Automatische Autorenunterstützung für das Postediting in der Maschinellen Übersetzung

Melanie Siegel

Einleitung

Die Maschinelle Übersetzung (MÜ) ist in den letzten Jahren so weit fortgeschritten, dass eine Übersetzung technischer Dokumentation sinnvoll ist, wenn die Ergebnisse einen Postediting-Prozess durchlaufen. Das Postediting der Ergebnisse von MÜ ist daher eine häufig angewandte Methode, um die gewünschte Übersetzungsqualität zu erhalten. In unserem Vortrag stellen wir Verfahren der MÜ vor und zeigen, welchen Einfluss die Verfahren auf den Postediting-Prozess haben. Wir stellen unterschiedliche Postediting-Prozesse vor, wie monolinguale und multilinguale Prozesse. Wir zeigen dann, wie das Postediting durch automatische Prüf- und Korrekturverfahren effizient unterstützt werden kann, die auf Experimenten mit Übersetzern basieren.

Verfahren der MÜ und deren Einfluss auf den Postediting-Prozess

Zwei grundsätzliche Verfahren liegen den heutzutage häufig genutzten MÜ-Systemen zugrunde: Statistische Verfahren zur MÜ und regelbasierte Verfahren zur MÜ (siehe Koehn 2010 und Somers 2003).

Die statistischen Verfahren (SMT) basieren auf einer großen Menge von bereits übersetzten Sätzen. Diese Daten stehen erst in den letzten Jahren in großem Maße, z.B. als Translation Memories, zur Verfügung. Aus den Sätzen werden in der Trainingsphase mit statistischen Methoden Phrasen und ihre Übersetzungen in den Sätzen der Zielsprache extrahiert. Diese Phrasen sind dann die Grundlage für die maschinelle Übersetzung neuer Sätze. Wenn die Trainingsdaten aus einer ähnlichen Domäne stammen wie die Sätze, die übersetzt werden sollen, haben die statistischen Verfahren hier einen klaren Vorteil in der Terminologieübersetzung. Zum Beispiel wird Bank dann im Kontext der Geldwirtschaft ins Englische mit „bank“ übersetzt und nicht etwa mit „bench“. Linguistische Information wie z.B. Grammatik spielt jedoch keine Rolle in diesen Verfahren. Daher muss der Postediting-Prozess sich bei statistischen Verfahren auf grammatische Korrektheit der zielsprachlichen Sätze konzentrieren. Ein wichtiges Problem ist auch der Negationsskopus. Da keine Satzanalyse gemacht wird, kann es passieren, dass Negation nicht richtig übersetzt wird. Ein Beispiel aus einem Experiment mit dem SMT-System Moses (Google Translate gibt ähnliche Ergebnisse):

DE: „Ich habe meinen Computer neu gestartet, aber ich kann weiterhin Websites mit kompromittierendem Inhalt öffnen.“

EN: “I have restarted my computer, but I still can not open websites with content

kompromittierendem."

Die regelbasierten Verfahren (RBMT) basieren auf einer linguistischen Analyse des Ausgangssatzes und einer abstrakten Repräsentation der Bedeutung. Diese Repräsentation wird in einem Transfer-Prozess in eine Repräsentation der Zielsprache überführt. Mit der Grammatik der Zielsprache wird daraus dann ein Satz generiert. Der Vorteil dieses Verfahrens ist, dass grammatisch korrekte Sätze in der Zielsprache generiert werden. Das Verfahren ist satzbasiert und hat keinen Zugriff auf Information aus voranstehenden Sätzen. Das Postediting muss daher Kontextinformation berücksichtigen, wenn z.B. im englischen Satz „it" steht und in der deutschen Übersetzung „es", was aber im Kontext „sie" sein sollte. Die Terminologie wird hier mit einem Lexikon übersetzt, was zu Terminologie-Übersetzungsfehlern führen kann.

Monolinguales und multilinguales Postediting

Grundsätzlich ist es sinnvoll, beim Postediting sowohl den Satz der Ausgangssprache als auch den der Zielsprache zu berücksichtigen. Dies ist jedoch nicht immer möglich, wenn die posteditierende Person die Ausgangssprache nicht kennt oder wenn automatische Postediting-Verfahren angewendet werden, die nur auf den zielsprachlichen Sätzen operieren. Daher können einige Probleme der MÜ mit monolingualem Postediting nicht gelöst werden. Dazu gehören Probleme in der Terminologie, aber auch der Negationsskopus. Monolinguales Postediting kann aber grammatische Probleme der SMT wie Kongruenz oder Kommasetzung gut lösen. Auch andere Probleme wie die Korrektur von Tempus können gut monolingual gelöst werden: Das Futur wird im Deutschen häufig mit Präsensformen ausgedrückt, aber durch die MÜ entstehen Futur-Formen (die in der Ausgangssprache verwendet werden).

Unterstützung des Postediting-Prozesses durch automatische Prüf- und Korrekturverfahren

Automatische Prüf- und Korrekturverfahren sind aus der Autorenunterstützung für Technische Dokumentation entstanden, wie sie z.B. von Acrolinx angeboten wird. Sie arbeiten auf dem zielsprachlichen Text, sind also monolingual. Dennoch gibt es eine Reihe von automatisch prüfbaren Regeln zum Postediting, die hier aufgelistet werden:

Artikellose Nominalphrasen vermeiden: Gerade bei der Übersetzung aus Sprachen, die keine Artikel verwenden (wie z.B. Japanisch) müssen Artikel im Postediting-Prozess häufig eingefügt werden.

Futur vermeiden: Im Deutschen werden häufig Präsensformen verwendet, um Ereignisse, die in der Zukunft stattfinden, zu beschreiben. Das ist in anderen Sprachen nicht der Fall, so dass die Futur-Formen übersetzt werden.

Imperativ-Korrektur: Im Spanischen wird z.B. der Konjunktiv für den Imperativ verwendet. Ein RBMT-System überträgt die Konjunktiv-Form, die im Deutschen falsch ist.

Nominalkomposita zusammenschreiben: Im Englischen werden Nominalkomposita mit Leerzeichen gebildet, was im Deutschen nicht korrekt ist und im Posteditng-

Prozess korrigiert werden muss.

umständliche Formulierungen vermeiden: Durch die Übersetzung entstehen im Deutschen umständliche Formulierungen wie z.B. „eine Einstellung vornehmen“. Diese Formulierungen können im Postediting-Prozess vereinfacht werden.

veraltete Wörter vermeiden: Abhängig von den Trainingsdaten im SMT und den Lexika im RBMT werden Wörter wie „vermögen“ oder „mittels“ übersetzt, die im Postediting-Prozess geändert werden.

Verbkorrektur: Einige systematische domänenbasierte Übersetzungsfehler von Verben können auch monolingual korrigiert werden. Z.B. wurde in unseren Übersetzungsexperimenten im Software-Kontext häufig das Verb „schleppen“ übersetzt, das systematisch mit „ziehen“ korrigiert werden konnte.

von-Serie vermeiden: Durch die Übersetzung entstehen im Deutschen komplexe Nominalphrasen, die besser durch ein Kompositum ersetzt werden. Z.B.: „Adresse von 32 Bit von Internet“ - „32-Bit-Internet-Adresse”.

Zusammenfassung

Maschinelle Übersetzung hat in den letzten Jahren große Fortschritte gemacht und ist daher auch in der Technischen Dokumentation nutzbar. Dennoch ist Postediting notwendig, um einen ausreichenden Qualitätsstandard der Übersetzungen zu erreichen. Es ist notwendig, sich vor dem Postediting über das MÜ-Verfahren zu informieren, um die Probleme des gewählten Verfahrens gezielt anzugehen. Auch wenn multilinguales Postediting wünschenswert ist, kann man mit einem Blick auf den zielsprachlichen Text viele Korrekturen durchführen. Automatische Prüf- und Korrekturverfahren sind hier hilfreich – wie bei der Autorenunterstützung der technischen Dokumentation durch automatische Tools.

Literatur

Koehn, P. (2010): Statistical Machine Translation. Cambridge University Press, New York.

Somers, H. (2003): Machine translation: Latest developments. In R. Mitkov (Hrsg.), The Oxford Handbook of Computational Linguistics, Chapter 28, S. 512-528. Oxford University Press.

für Rückfragen: melanie.siegel@acrolinx.com