn sie zusammengeschrieben werden, muss die Analyse der MÜ die Teile identifizieren. Das ist aber nicht immer eindeutig im Deutschen, wie das Beispiel „Blumentopferde" - „Blumento-Pferde", „Blumentopf-Erde" zeigt. Wenn auch im Deutschen wie im Englischen ein Leerzeichen zwischen den Teilen des Kompositums steht, dann ist die MÜ-Analyse of überfordert, weil die Beziehung zwischen den Nomen unklar bleibt. Z.B.: „bei den heutzutage verwendeten Longlife Kerzen" - „at the nowadays used ones"
- **Metaphorik**. Bildhafte Sprache lässt sich nicht wörtlich übertragen. Ein Beispiel aus den Experimenten: „Man ist daher leicht geneigt" – „One is therefore slightly only still to"
- **Pronomen**. Das Pronomen „Sie" meint im Deutschen sowohl die 3. Person Singular als auch die 2. Person Singular, abhängig von der Großschreibung. Wenn das „Sie" aber am Satzanfang steht, bleibt unklar, welche Variante gemeint ist. Beispiel: „Sie haben es fast geschafft" – „her it have created almost".

Zusammenfassung und nächste Schritte

Wir haben ein Regelset für die automatische Autorenunterstützung aufgestellt. Dieses Regelset basiert auf Untersuchungen zu Problemen humaner Übersetzer, auf Experimenten mit MÜ und Umformulierungen und auf Untersuchungen zu Translation Mismatches in der MÜ.

Ein nächster Schritt wird sein, das entstandene Regelset in Experimenten mit verschiedenen MÜ-Systemen zu validieren. Die Übersetzungen werden dieses Mal von professionellen Übersetzern und Übersetzerinnen validiert. Leitfragen der Untersuchungen sind:

- Welche Regeln für das Pre-Editing sind am relevantesten?
- Gibt es weitere Regeln, die zum Regelset hinzugefügt werden?
- Sind die Regeln abhängig von der Zielsprache?
- Sind die Regeln abhängig vom MÜ-System?

Die Regeln für das Pre-Editing können zum Teil automatische Vorschläge für die Umformulierung geben. Wir suchen nach einem Weg, aus diesen Vorschlägen ein automatisches Pre-Editing zu erzeugen.