January 15, 2016

Do-it-yourself-MÜ

Uwe Muegge

Foto: (c) Ingo Bartussek – fotolia.com

Mit statistischer Maschinenübersetzung schneller zu besseren Ergebnissen

Do-it-yourself-MÜ

Cloudbasierte SaaS-Lösungen, bei denen die Software im Abonnement angeboten wird, eröffnen seit einiger Zeit auch kleinen Unternehmern und Einzelübersetzern die Möglichkeit, konfigurierbare Maschinenübersetzung zu überschaubaren Kosten zu nutzen. Uwe Muegge beschreibt die Funktionsweise der leistungsfähigen Technologie und stellt exemplarisch einige Anbieter vor.

Cloudbasierte statistische Maschinenübersetzungsdienste (SaaS-Lösungen, von „Software as a Service"), bei denen Benutzer auf der Grundlage der eigenen Translation Memories kundenspezifische MÜ-Engines erstellen können, halten Einzug in die Übersetzungsbranche. Die meisten bieten ihre Software als Abonnement an, so dass die Anlaufkosten für die DIY-Maschinenübersetzung wesentlich geringer sind als die einer traditionell lizensierten MÜ-Lösung. Dadurch ist diese sehr leistungsfähige Übersetzungstechnologie selbst für kleine Unternehmen erschwinglich.

Bei der statistischen Maschinenübersetzung (SMÜ), der momentan dominanten Technologie für maschinelle Übersetzung, werden umfangreiche Textkorpora – sowohl zweisprachige (in der Ausgangs- und Zielsprache) als auch einsprachige (nur in der Zielsprache) – für das Training von Übersetzungsengines eingesetzt. Mit dem Google Übersetzer steht heute ein auf der statistischen Maschinenübersetzung beruhender Online-Dienst zur Verfügung, der die kostenlose Übersetzung zwischen 90 Sprachen ermöglicht. Allerdings war und ist die Qualität, die generische statistische MÜ-Dienste wie der Google Übersetzer bei der Übersetzung komplexer Fachtexte liefern, nur mit erheblichem Aufwand auf ein mit einer Humanübersetzung vergleichbares Niveau zu bringen.

Im Laufe der letzten Jahre wurden zahlreiche Online-Dienste gegründet, die ihren Benutzern ein viel höheres Maß an Kontrolle über die maschinell erstellten Übersetzungen geben, als das mit dem Google Übersetzer oder dem Microsoft Bing Übersetzer möglich ist. Das Geschäftsmodell dieser Dienstleister ist einfach: erstens, ein leistungsfähiges statistisches Maschinenübersetzungssystem über das Internet zugänglich machen und zweitens, es den Benutzern ermöglichen, durch das Hochladen ihrer

Translation Memories eigene, kundenspezifische Übersetzungs-Engines zu erstellen. Und das Beste: Auf diesem Wege können die Kosten für die Benutzung eines an die kundenspezifischen Bedürfnisse angepassten statistischen Maschinenübersetzungssystems im Vergleich zu anderen qualitativ hochwertigen MÜ-Lösungen drastisch gesenkt werden.

Traditionelle Ansätze für die Anpassung von SMÜ

Google Translator Toolkit

Der Google Übersetzer wird heute von über 200 Millionen Benutzern pro Tag eingesetzt, was den statistischen MÜ-Dienst zum meistgenutzten Übersetzungsservice der Welt macht. Im Jahr 2009 gab Google den Startschuss für das Translator Toolkit: ein kostenloses cloudbasiertes Translation-Memory-System, das Anwendern das Post-Editing der vom Google Übersetzer erzeugten maschinellen Übersetzungen ermöglicht. Benutzer können auch ihre eigenen Terminologiedatenbanken in den Google Übersetzer hochladen, aber diese Glossare können nur für das Post-Editing verwendet werden und nicht für die terminologische Anpassung der maschinellen Übersetzungen, die der Google Übersetzer erstellt.

Mit anderen Worten: Benutzer des Google Translator Toolkit können lediglich die vom Google Übersetzer erstellten Übersetzungen nachbearbeiten und im Idealfall früher vorgenommene Post-Edits wiederverwenden. Das Google Translator Toolkit bietet also derzeit keine Funktionalität, mit der auf die Art, wie der Google Übersetzer Übersetzungen erstellt, Einfluss genommen werden könnte.

Lizenzieren einer SMÜ-Software

Bis vor kurzem waren der Erwerb einer Lizenz für ein statistisches Maschinenübersetzungssystem und die Installation der Software auf dem eigenen Server die einzige Möglichkeit, die Leistungsfähigkeit einer SMÜ-Lösung in vollem Umfang zu nutzen. Der Marktführer in dieser Nische war Language Weaver, ein Unternehmen, das 2010 von SDL übernommen wurde. SDL bietet den Language Weaver heute als LW Enterprise Translation Server an. Wie der Name des Produkts bereits suggeriert, ist der Enterprise Translation Server in erster Linie für große Unternehmen und Behörden gedacht, die neben der Software auch deren Installation und die kundenspezifische Anpassung einkaufen.

Der große Nachteil dieser Art von SMÜ-Lösung sind die hohen Anlaufkosten: Neben den Kosten für die Softwarelizenz und die kundenspezifische Anpassung der Software, das sogenannte Training, fallen weitere nicht unerhebliche Kosten an; beispielsweise für den Unterhalt der Server, auf denen die Software läuft.

SMÜ-System Moses

Moses wurde 2005 als wissenschaftliches Experiment an der University of Edinburgh ins Leben gerufen und hat sich seitdem zu einem Open-Source-Projekt entwickelt, das von zahlreichen Institutionen auf der ganzen Welt unterstützt wird. Das Schöne an Moses ist, dass dieses SMÜ-System nicht nur kostenlos zur Verfügung steht, sondern auch für beliebige Sprachkombinationen geeignet ist. Doch das Ganze hat einen großen Nachteil: Moses ist nicht gerade eine leicht zu bedienende Anwendung. Genau gesagt handelt es sich bei Moses nicht einmal um eine Anwendung, sondern um eine Sammlung von Softwaremodulen, die auf einem Linux-Server – am besten einem Computer-Cluster – ausgeführt werden.

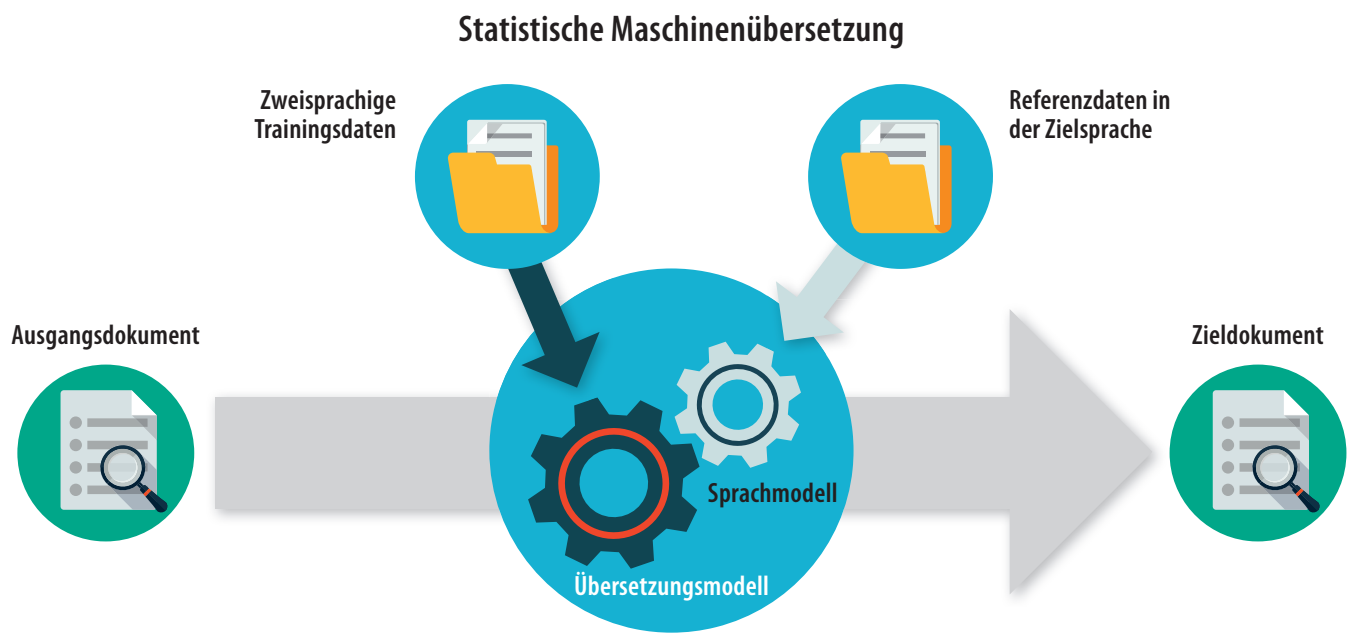
Hinzu kommt, dass die Benutzung von Moses fundierte Kenntnisse in Computerlinguistik und Informatik voraussetzt. Obwohl das SMÜ-System selbst kostenlos ist, stellt Moses also hohe Anforderungen an speziell geschultes Personal und eine aufwändige IT-Infrastruktur, was diese SMÜ-Lösung für die meisten kleinen und mittleren Unternehmen uninteressant macht.

So funktioniert statistische Maschinenübersetzung im DIY-Verfahren

Eine cloudbasierte Lösung verwenden

Als vor ein paar Jahren eine Handvoll Unternehmen damit begann, statistische Maschinenübersetzung als Webanwendung (SaaS) anzubieten, die Benutzer selbst ohne besondere Kenntnisse an ihre kundenspezifischen Bedürfnisse anpassen konnten, war das ein echter Paradigmenwechsel. Denn bei diesen cloudbasierten SMÜ-Diensten fallen für die Anwender keine großen Investitionen für Software, Hardware und Personal an. In diesem SaaS-Modell müssen sich die Benutzer weder mit den Komplexitäten der statistischen Maschinenübersetzung auseinandersetzen noch sie auch nur verstehen.

In einer Do-it-yourself-SMÜ-Umgebung laden die Benutzer ihre eigenen Daten (Translation-Memorys, Glossare, Referenztexte in der Zielsprache) für die Erstellung der kundenspezifischen MÜ-Engines hoch; danach müssen lediglich die zur Übersetzung bestimmten Dokumente in der Ausgangssprache hoch- und die übersetzten Dokumente heruntergeladen werden. Für diese Transaktionen stellen die Betreiber dieser SMÜ-Dienste eine einfach zu bedienende Projektmanagementsoftware zur Verfügung, auf die mit den üblichen Webbrowsern zugegriffen wird.



Beim Kunden muss also keine spezielle Übersetzungssoftware installiert und gepflegt werden.

Die eigenen Daten nutzen

Im Unterschied zum Google Translator Toolkit, bei dem Benutzer keine über das Post-Editing hinausgehende Möglichkeit zur Beeinflussung der Übersetzung haben, machen es DIY-MÜ-Systeme den Benutzern leicht, kundenspezifische Übersetzungsengines zu erstellen. Sinn und Zweck dieser MÜ-Engines ist es, Übersetzungen zu erstellen, die von Anfang an den terminologischen und stilistischen Konventionen des Auftraggebers entsprechen. Mit anderen Worten: Nach dem Erstellen der kundenspezifischen Übersetzungs-Engines sollten die von einem DIY-SMÜ-System erstellten maschinellen Übersetzungen qualitativ besser sein und weniger Nachbearbeitung erfordern als Übersetzungen, die von einem generischen SMÜ-System wie etwa dem Google Übersetzer erzeugt werden.

Die Anpassung eines SMÜ-Systems, das Training, erfolgt mit kundenspezifischen Sprachdaten. Für das Training des Übersetzungsmodells, das eine Reihe von möglichen Übersetzungen erzeugt, lädt der Benutzer möglichst umfangreiche zweisprachige Korpora hoch, in der Regel Translation-Memorys und Glossare. Für das Training des Sprachmodells, das in einem SMÜ-System dafür zuständig ist, die wahrscheinlich beste unter den vom Übersetzungsmodell erzeugten Übersetzungen zu bestimmen, lädt der Benutzer ein einsprachiges Textkorpus in der Zielsprache hoch.

Es gibt zwei wichtige Punkte, über die man sich als potentieller Benutzer einer statistischen Maschinübersetzungslösung im Klaren sein sollte: Der erste Punkt ist, dass für die kundenspezifische Anpassung eines SMÜ-Systems relativ umfangreiche Trainingskorpora benötigt werden, wobei 10.000 Sätze pro Korpus die untere Grenze sind. Das bedeutet, dass für die Erstellung einer brauchbaren kundenspezifischen SMÜ-Engine in jedem Sprachpaar ein umfangreicher Bestand an Translation-Memorys vorhanden sein muss. Und nicht nur das: Die für das SMÜ-Training vorgesehenen Translation-Memorys sollten einen Qualitätssicherungsprozess durchlaufen haben (also zumindest revidiert sein), denn nur mit hochwertigen Trainingsdaten können SMÜ-Systeme hochwertige Übersetzungen erstellen.

Der zweite Punkt ist die Tatsache, dass die kundenspezifische Anpassung eines SMÜ-Systems sehr punktuell sein kann. Mit anderen Worten: Es ist sinnvoll, nicht nur eine Übersetzungs-Engine pro Sprachkombination und Unternehmen einzurichten, sondern zumindest eine für jede Textsorte (z.B. Marketing, technische Dokumentation, Recht, Software). Es ist durchaus möglich, noch weiter spezialisierte SMÜ-Engines zu trainieren, etwa für einzelne Produktfamilien. Generell gilt, dass das Training der Übersetzungs-Engines keine einmalige Angelegenheit ist, sondern in Abhängigkeit vom jeweiligen Übersetzungsvolumen in regelmäßigen Abständen erfolgen sollte. Nur mit laufend aktualisierten Übersetzungs-Engines können SMÜ-Systeme mit der inhaltlichen Entwicklung der Ausgangstexte Schritt halten.

Eine Maschinenübersetzungslösung innerhalb von Tagen oder gar Stunden realisieren

Wie viele andere SaaS-Lösungen kann die Einrichtung eines DIY-SMÜ-Systems in der Regel in wesentlich kürzerer Zeit bewerkstelligt werden als die Implementierung einer traditionellen statistischen Maschinenübersetzungslösung. Denn bei DIY-SMÜ-Systemen kann die Einführung in nur drei einfachen Schritten erfolgen: 1. Einrichten ei-

nes Benutzerkontos; 2. Hochladen der zweisprachigen und einsprachigen Trainingskorpora; 3. Warten, bis das Training der Übersetzungsengines abgeschlossen ist. Je nach Dienstleister, Anzahl der zu erstellenden Übersetzungs-Engines und Größe der Korpora kann die kundenspezifische Anpassung eines SMÜ-Systems von weniger als einer Stunde bis mehrere Tage in Anspruch nehmen.

In der Regel müssen bei der Einführung eines DIY-SMÜ-Systems weder größeren Veränderungen an der IT-Infrastruktur des Kundenunternehmens vorgenommen noch langfristige Vertragsverpflichtungen eingegangen werden. Es ist in diesem Zusammenhang auch erwähnenswert, dass viele Anbieter von DIY- SMÜ-Lösungen Interessenten den kostenlosen Test ihres Systems ermöglichen.

Tipps für die erfolgreiche Implementierung Ihres eigenen DIY-SMÜ-Systems

Pflegen Sie Ihre Sprachdaten

Bei der statistischen Maschinenübersetzung ist ein großer Bestand an Translation-Memorys unverzichtbar, da die Übersetzungs-Engines auf der Grundlage dieser Translation-Memorys und eventuell vorhandener Glossare kundenspezifisch angepasst werden. Dabei gilt das GiGo-Prinzip: Garbage in, garbage out! Um mit einem SMÜ-System möglichst hochwertige Übersetzungen erstellen zu können und damit den Aufwand für das Post-Editing zu minimieren, sollten nur aktualisierte (Stichwort: Terminologie!) und auf Fehlerfreiheit geprüfte Translation-Memorys für das Training von Übersetzungs-Engines verwendet werden.

Investieren Sie in Schulung

Wie bereits erwähnt, braucht man keinen Doktorgrad in Computerlinguistik, um ein cloudbasiertes DIY-SMÜ-System zu bedienen. Tatsächlich müssen die Anwender derartiger Systeme nicht einmal Übersetzungsexperten sein, um die Grundfunktionen (Anpassung und Übersetzung) dieser Systeme zu nutzen.

Wer allerdings den maximalen Nutzen aus einem DIY-SMÜ-System ziehen will und danach strebt, die Effizienz des SMÜ-Systems und sämtlicher damit verbundener Prozesse zu optimieren, sollte für das Übersetzungsmanagement entsprechend geschult sein, um festzulegen, welche Projekte ausschließlich mit SMÜ (und mit welcher Engine) übersetzt werden, welche maschinell mit anschließendem Post-Editing übersetzt werden und welche den traditionellen Humanübersetzungsprozess durchlaufen sollen.

Wo seitens des Auftraggebers der Wunsch nach der Nutzung von Maschinenübersetzung besteht, ist es auf jeden Fall hilfreich, darauf hinzuweisen, dass in diesem Fall die Ausgangstexte möglichst bereits übersetzungsfreundlich erstellt sein sollten (auch hierzu gibt es entsprechende Softwareunterstützung). Durchgängiger Stil und konsistente Verwendung korrekter Terminologie im Ausgangsdokument trägt bei jedem Übersetzungsprojekt zur Steigerung der Übersetzungsqualität bei. Für die Bearbeitung mit einem statistischen Maschinenübersetzungssystem ist stilistische und terminologische Konsistenz jedoch besonders wichtig.

Halten Sie die Erwartungen realistisch

Do-it-yourself-SMÜ ist eine interessante, relativ neue Technologie, die hochwertige automatische Übersetzungen einem viel breiteren Nutzerkreis zugänglich macht.

Drei Do-it-yourself-SMÜ-Anbieter für den leichten Einstieg

Microsoft Translator Hub

Wie Google das Translator Toolkit ins Leben gerufen hat, um die Qualität des kostenlosen Google Übersetzers zu verbessern, so ist es das erklärte Ziel von Microsoft, mit dem Microsoft Translator Hub die Genauigkeit des maschinellen Bing Übersetzers zu steigern. Im Unterschied zum Google Translator Toolkit ist der Microsoft Translator Hub ein DIY-SMÜ-Dienst, mit dem Benutzer ihre eigenen maßgeschneiderten MÜ-Engines erstellen können. Die Benutzung des Microsoft Translator Hub ist bis zwei Millionen Zeichen pro Monat kostenlos.
hub.microsofttranslator.com

KantanMT

KantanMT mit Sitz in Dublin/Irland ist einer der neuen kommerziellen DIY-MÜ-Dienste, die im Laufe der letzten Jahre ihre Pforten geöffnet haben. Im Unterschied zu anderen MÜ-Dienstleistern verlangt KantanMT keine gesonderten Gebühren für die erstmalige Einrichtung von kundenspezifischen MÜ-Engines, sondern rechnet monatlich nach dem in Anspruch genommenen MÜ-Engine-Kontingent ab, was den Einstieg in diese Technologie relativ günstig macht. KantanMT bietet Interessenten eine kostenlose Testphase von 14 Tagen an.
www.kantanmt.com

SmartMate

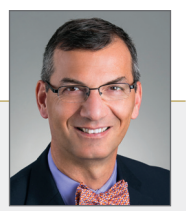
SmartMate ist mehr als nur ein DIY-SMÜ-Dienst, denn bei diesem Onlines-Service handelt es sich um eine komplette cloudbasierte Übersetzungsmanagementlösung einschließlich Online-Editor und Terminologiemanagementfunktionalität. Der monatliche Subskriptionspreis richtet sich nach der Anzahl der übersetzten Wörter und der Anzahl der Benutzer der Editierumgebung. Wie KantanMT bietet auch SmartMate eine kostenlose Testphase an – allerdings nur für die Dauer von fünf Tagen. SmartMate wird von Capita Translation and Interpretation (London) betrieben.
www.smartmate.co

Aber wer denkt, dass DIY-SMÜ die Humanübersetzung ersetzt, liegt falsch. Auch wenn die statistische Maschinenübersetzung sehr leistungsfähig ist, hat sie grundsätzlich Grenzen: Beinhalten die zu übersetzenden Texte beispielsweise Satzstrukturen oder Terminologie, die nicht in den Trainingskorpora enthalten sind – was der Normalfall ist, wenn die Autoren keine kontrollierte Sprache verwenden – sind Übersetzungsfehler unvermeidlich.

Andererseits sind von DIY-SMÜ-Systemen erzeugte Übersetzungen immer dann eine gute Lösung, wenn Qualität nicht im Vordergrund steht und sonst aus Kosten- oder Zeitgründen nur eine Alternative zur Verfügung steht: nämlich gar keine Übersetzung. Und, um es noch einmal und ganz betont zu sagen: Maschinelle Übersetzungen, die von einem kundenspezifisch angepassten DIY-SMÜ-System erstellt werden, eignen sich hervorragend für das Post-Editing durch entsprechend geschultes Personal.

Zusammenfassung

Die statistische Maschinenübersetzung mit DIY-Diensten ist eine vielversprechende Technologie – insbesondere für alle, die bereits erste Erfahrungen mit maschineller Übersetzung gesammelt haben und die Qualität von maschinell erstellten Übersetzungen verbessern möchten. Cloudbasierte DIY-SMÜ-Dienste erleichtern nicht nur das Erstellen kundenspezifischer Maschinenübersetzungen, sondern machen hochwertige Maschinenübersetzung auch erschwinglicher. Mit DIY-SMÜ-Diensten steht die momentan leistungsfähigste Maschinenübersetzungstechnologie nun auch Kleinunternehmen oder auf Post-Editing spezialisierten freiberuflichen Übersetzern zur Verfügung.



Uwe Muegge
Uwe Muegge hat mehr als 15 Jahre Erfahrung in der internationalen Übersetzungs- und Lokalisierungsbranche und war sowohl auf Anbieter- als auch auf Kundenseite in Führungsfunktionen tätig. Zurzeit ist er bei Z-Axis Tech Solutions, einem Anbieter von Sprach-, Consulting- und IT-Dienstleistungen mit Sitz in San José/USA, für maschinelle Übersetzung, kontrollierte Sprachen und Terminologiemanagement zuständig.
info [at] zaxistech.com
www.zaxistech.com

Deutsche, an den MDÜ-Heftschwerpunkt angepasste und überarbeitete/aktualisierte Version des Artikels „Do-it-yourself MT: Taking (statistical) machine translation to the next level" aus tcworld Magazine Juli 2013.