Erschienen in: Stickel, Gerhard (Hrsg.): Translation and Interpretation in Europe. Contributions to the Annual Conference 2013 of EFNIL in Vilnius. - Frankfurt am Main/Berlin/Bern/Bruxelles/New York/Oxford/Wien: Lang, 2014.
S. 79-103. (Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 102)

Antonina Werthmann/Andreas Witt

Maschinelle Übersetzung – Gegenwart und Perspektiven

Abstract

Communication across all language barriers has long been a goal of humankind. In recent years, new technologies have enabled this at least partially. New approaches and different methods in the field of Machine Translation (MT) are continuously being improved, modified, and combined, as well. Significant progress has already been achieved in this area; many automatic translation tools, such as *Google Translate* and *Babelfish*, can translate not only short texts, but also complete web pages in real time. In recent years, new advances are being made in the mobile area; Google's Translate app for Android and iOS, for example, can recognize and translate words within photographs taken by the mobile device (to translate a restaurant menu, for instance). Despite this progress, a "perfect" machine translation system seems to be an impossibility because a machine translation system, however advanced, will always have some limitations. Human languages contain many irregularities and exceptions, and consequently go through a constant process of change, which is difficult to measure or to be processed automatically. This paper gives a short introduction of the state of the art of MT. It examines the following aspects: types of MT, the most conventional and widely developed approaches, and also the advantages and disadvantages of these different paradigms.

1. Allgemeines

Im Herbst 2013 widmete das deutsche Nachrichtenmagazin "Der Spiegel" den zweiseitigen Artikel "Lost in Translation" dem Thema Maschinelles Übersetzen. Im Mittelpunkt des Berichts stand die Tätigkeit der Übersetzungsabteilung von Google Inc., die von Franz Josef Och geleitet wird. Es hieß in dem Beitrag, dass der kostenlose Textübersetzungsservice von Google bereits 2012 im Internet rund 200 Millionen Mal genutzt wurde, doch Franz Josef Och und sein Translate-Team strebten nach Höherem und hätten eine App entwickelt, die das Mobiltelefon zu einer sprechenden Übersetzungsmaschine mache. Obwohl die Einsatzmöglichkeiten und die Leistung der mobilen App noch begrenzt seien, habe das Translate-Team große Ambitionen: Das babylonische Sprachgewirr der Menschheit mit den Mitteln der Technik endlich zu beheben und alle Sprachbarrieren aufzulösen. Nun sind Google nicht die ersten, die Mobiltelefone und Sprachtechnologie zusammenbringen. So hat Apple bereits 2011 sein iPhone 4s mit dem Siri-System zur Spracherkennung

ausgeliefert. Aber trotz dieser Fortschritte bleiben der sprachgesteuerte Zugang zu Maschinen und insbesondere die automatische Übersetzung von einer Sprache in eine andere nach wie vor ein Problem, das lange noch nicht als gelöst bezeichnet werden kann.

Derzeit gibt es weltweit ca. 6.500–7.000 verschiedene Sprachen (vgl. Haspelmath 2008), die in ca. 200 Ländern[1] gesprochen werden. Allein in Europa existieren 37 Nationalsprachen (Pan 2008), von denen in der Europäischen Union derzeit 24 als Amtssprachen gelten. Hinzu kommen noch über 60 Regional- oder Minderheitensprachen.[2]

Im Zeitalter der Globalisierung wird es für die Menschen immer wichtiger mit mehreren Sprachen umgehen zu können. Es wäre daher sehr hilfreich, Geräte zur Verfügung zu haben, mit denen jede Sprache in jede gewünschte andere Sprache übersetzt werden könnte. Wenn ein solches Gerät in der Lage wäre, zumindest die 10 meistgesprochenen Sprachen ineinander zu übersetzen, wäre die Kommunikation bereits zwischen der Hälfte der Erdbevölkerung problemlos möglich.[3] Es ist deswegen nicht verwunderlich, dass Fortschritte auf dem Gebiet der Maschinellen Übersetzung (MÜ) nicht nur für die Wissenschaft, sondern auch für die Wirtschaft, die Politik, aber auch das Militär höchst erstrebenswert waren und sind.

Bereits in den 1940er Jahren wurden die ersten Versuche unternommen, Systeme zu bauen, die automatisch die Übersetzung natürlicher Sprache übernehmen können. Die Fortschritte in diesem Gebiet erfolgten sehr langsam und für lange Zeit waren die Ergebnisse der MÜ absolut inakzeptabel. Heute, im Jahr 2014, gibt es aber eine Vielzahl von Systemen für die MÜ, die in bestimmten Domänen und für gewisse Anwendungen durchaus gute Ergebnisse liefern. Sie nutzen verschiedene Techniken und Methoden, z.B. viele Systeme arbeiten mit regelbasierten oder statistischen Methoden, wobei zunehmend auch Mischformen Verwendung finden. Beim Aufbau der ersten MÜ-Systeme wurden vor allem regelbasierte Ansätze genutzt. Seit den Nullerjahren gewinnen aber

---

1 Von den Vereinten Nationen werden 193 Staaten anerkannt. Hinzu kommen noch ein paar weitere Territorien, bei denen der Status *Staat* umstritten ist und die nicht Mitglieder in der UN sind.

2 Nach Angaben der Föderalistischen Union Europäischer Volksgruppen (FUEN), zuletzt abgerufen am 22.01.2014.

3 "Die Sprachen sind sehr ungleich über die Sprecher verteilt: Etwa die Hälfte aller Menschen sprechen eine der zehn meistgesprochenen Sprachen (Chinesisch, Spanisch, Englisch, Hindi/Urdu, Bengali, Arabisch, Portugiesisch, Russisch, Japanisch, Panjabi)." (Haspelmath 2008, 140).

statistikbasierte Ansätze eine immer größere Bedeutung. Sie finden sich z.B. in den Systemen GIZA++, SRILM, Moses und Thot (vgl. Koehn 2010). Mit *Google Translate*, das auch statistik basiert arbeitet, kann man derzeit Texte zwischen 71 Sprachen übersetzen, wobei die Qualität der Übersetzung stark von den Sprachpaaren abhängt, aber auch die Textsorte und die Komplexität der Satzstrukturen haben einen starken Einfluss auf die Übersetzungsgüte: Je ähnlicher die Sprachen einander sind und je weniger komplex die Sätze der Ausgangssprache sind, desto besser ist meist die Qualität der Übersetzung.

Ein MÜ-System kann sowohl als eine selbstständige Anwendung zur automatischen Übertragung eines geschriebenen Textes einer Ausgangssprache in eine Zielsprache wie auch als Komponente eines anderen sprachtechnologischen Systems eingesetzt werden. Als Teilsystem können MÜ-Systeme z.B. beim Maschinellen Dolmetschen (MD) angewendet werden, nachdem ein System zur automatischen Spracherkennung die gesprochenen Daten verarbeitet und aufbereitet hat. Die Aufgabe der MÜ wird jedoch beim MD verkompliziert, da zusätzliche Informationen der gesprochenen Sprache, wie Intonation, Pausen, Füllwörter etc., im Übersetzungsprozess berücksichtigt werden müssen. Ein Beispiel für das MD war das Forschungsprojekt "Verbmobil" (1993–2000) zur sprecherunabhängigen maschinellen Übersetzung der Spontansprache zwischen Deutsch, Japanisch und Englisch (siehe Wahlster 2000). Der Forschungsprototyp von *Verbmobil*, der im Jahr 2001 mit dem Deutschen Zukunftspreis ausgezeichnet wurde, wurde zwar nie vermarktet, lieferte aber neue wichtige Erkenntnisse für dieses Gebiet.[4]

Trotz der großen Fortschritte, die in diesem Bereich zu verzeichnen sind, ist es jedoch noch ein weiter Weg, bis beliebige Ausgangstexte in eine beliebige Zielsprache übersetzt werden können. Es ist derzeit auch nicht abschätzbar, ob dies jemals möglich sein wird. Nichtsdestotrotz steigt die Nachfrage nach MÜ-Anwendungen, die in sehr unterschiedlichen Bereichen eingesetzt werden (vgl. Krenz/Ramlow 2008), stetig. Ihre schnelle und relativ kostengünstige Verfügbarkeit lässt die Menschen oft über die Mängel bezüglich der Genauigkeit und Zuverlässigkeit hinwegsehen.[5]

---

4 Nebenbei war Verbmobil auch ein Forschungsprojekt, das seine Mitarbeiterinnen und Mitarbeiter hervorragend auf die akademische und außerakademische Forschung vorbereitete. Auch der oben erwähnte Leiter der *Google Translate* Teams war in seiner Promotionszeit mit dem Projekt befasst.

5 Die Alternative für die Nutzer/innen ist ja meist nicht eine professionelle menschliche Übersetzung, sondern schlicht: keine Übersetzung.

2. Abgrenzung: Maschinelle Übersetzung vs. maschinengestützte Übersetzung

Zu Beginn ist es wichtig, zwei Begriffe voneinander zu unterscheiden: maschinelle Übersetzung und maschinengestützte Übersetzung (engl. *Computer Aided Translation*, abgekürzt CAT):

- MÜ: Die Voraussetzung einer MÜ ist, dass der Übersetzungsprozess ohne jede menschliche Hilfe abläuft und qualitativ hochwertige Übersetzungen liefert, d.h. die MÜ sollte sehr nah an die Übersetzungsleistung eines menschlichen Übersetzers herankommen.
- CAT: Eine maschinengestützte Übersetzung liegt dann vor, wenn der eigentliche Übersetzungsprozess eines menschlichen Übersetzers von verschiedenen Computerprogrammen, wie elektronischen Wörterbüchern, Terminologiedatenbanken und insbesondere Translation-Memory-Systemen (TMS), unterstützt wird.

Bei der TMS-gestützten Übersetzung werden Sätze und Satzfragmente, die meist von professionellen Übersetzer/innen erstellt wurden, zusammen mit ihren Übersetzungsäquivalenten in einer Datenbank gespeichert. Sollen neue Sätze übersetzt werden, werden diese mit den bereits im System gespeicherten Sätzen verglichen. Finden sich die neuen Sätze bereits in ähnlicher oder sogar identischer Form im System, werden dem Übersetzenden die alten Übertragungen präsentiert, woraufhin man entscheiden kann, ob die frühere Übersetzung auch in dem neuen Kontext passt. Falls ja, kann man die alte Übersetzung wiederverwenden, anderenfalls muss eine neue Übersetzung erstellt werden, die selbstverständlich dann im TMS gespeichert wird. Die Wiederverwendung der repetitiven Inhalte vereinfacht und beschleunigt den Übersetzungsprozess insbesondere bei domänenspezifischen Texten. Außerdem erhöht TMS die Konsistenz und die Qualität der Übersetzung (mehr zu TMS siehe Somers 2003). Einige Beispiele für TM-Systeme sind *Déjà Vu*, *OmegaT*, *SDL Trados Studio* etc.

3. Probleme in der MÜ

Nach den Ursachen der schlechten Realisierbarkeit eines MÜ-Systems muss nicht lange gesucht werden. Die Probleme liegen in erster Linie nicht an den angewendeten Technologien, sondern vor allem an der Komplexität der natürlichen Sprache und ihrer Verarbeitung. Im Laufe der Zeit wurde es immer klarer, dass

die linguistischen Probleme im MÜ-Bereich weitaus größer waren, als zuerst angenommen. Die automatische Bearbeitung bereits einer Sprache ist komplex und aufwändig: Mehrere Sprachen vervielfachen die Komplexität und damit auch das Fehlerpotenzial bei der MÜ.

Problematisch ist z.B., dass die Wörter einer Sprache nicht einfach einzeln in eine andere Sprache übertragen und dann aneinandergereiht werden können: Beim Übersetzungsprozess müssen aus den Wörtern Phrasen gebildet werden, die ihrerseits in die Sätze hierarchisch eingefügt werden, die wiederum den Text bilden, der erneut durch einen mehr oder weniger hohen Grad an Variabilität und Flexibilität der Abfolge der Wörter gekennzeichnet ist. Hinzu kommt, dass viele Wörter in Abhängigkeit von unterschiedlichen Kontexten verschiedene Bedeutungen haben können. Da die lexikalischen Ambiguitäten in den verschiedenen Sprachen unterschiedlich sind, muss das jeweilige Übersetzungsäquivalent für die passende Bedeutung eines Lexems gefunden werden.

Ein weiteres Problem für die MÜ besteht darin, dass natürliche Sprachen zwar Regeln haben, die in Grammatiken erfasst sind, das Regelwerk jedoch viele Variationen beinhaltet, d.h. Ausnahmen, spezifische Bedeutungen etc., deren Berücksichtigung und Implementierung in Computersystemen äußerst komplex sind.

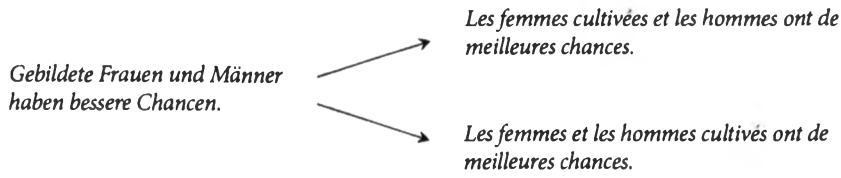
Eines der zentralen und wichtigen Probleme der MÜ ist aber die bereits erwähnte Mehrdeutigkeit bzw. Ambiguität, die in verschiedenen Sprachebenen auftreten kann. Exemplarisch werden hier zwei Arten der Mehrdeutigkeit veranschaulicht:

Lexikalische Mehrdeutigkeit von zu übersetzenden Wörtern: Lexikalische Mehrdeutigkeit liegt dann vor, wenn eine Wortform für zwei oder mehrere Begriffe steht, z.B.

dt. *die Bank* → engl. *bench*
dt. *die Bank* → engl. *bank*
dt. die *Mutter* → engl. *mother*
dt. die *Mutter* → engl. *nut*

Ein weiteres Problem stellt ***strukturelle Mehrdeutigkeit*** dar, die immer dann entsteht, wenn z.B. ein Element nach dem Regelwerk einer Ausgangsprache von verschiedenen Elementen regiert werden kann und dies dem Syntaxmuster der Zielsprache nicht entspricht.

Ein Beispiel aus Eberle (2008) zeigt einen deutschen Satz, der zwei Lesarten hat und ins Französische auf zwei verschiedene Weisen übersetzt werden kann:



Der deutsche Satz wurde ins Französische von mehreren MÜ-Systemen exemplarisch übersetzt, die in der folgenden Tabelle aufgelistet sind:

On-line Übersetzer	Ansatz	Web-Adresse
Babelfish	Basiert auf einer ältere Version des Systems SYSTRAN, das in regelbasiert war. In letzteren Jahren wird Systran zu einem hybriden Ansatz weiterentwickelt.	http://www.babelfish.com/
Bing Translator	Statistisches MS-System mit sprachspezifischen Regelkomponenten für das Zerlegen und Zusammensetzen von Sätzen.	http://www.bing.com/translator/
Google Translate	Rein statistisches MÜ-System	http://translate.google.ch/
Pons	Abhängig von der Auswahl der Sprache werden Übersetzungssysteme von Lingenio[6] oder Bing benutzt.	http://en.pons.eu/text-translation
Promt	Hybrides MÜ-System: regelbasiert und statistisch	http://www.online-translator.com/

An den unterschiedlichen, folgend aufgelisteten Übersetzungen sieht man, dass die Mehrdeutigkeit des deutschen Ausgangssatzes die Übersetzung ins Französische erschwert: Im deutschen Satz ist es nicht eindeutig, ob sich die Attribution *gebildet* nur auf das Nomen *Frauen* oder auf die Koordination *Männer und Frauen* bezieht. Bei der Übersetzung ins Französische muss aufgrund der vom Deutschen abweichenden Kongruenzregel zwischen den Bezugswörtern disambiguiert werden:

Babelfish: *"Hommes et femmes instruites ont une meilleure chance."*
Bing: *"Hommes et femmes instruites ont une meilleure chance."*
Google Übersetzter: *"Les femmes et les hommes instruits ont de meilleures opportunités."*
Pons: *"Des femmes et des hommes cultivés ont de meilleures chances."*
Promt: *"Les femmes formées et les hommes ont les meilleures chances."*

6 Lingenio ist ein regelbasiertes MÜ-System, das der Transfer-Ansatz benutzt und nur kommerziell angeboten wird.

Eine besondere Schwierigkeit bereiten der MÜ solche Sprachpaare, die große strukturelle Unterschiede in ihren Grammatiken aufweisen. Ihre automatische Übersetzung bereitet naturgemäß größere Probleme als die von Sprachpaaren, die ähnliche Satzstrukturen besitzen. Beispielsweise gelingen die Übersetzungen mit *Google Translate* zwischen dem Englischen und Spanischen wesentlich besser als z.B. zwischen dem Englischen und Japanischen (vgl. Schulz 2013). Auch das Deutsche besitzt einige grammatische Strukturen, wie die Satzstruktur mit linken und rechten Satzklammern, für die eine parallele Entsprechung nur in sehr wenigen anderen Sprachen zu finden ist. Einen besonderen Platz bei der maschinellen Übersetzung des Deutschen nehmen aber auch die Verben mit abtrennbaren Präfixen ein. Ihre falsche Erkennung, die bei der statistischen MÜ oft auftritt, kann die Bedeutung des Satzes stark verfälschen. Die Übersetzung eines deutschen Satzes mit verschiedenen Online-MÜ-Systemen exemplifiziert dies: Das Verb *aufstehen, welches ein abtrennbares Präfix hat*, wird je nach Position im Satz unterschiedlich übersetzt, z.B.:

Ein Bäcker steht gewöhnlich schon um 2:30 Uhr auf.

Babelfish:	*A Baker is usually around 2:30.*
Bing:	*A Baker is usually around 2:30.*
Google Übersetzer:	*A baker is usually already on at 2:30 clock.*
Poins:	*A baker gets up usually already at 2:30 o'clock.*
Promt:	*A baker normally already gets up at 2:30 o'clock.*

[..., weil] ein Bäcker gewöhnlich schon um 2:30 Uhr aufsteht.

Babelfish:	*"[because] Baker is usually around 2:30 on."*
Bing:	*"[because] Baker is usually around 2:30 on."*
Google Übersetzer:	*"[because] a baker gets up usually already at 2:30 clock."*
Poins:	*"[because] a baker gets up usually already at 2:30 o'clock."*
Promt:	*"[because] a baker normally already gets up at 2:30 o'clock."*

Weiterhin zeichnet sich z.B. die deutsche Sprache durch eine sehr produktive Verwendung von Komposita aus, die dann in anderen Sprachen nur mit komplexen Mehrwortkombinationen übersetzt werden können, z.B. siehe Tabelle 1.

Auch die korrekte Übersetzung von anaphorischen Wörtern bereitet der MÜ Probleme. Die Ursache hierfür ist, dass bei der MÜ die Analyse der Ausgangssprache oft auf die Wort- und Phrasenebene beschränkt bleibt, was oft für die Anaphernresolution nicht ausreichend ist, z.B.:

Das Mädchen, das auf seine Mutter wartet, spielt mit seinem Hamster.

Google Übersetzer:	*"The girl who waits for his mother, playing with his hamster."*
Poins:	*"The girl who waits for his mother plays with his hamster."*
Bing:	*"The girl, who is waiting for his mother, plays with his Hamster."*
Promt:	*"The girl who waits for his mother plays with his hamster."*

Tab. 1: Beispiele von deutschen Komposita mit entsprechenden Übersetzungen ins Englische und Französische

dt.	engl.	franz.
Kartoffelbrei	mashed potatoes	purée de pommes de terre
Energiewende	energy transition	transition énergétique
Textverarbeitungssystem	text processing system	système de traitement de texte
Dienstleistungsabkommen	service agreement	accord sur les prestations de services
Rindfleischetikettierungs-überwachungsaufgaben-übertragungsgesetz[7]	beef labeling regulation and delegation of supervision law[8]	loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine[9]

Schließlich hängt die Qualität der MÜ von dem zu übersetzenden Text selbst ab: Zahlreiche Fehler hinsichtlich der Rechtschreibung und Grammatik sowie lange komplexe Sätze führen bei den meisten Systemen zu Übersetzungsfehlern.

4. Ansätze und Methoden der Maschinellen Übersetzung

Die MÜ liegt an der Schnittstelle von Linguistik, Computerlinguistik, Übersetzungswissenschaften und Informatik. Diese Interdisziplinarität erklärt auch die Tatsache, dass eine Vielzahl von verschiedenen Methoden und Ansätzen zur MÜ entwickelt wurde, die hauptsächlich in regelbasierte und korpusbasierte Strategien (vgl. Tripathi/Sarkhel 2010) unterteilt werden können. Während bei den regelbasierten Übersetzungsstrategien die Übersetzung vor allem auf manuell erstellten, mehrsprachigen Wörterbüchern und grammatischen Regeln beruht,

7 Das Wort stammt aus der Rechts- und Verwaltungssprache und galt für längere Zeit als Beleg für das angeblich längste Wort der deutschen Sprache. In 1999 wurde es als eines der Wörter des Jahres nominiert (Bär 2003, 247). In 2013 wurde das aus Mecklenburg-Vorpommern stammende Gesetz mit dem vollen Namen „Rinderkennzeichnungs- und Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz" nach Beschluss des Schweriner Landtags allerdings aufgehoben (Spiegel Online 2013).

8 Übersetzt mit Deutsch-Englisches WÖRTERBUCH: www.deutschenglischeswoerterbuch.com.

9 Vgl. Andries (2008, 347).

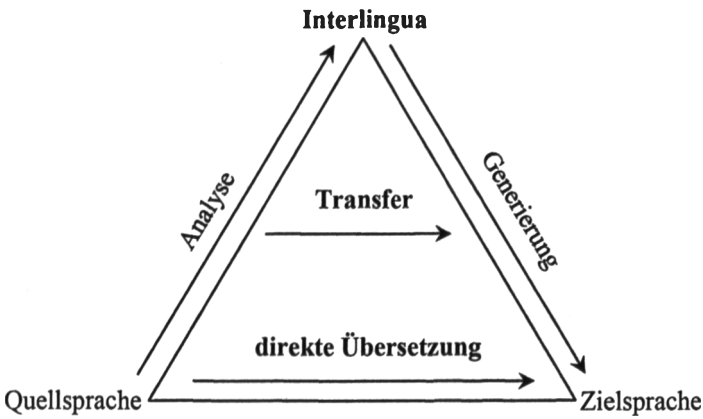
die aus detaillierter linguistischer Analyse und Modellierung sowohl der Quell- als auch der Zielsprache resultieren, stützt sich der Übersetzungsprozess der statistischen Strategien vor allem auf die Analyse einer großen Menge von ein- und mehrsprachigen Korpora, mit deren Hilfe ein System die Wahrscheinlichkeiten möglicher "Übersetzungen" aus den Wortentsprechungen und -mustern berechnet und die wahrscheinlichste Übersetzung auswählt.

4.1 Regelbasierte Ansätze

Die regelbasierten Ansätze (engl. *Rule-Based Machine Translation*, abgekürzt RBMT) gehören zu den klassischen Ansätzen der MÜ (Stein 2009). Die Sprache wird auf verschiedenen linguistischen Ebenen analysiert, z.B. werden auf der morphologischen Ebene die Kasusflexionen der Nomen und auf der syntaktischen Ebene die Wortstellungsinformationen ausgewertet. Die Ausgangssprache wird dann in eine – je nach MÜ-Strategie mehr oder weniger abstrakte – Repräsentation überführt, auf der dann Übersetzungsregeln angewendet werden.

Die regelbasierten Ansätze werden in weitere drei MÜ-Strategien unterteilt: direkte Übersetzung, Transfer und Interlingua. Sie unterscheiden sich voneinander vor allen in folgenden Ansatzpunkten: Zahl der Sprachpaare, die bei der Übersetzung berücksichtigt werden; Art und Weise, wie die gewonnenen Informationen aus der Quellsprache analysiert werden; und anschließend wie diese für die Übersetzung abstrakt repräsentiert werden. In Vauquois' Dreieck werden diese unterschiedlichen Übersetzungsstrategien (siehe Abschnitte 4.1.1 und 4.1.2) in einem Architekturschema wie in Abbildung 1 dargestellt.

Abb. 1: MÜ Vauquois' Dreieck der MÜ nach Hutchins/Somers (1992, 107)



Während sich die direkte Übersetzung mit ihrer minimalen Analyse auf der untersten Ebene des Architekturschemas befindet, liegt die Interlingua-Übersetzung, die versucht alle Einzelsprachen in einer gemeinsamen sprachunabhängigen Repräsentation abzubilden, an der Spitze des Vauquois-Dreiecks. In der Mitte des Architekturschemas ist der Transfer-Ansatz zu finden, der als eine Art Mittelweg zwischen der direkten Übersetzung und dem Interlingua-Ansatz angesehen werden kann. Hierbei wird zwar ebenfalls eine Zwischenrepräsentation erzeugt, jedoch ist diese im Unterschied zu einer Interlingua sprachabhängig. Das bedeutet z.B. dass Mehrdeutigkeiten, die in der gleichen Form in Ausgangs- und Zielsprache auftreten, nicht disambiguiert werden müssen. Transfer und Interlingua zählen sich zur indirekten Übersetzung. Verglichen mit der direkten Übersetzung sind beide komplexer und werden auch als Strategien der 2. Generation bezeichnet (Eberle 2008).

4.1.1 Direkte Übersetzung

Die direkte Übersetzung ist der älteste und einfachste Ansatz der MÜ. Es handelt es sich hier um einen mit vielen Einschränkungen durchlaufenden Wort-zu-Wort-Übersetzungsprozess, bei dem keine oder nur eine minimale strukturelle und semantische Analyse der Quellsprache geleistet wird (Stein 2009).

Während des Übersetzungsprozesses wird im ersten Schritt der Quelltext in Wörter bzw. Phrasen segmentiert, die dann im nächsten Schritt mit den Einträgen des bilingualen Lexikons bzw. Wörterbuches verglichen werden. Es wird nach möglichst genauen Übereinstimmungen gesucht, mit denen die Wörter der Ausgangssprache mit Entsprechungen aus der Zielsprache ersetzt werden.

Die Qualität der direkten Übersetzung ist nicht besonders hoch, fällt aber natürlich von Sprach- zu Sprachpaar unterschiedlich aus. Je ähnlicher die zu übersetzenden Sprachen hinsichtlich ihrer grammatischen und lexikalischen Strukturen sind, desto bessere Ergebnisse bei der Übersetzung können erzielt werden.

Die Realisierung des Ansatzes ist zwar relativ einfach, aber trotzdem aufwändig. Es kann immer nur ein Sprachpaar berücksichtigt werden: Bei n Sprachen, die ineinander übertragen werden sollen, benötigt der Ansatz $n \times (n - 1)$ Regelsätze. Aufwändig ist es vor allem, große Wörterlisten mit allen möglichen Wortformen eines Eintrags herzustellen. Für jede weitere Sprache müssen dann weitere Wörterlisten erstellt werden. Die Nutzung eines zusätzlichen morphologischen Analyseschritts kann die Menge der gespeicherten Einträge im Wörterbuch reduzieren und das System verbessern,

indem – z.B. nach der Textsegmentierung – die Wörter auf ihre Stammformen zurückgeführt werden, um alle Variationen eines Wortes auf einen Eintrag zu reduzieren. Die folgende Abbildung zeigt die Funktionsweise dieses ältesten und einfachsten Ansatzes der MÜ:

Abb. 2: Schema für direktes MÜ-System (Carstensen et al. 2004, 566)

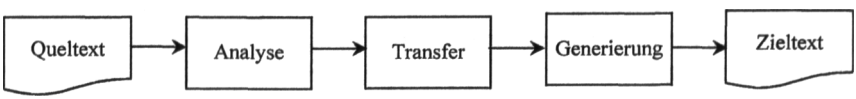
4.1.2 Indirekte Übersetzung

Transfer

Beim Transfer nimmt der Aufwand des Übersetzungsprozesses zu: Im Unterschied zur direkten Übersetzung läuft er nicht auf der Wortebene ab, sondern auf der Satzebene. Die Wörter werden nicht einzeln und isoliert, sondern als Teil der syntaktischen Strukturen betrachtet. Der Übersetzungsprozess vollzieht sich in drei Phasen: Analyse, Transfer und Generierung (vgl. Jurafsky/ Martin 2009).

Im Analyseschritt wird der Quelltext geparst, in Sätze und Konstituentenstrukturen segmentiert und schließlich analysiert. Die Sätze des Quelltexts werden nach den durchgeführten morphologischen, syntaktischen oder weitergehenden Analysen in eine abstrakte Struktur übertragen. Im nächsten Transfer schritt werden die gewonnenen abstrakten Strukturen in die abstrakten Strukturen der Zielsprache transferiert. Schließlich wird im Generierungsschritt aus der vom Transfer erzeugten abstrakten Repräsentation von Strukturen der Zielsprache wieder eine natürlichsprachliche Ausgabe von Sätzen der Zielsprache generiert, siehe Abbildung 3.

Abb. 3: Schema für Transferbasierte Übersetzung (Carstensen et al. 2010, 646)



Der Aufwand beim Transfer-Ansatz ist mit dem von direkter Übersetzung vergleichbar: Für jedes Sprachpaar in einem multilingualen Übersetzungssystem muss die Transferkomponente neu erstellt werden, wodurch bei *n* Sprachen

man Minimum $n \times (n - 1)$ Transfermodule braucht (vgl. Mishra 2010), weil die Repräsentationen nicht von universeller Natur, sondern sprachpaar- und sprachrichtungsabhängig sind (Schäfer 2002).

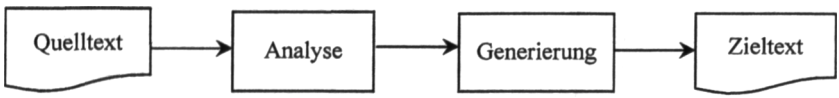
Die Vorteile des Transfer-Ansatzes gegenüber der direkten Übersetzung bestehen darin, dass hierbei die syntaktische und semantische Analyse der Sprache in den Übersetzungsprozess einbezogen wird, wodurch die Ergebnisse deutlich verbessert werden können (Stein 2009).

Interlingua

Beim Interlingua-Ansatz ist der Aufwand für die Analyse der Ausgangssprache im Vergleich zum Transfer-Ansatz und zur direkten Übersetzung am größten. Die Realisierung dieses Ansatzes besteht im Wesentlichen aus zwei Schritten: Analyse und Generierung.

Ähnlich wie beim Transfer erfolgt bei der Interlingua zuerst die Analyse der Daten: Die Quellsprache wird geparst, segmentiert und analysiert, wonach eine – als Interlingua bezeichnete – Zwischenrepräsentation des Ausgangssatzes oder des Ausgangstextes erzeugt wird. Diese ist sprachpaarunabhängig und wird in einer formalen Sprache zur Wissensmodellierung repräsentiert. Im zweiten Schritt findet die Generierung der Zielsprache aus der Zwischenrepräsentation in die Quellsprache statt, siehe Abbildung 4.

Abb. 4: Schema für Interlingua-basierte Übersetzung (Carstensen et al. 2010, 646)



Der Vor- und gleichzeitig Nachteil des Interlingua-Ansatzes ist, dass die Zwischenrepräsentation unabhängig von Quellsprache ist. Dies erschwert die Analyse der Quellsprache, da alle sprachlichen Unschärfen, Vagheiten und Ambiguitäten aufgelöst werden müssen. Andererseits kann diese Repräsentation dann für die Generierung von unterschiedlichen Zielsprachen verwendet werden. Der Weg der Übertragung der Quellsprache in die Zwischenrepräsentation, die ein Ausgangspunkt für die Generierung des Satzes in die Zielsprache bildet, macht den Ansatz dann effizient, wenn viele verschiedene Zielsprachen bedient werden sollen. Ein System, das auf diesem Verfahren basiert, lasst sich naturgemäß leichter um weitere Sprachpaare ergänzen (vgl. Ramlow 2009). Für n Sprachen werden für alle möglichen Übersetzungsrichtungen lediglich $2n$ sprachabhängige Module benötigt (vgl. Mishra 2010).

Die bereits erwähnte besondere Schwierigkeit dieses Ansatzes besteht darin, eine sprachunabhängige universale Zwischenrepräsentation mit einer formalen Sprache zu definieren, die über eigene strukturelle und lexikalische Elemente verfügt und alle möglichen Ausdrucksmöglichkeiten von allen zu übersetzenden Sprachpaaren berücksichtigt und formalisiert (vgl. Hutchins/Somers 1992; Schubeft 1995; Ramlow 2009). Bis heute wurde noch keine ideale Repräsentationssprache hierfür entwickelt (Stein 2009). Die existierenden Systeme, die diesen Ansatz nutzen, sind häufig experimentelle Systeme (Schäfer 2002; Ramlow 2009), die nur für bestimmte Domänen (wie z.B. KANT[10]) oder für Zielsprachen mit kontrolliertem Vokabular und stark vereinfachten Satzstrukturen konzipiert sind (wie KANT und DLT[11], vgl. AlAnsary 2011).

4.2 Korpusbasierte statistische MÜ

Der Nachteil der regelbasierten MÜ besteht vor allem darin, dass ihre Realisierung mit großem Aufwand der Erstellung von zweisprachigen Wörterbüchern und oft manuellem Zusammenstellen von Regelsammlungen verbunden ist. Aus diesem Grund treten seit 1989 in der MÜ korpusbasierte Systeme immer mehr in den Vordergrund, mit denen eine hohe Genauigkeit mit wenigem Aufwand erreicht werden kann (Tripathi/Sarkhel 2010).

Grundsätzlich beruht korpusbasierte MÜ auf Informationen, die mit empirischen Auswertungsmethoden aus parallelen und alignierten Textkorpora extrahiert werden. Es handelt sich dabei um zwei- oder auch mehrsprachige Korpora mit den gleichen Inhalten. Bei der sogenannten Alignierung werden die Texte zusätzlich mit Markierungen versehen, die inhaltsidentische Entsprechungen zwischen Korpora auf der Wort-, Phrasen-, Satzebene, etc. auszeichnen. Diese können sowohl manuell als auch automatisch oder halbautomatisch erstellt werden. Die manuelle Alignierung zeichnet sich dadurch aus, dass die Entsprechungen zuverlässig ermittelt werden, was die Qualität der Übersetzung deutlich verbessern kann. Allerdings ist die manuelle Alignierung nicht effizient und kann daher für große Menge von Texten, wie sie für die

10 Knowledge-based, Accurate Natural-language Translation: Das einzige Interlingua-basierte System, das in Carnegie-Melon University entwickelt wurde und kommerziell angesetzt werden konnte (AlAnsary 2011).

11 Distributed Language Translation: Ein Interlingua-basiertes System, das von 1979 bis 1990 innerhalb eines Forschungsprojekts in Utrecht, Niederlande entwickelt wurde, ist aber Prototyp geblieben.

korpusbasierte MÜ benötigt werden, kaum angewendet werden. Die manuell alignierten Daten werden im Rahmen der Evaluierung oder für das Trainingskorpus im statistischen Alignierungssystem genutzt.

Bei der automatischen Alignierung werden das Ausgangsdokument und seine Übersetzung in Übersetzungseinheiten segmentiert und die entsprechenden Übersetzungseinheiten werden dann mit Hilfe statistischer und linguistischer Algorithmen einander zugeordnet bzw. zusammengefügt. Das Verfahren ist effizient, kann jedoch Fehler und Überlappungen zwischen den Übersetzungseinheiten erzeugen. Der Grund dafür ist zum einen, dass ein Wort bzw. ein Satz der Quellsprache nicht immer einem Wort bzw. einem Satz der Zielsprache entspricht. Zum anderen kann sich die Reihenfolge der Textfragmente und Sätzen bei der Übersetzung verschieben oder ändern (vgl. Carstensen et al. 2010). Diese Unregelmäßigkeiten können dann bei einer halbautomatischen Alignierung behoben werden, indem der automatische Prozess von einem Menschen überwacht wird, der die Ergebnisse bei Bedarf nachbessert, damit möglichst viele und vor allem sinnvolle Beispiele ermittelt werden können (vgl. Kay/Röscheisen 1993).

Im Grunde verlaufen alle korpusbasierten MÜ-Ansätze nach ähnlichem Prinzip: Zuerst wird nach entsprechenden Beispielen aus dem Korpus der Quellsprache gesucht, wonach bei Erfolg die am besten übereinstimmenden zielsprachlichen Äquivalente geliefert werden. Unter den korpusbasierten Übersetzungsstrategien werden weitere Ansätze unterschieden: reinstatistik-, beispiel- und kontextbasiert (vgl. Ramlow 2009; Stein 2009; Tripathi/Sarkhel 2010). Die meisten korpusbasierten Ansätze, die heutzutage untersucht und angewendet werden, sind entweder reinstatistik- oder beispielbasiert (Kit et al. 2002).

4.2.1 Reinstatistikbasierte Übersetzung

Die rein statistisch arbeitende MÜ (engl. *Statistical Machine Translation*, abgekürzt SMT) gehört zu einem der meist verbreiteten korpusbasierten Lernverfahren, in dem sich das System auf eine Auswertung von großen parallelen und alignierten Sprachdaten stützt und berechnet, wie groß die Wahrscheinlichkeit ist, dass ein ausgangssprachlicher Term, z.B. ein Wort bzw. eine Phrase oder ein Satz, mit einem zielsprachlichen Term übersetzt wird. In die Wahrscheinlichkeitsberechnungen kommen solche Aspekte wie Worthäufigkeit, Position des Wortes im Satz, die Satzlänge, etc. Das Ergebnis ist ein im Training selbstständig erlerntes Übersetzungsmodell, das eine Liste von Übersetzungsmöglichkeiten

für jeden ausgangssprachlichen Term mit der berechneten Wahrscheinlichkeit seiner zielsprachlichen Entsprechungen enthält. In die Erstellung des Übersetzungsmodells werden weder Wörterbücher, noch linguistische Methoden oder explizites linguistisches Wissen einbezogen. Das Verfahren basiert ausschließlich auf den statistischen Wahrscheinlichkeitsberechnungen der großen bilingualen Korpora. Abhängig davon, auf welcher Ebene ein Übersetzungsmodell erstellt wird, werden für seine Erstellung die Wahrscheinlichkeiten für einzelne Wörter, Phrasen oder ganze Sätze berechnet.

Wortbasierte statistische MÜ

Wortbasierte statistische MÜ stützt sich auf die Analyse von Korpora auf der Wortebene. Die Grundidee des Modells geht auf das Projekt *Candide* von IBM zurück, das in 80er- und 90er-Jahren durchgeführt wurde (vgl. Koehn 2010, 81). Sie besteht darin, dass ein Satz nicht als Ganzes, sondern zuerst geteilt und erst dann auf Wortebene übersetzt wird. Am Beispiel des deutschen Satzes "Das Haus ist klein." soll das Übersetzungsmodell auf Wortbasis veranschaulicht werden: Bei der Übersetzung ins Englische finden sich für jedes deutsche Wort fünf verschiedene Wörter im Englischen. Nach dem Übersetzungsmodell wird die Wahrscheinlichkeit für jede Übersetzung berechnet, indem man die Zahl der einzelnen Möglichkeiten durch die Gesamtzahl aller Möglichkeiten teilt. Wahrscheinlichkeiten erhalten stets einen Wert zwischen 0 und 1.

Tab. 2: Berechnung der Übersetzungswahrscheinlichkeiten für jedes Wort im Satz "Das Haus ist klein." (Koehn 2010, 84)

das		**Haus**		**ist**		**klein**	
Übersetzung (Ü)	**Wahrscheinlichkeit (P)**	**Ü**	**P**	**Ü**	**P**	**Ü**	**P**
the	0.7	house	0.8	is	0.8	small	0.4
that	0.15	building	0.16	's	0.16	little	0.4
which	0.075	home	0.02	exists	0.02	short	0.1
who	0.05	household	0.015	has	0.015	minor	0.06
this	0.025	shell	0.005	are	0.005	petty	0.04

In der Tabelle 2 ist abzulesen, dass die wahrscheinlichste Übersetzung für das deutsche Wort *Haus* das englische *house* ist: In über 80% der Fälle wurde das erste durch das zweite ersetzt. Der Wert 0.8 für die Wahrscheinlichkeit ist ein relativer Wert, der sich auf die 10.000 eingegebenen deutsch-englischen Beispielsätze bezieht.

Nach diesem Verfahren soll einem Wort aus der Ausgangssprache immer ein Wort aus der Zielsprache entsprechen, was der Realität nicht immer entspricht. Die Beispiele in Tabelle 1 haben bereits veranschaulicht, dass es nicht selten der Fall ist, dass ein Wort nur mit mehreren Wörtern übersetzt werden kann oder umgekehrt.

Phrasenbasierte statistische MÜ

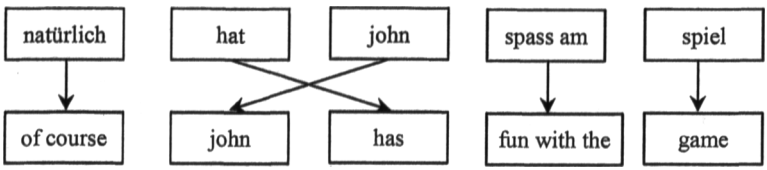
Das wortbasierte statistische MÜ-Verfahren hat einen großen Nachteil: Das Übersetzungsmodell wird mit der Annahme erstellt, dass die einzelnen Wortübersetzungen unabhängig voneinander existieren. Die umliegenden Kontextinformationen werden bei der Berechnung der wortbasierten Wahrscheinlichkeit nicht berücksichtigt, obwohl von diesen die Übersetzung eines Wortes stark abhängt.

Bessere Ergebnisse in der statistischen MÜ können die Systeme mit der Berechnung der Wahrscheinlichkeit auf der Phrasenebene erzielen, bei denen in ein Übersetzungsmodell nicht nur einzelne Wörter, sondern auch Phrasen, die aus mehreren Wörtern bestehen, als Übersetzungseinheiten einbezogen werden. Als Phrasen werden hierbei Mehrwortsequenzen bezeichnet. Es muss jedoch bemerkt werden, dass die Phrasen in der MÜ nicht linguistisch motiviert sind und nicht bzw. nur vereinzelt und zufälligerweise den syntaktischen Phrasen entsprechen[12] (vgl. Koehn 2010, 128).

Nachdem die Phrasen aus der Quellsprache in die Zielsprache übersetzt worden sind, stehen sie in der gleichen Reihenfolge wie auch in der Quellsprache. Da die Wortreihenfolgen in den Sätzen der Ausgangs- und Zielsprache jedoch stark voneinander abweichen können, werden in einem nächsten Schritt die Wörter in zielsprachlicher Reihenfolge mithilfe eines Sprachmodells neu geordnet.

Ein Beispiel für die Segmentierung und Übersetzung eines Satzes ist in Abbildung 5 dargestellt.

Abb. 5: Beispiel zur phrasenbasierten statistischen MÜ: Der Ausgangsatz ist in Phrasen segmentiert, auf der Phrasenebene übersezt und neugeordnet (Koehn 2010, 128)



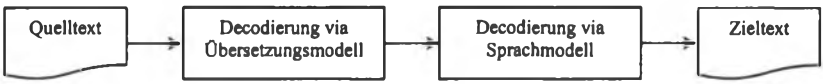
12 Nach Koehn et al. (2003, 50) wird eine syntaktische Phrase als eine Wortsequenz definiert, die in einer syntaktischen Baumstruktur als ein Teilbaum erfasst wird.

Ein Vorteil der phrasenbasierten gegenüber der wortbasierten statistischen MÜ ist, dass mehrere Wörter mit einem einzelnen übersetzt werden können und umgekehrt. Ein weiterer Vorteil ist, dass der Kontext der einzelnen Wörter in die Übersetzung einbezogen wird, wodurch die Ambiguitäten besser behandelt und gelöst werden können. In Abbildung 5 wird z.B. die deutsche Phrase *spass am* als *fun with the* übersetzt. Wenn *with* als einzelnes Wort übersetzt worden wäre, wäre es wahrscheinlicher es mit *on* oder *at* zu übersetzen, was hier zu einer falschen Übersetzung geführt hätte.

Neben dem Übersetzungsmodell gehört zu einem statistischen MÜ-System eine weitere wichtige Komponente: das sogenannte Sprachmodell. Dies hat die Aufgabe, aus den übersetzten Sequenzen die wohlgeformte Wortfolge der Zielsprache zu berechnen (Carstensen et al. 2010; Kuhlen et al. 2013). Sprachmodelle werden auf Basis von einsprachigen Korpora gebildet.

Der Übersetzungsvorgang eines reinstatistischen MÜ-Verfahrens besteht damit aus zwei Hauptschritten (siehe Abbildung 6): Im ersten Schritt, der Decodierung, werden mithilfe des Übersetzungsmodells für alle einzelnen Terme der Ausgangssprache die möglichen Übersetzungen in der Zielsprache mit den dazugehörigen Wahrscheinlichkeiten ermittelt. Im zweiten Schritt werden mithilfe des Sprachmodells der Zielsprache die auch kontextuell wahrscheinlichsten Übersetzungsalternativen ausgewählt und in die wahrscheinlichste zielsprachliche Wortreihenfolge geordnet, sodass ein möglichst grammatisch korrekter und natürlich wohlgeformter Zielsprachensatz entsteht.

Abb. 6: Schema für die statistische Übersetzung (Carstensen et al. 2010, 650)



Zu den Vorteilen der statistischen Ansätze der MÜ zählt vor allem, dass diese sowohl im Kosten- als auch im Zeitaufwand verglichen mit regelbasierten Ansätzen günstiger sind. Sie verzichten auf die mühsame manuelle Erstellung von komplizierten Grammatikregeln und auf die Auflistung ihrer Ausnahmen, die das linguistische Wissen über jede Ziel- und Quellsprache erfassen sollen. Die Erweiterung eines Systems um weitere Sprachen ist relativ einfach zu bewerkstelligen: Die meisten Algorithmen in der statistisch basierenden MÜ sind sprachunabhängig, deswegen können die darauf basierenden Systeme schnell durch Hinzunahme von neuen Daten um weitere Sprachen erweitert werden. Voraussetzung hierfür ist natürlich, dass großen Mengen von alignierten parallelen Textdaten für jedes Sprachpaar aufgebaut wurden (vgl. Koehn 2005;

Stein 2009). Schnelle Realisierung (bereits in wenigen Tagen, vgl. z.B. Oard/Och 2003), geringe Investitionskosten und hohe Funktionssicherheit machen den statistischen Ansatz effizient. Die bekannten Online-Übersetzungsysteme der Firmen *Google* und *Microsoft* sind auf einem statistischen Ansatz aufgebaut.

Die Schwäche des reinstatistischen Ansatzes liegt im Erstellungsaufwand von großen Mengen der auf geeigneter Weise aufbereiteten Textkorpora, aus denen Informationen für zuverlässige Wahrscheinlichkeitsberechnungen erhalten werden können, sodass kein expliziertes linguistisches Wissen für Regelaufbau benötigt wird. Nach Bennet/Gerber (2003) ist ein rund 1 Millionen Wörter großes Korpus von zweisprachigen Satzpaaren notwendig, damit ein Trainings-Set für ein universell einsetzbares MÜ-System erstellt werden kann.

4.2.2 Beispielbasierte MÜ

Die beispielbasierte MÜ (engl. *Example Based Machine Translation*, abgekürzt EBMT) stützt sich auf dem Abrufen und Wiederverwenden von ähnlichen oder identischen Segmenten aus bereits übersetzten Inhalten. Die Idee von beispielbasierter MÜ wurde zuerst von Makoto Nagao im Jahr 1981 vorgeschlagen, die aber erst drei Jahre später veröffentlicht wurde (vgl. Somers 2003a). Der Ansatz entstand aus Technologiekonzept eines TMSs, bei dem die Sätze mit ihren jeweiligen Übersetzungen zur weiteren Wiederverwendung in späteren Übersetzungen gespeichert werden (Brown 1996; Eberle 2008; Tripathi/Sarkhel 2010). Sie werden vor allem zur Unterstützung eines menschlichen Übersetzers eingesetzt (Stein 2009), der für die Übersetzung relevante Sätze und Satzteilen vom System auswählt, um sie wiederzuverwenden zu können. Nagaos Idee war diese ebenfalls zum Zwecke der Automatisierung der Übersetzungsprozesse zu verwenden.

Die Grundlage des beispielbasierten Ansatzes bilden bilinguale parallele Korpora mit übersetzten Textdaten, die von Fachleuten übersetzt wurden und nicht von denjenigen, die lediglich Sprachkenntnisse haben. Nur in diesem Fall kann Wissen, die nicht in irgendwelcher Form formal codiert oder repräsentiert werden, aus den zweisprachig kodierten Texten extrahiert und benutzt werden. Dieses Wissen wird mit der Übersetzung von menschlichen Übersetzern mitgeliefert, indem sie die Sätze der Quellsprache an die übersetzten Sätze der Zielsprache verknüpfen (vgl. Kit et al. 2002). Die Qualität der Übersetzung kann auch dadurch steigen, wenn die zielsprachlichen Sätze aus der gleichen Textdomäne wie die Quellsprache stammen.

Die Übersetzung verläuft in vier Schritten (Kit et al. 2002): Im ersten Schritt erfolgt die Extrahierung von "relevanten" Beispielen aus einem parallelen Korpus

oder bilingualen Wörterbüchern und mithilfe automatischer Textalignierung.[13] Es wird ermittelt, welches Beispiel der Ausgangsprache zu welchem Beispiel der Zielsprache zugehört. Diese Beispiele werden dann in einer Datenbank gespeichert. Hier gilt: Je größer die Datenbank ist, desto besser funktioniert das System. Im zweiten Schritt werden die Eingaben in einer Datenbank verwaltet. Hier wird für die Speicherung, Ausgabe (einschließlich Hinzufügen, Löschen und Modifizierung) und den Abruf von Beispielen gesorgt, um die automatische Übersetzung durchzuführen oder sie zu unterstützen. Im dritten Schritt werden die Bespiele angewendet: Es wird zuerst in der Datenbank nach den Einträgen gesucht, die mit der Eingabe am ähnlichsten sind und zielsprachliche Ausdrücke extrahiert. Im vierten und letzten Schritt werden schließlich die entsprechenden Teilübersetzungen aus der Datenbank zu einer korrekten Übersetzung rekombiniert.

Im Unterschied zu dem reinstatistischen MÜ-Verfahren, wo die Übersetzung aufgrund der Wahrscheinlichkeitsberechnung der Übersetzungsmöglichkeiten des Quelltexts ermittelt wird, wird in dem beispielbasierten Ansatz auf bereits vorhandene mit ihren jeweiligen Übersetzungen gespeicherte Sätze, Redewendungen bzw. Sequenzen von Wörtern zurückgegriffen und nach möglichst genauen Matches für die Übersetzung gesucht, die dann zu vollständigen Texten der Zielsprache zusammengesetzt werden. Werden beim Vergleich keine Übereinstimmungen in den zuvor übersetzen und angelegten Korpusdaten gefunden, kann das System keine Übersetzung liefern, woraus sich der große Nachteil des Ansatzes ergibt.

Zu den Vorteilen dieses Ansatzes zählt unter anderem, dass das in Beispielen eingeschlossene Sprachwissen bei der automatischen Übersetzung einbezogen wird. Ähnlich wie bei anderen korpusbasierten Ansätzen erzielt das MÜ-System mit einem größeren Beispielkorpus in der Regel bessere Qualität der Übersetzung. Verglichen mit einem regelbasierten MÜ-System, dessen Erweiterung um jedes zusätzliche Sprachpaar die Erstellung von neuen zeitaufwendigen sprachspezifischen Regeln erfordert, ist der Ausbau eines beispielbasierten MÜ-Systems mit weiteren Korpora und Lexikons schneller und ist zudem mit geringerem Aufwand verbunden.

---

13 Die Alignierung kann auch manuell von den Expert/innen gemacht werden, die auch natürlich recht zuverlässig Beispiele finden können, die ihrerseits die Präzision erhöhen können, aber der Aufwand wäre zu groß, ein Korpora von mehreren Millionen Wörter für praktische Anwendungen zu bearbeiten.

4.2.3 Kontextbasierte MT

Die Nachteile sowohl von statistikbasierten als auch beispielbasierten Ansätzen bestehen darin, dass für ihre Umsetzung große Mengen von parallelen und alignierten Textdaten benötigt werden, von deren Quantität und Qualität die Zuverlässigkeit der Übersetzung abhängt. Die Verfügbarkeit von solchen Korpora ist sprachabhängig. Die Voraussetzung für die Zusammenstellung eines Korpus bilden Texte, für die Übersetzungen vorliegen. Einerseits gibt es Sprachpaare, die öfter übersetzt werden und für die daher mehr parallele Texte zu finden sind. Andererseits gibt es Sprachpaare, die selten übersetzt werden und zu denen nur wenige oder gar keine parallelen Texte vorliegen.

Mit einem kontextbasierten Ansatz (engl. *Context Based Machine Translation*, abgekürzt CBMT) wird versucht, dieses Problem umzugehen: Für seine Umsetzung sind keine parallelen Korpora erforderlich. Stattdessen nutzt der Ansatz umfangreiche einsprachige Korpora der Zielsprache sowie ein zweisprachiges Vollformenlexikon. Optional kann ein kleineres, einsprachiges ausgangssprachliches Korpus eingesetzt werden, mit dem die Übersetzungsqualität gesteigert werden kann (Carbonell et al. 2006). Der Aufwand ein monolinguales Textkorpus (50 GB-1 TB) zu erstellen ist nicht groß, da es beispielsweise ausschließlich auf vom Web heruntergeladenen und indizierten Texten aufgebaut werden kann. Allgemein gilt auch hier: Je größer das Korpus ist, desto präziser ist die spätere Übersetzung. Den größeren Aufwand nimmt allerdings die Erstellung eines bilingualen Vollformenwörterbuchs in Anspruch. Auch hier gilt: Je größer der Umfang der Wörterbücher ist, desto zuverlässiger ist die spätere Übersetzung.

Als Erstes wird der Quelltext in N-Gramme zerteilt, die in der Regel zwischen 4 und 8 Wörtern lang sind. Dann werden im Wörterbuch alle möglichen Übersetzungsvarianten für jedes Wort jedes N-Gramms ermittelt, übersetzt und in die zielsprachlichen N-Gramme übertragen. In diesem Schritt können für jedes einzelne N-Gramm der Quellsprache eine große Anzahl von allen möglichen zielsprachlichen N-Grammen entstehen, die dann im monolingualen Korpus der Zielsprache abgeglichen werden. Die Variante, die größere oder längere Treffer im Korpus hat, wird dann weitergeführt und die restlichen aussortiert. Damit werden ungültige Übersetzungen ausgefiltert. Für den Fall, dass keine überschneidenden N-Gramme im Korpus gefunden werden, kann das System mit einem Synonymgenerator erweitert werden, wodurch die Wahrscheinlichkeit, sich überschneidende N-Gramme zu finden, erhöht wird (vgl. Carbonell et al. 2006).

Der Vorteil der kontextbasierten MÜ ist, dass die Wörter in ihrem Kontext übersetzt werden, wodurch viele Ambiguitäten gelöst werden können. Beispiele

der kontextbasierten MÜ sind CONTRAST (Isahara/Uchida 1995) und REFTEX (Kjærsgaard 1987).

4.3 Andere Ansätze

Wie bereits eingangs erwähnt wurde, scheint natürliche Sprache zu komplex zu sein, um ganzes sprachliches Wissen, das sie umfasst, durch Regeln zu beschreiben oder statistisch zu berechnen. Jeder der oben beschriebenen Ansätze stößt früher oder später an seine Grenzen des Möglichen. Deswegen wird versucht, sogenannten hybride Systemen zu entwickeln, bei denen mit verschiedenen Übersetzungssystemen parallel gearbeitet wird, um die Vorteile verschiedener Ansätze in einem einzelnen System zu vereinen und die Probleme der einen oder anderen Ansätzen zu lösen. Eines der Beispiele für die Realisierung eines hybriden Ansatzes ist das MÜ-System SYSTRAN, das anfangs als regelbasiertes System entwickelt wurde und dann 2009 mit statistischen Komponenten ergänzt (*SYSTRAN annual financial report 2009*) wurde. Ein anderes Beispiel ist der hybride Ansatz von Smith/Clark (2009), welcher einen rein statistischen Ansatz mit einem beispielbasierten kombiniert.

Eine weitere Richtung nimmt die MÜ-Forschung mit der Durchsetzung der Erkenntnis, dass ein automatischer Übersetzungsprozess nur dann erfolgreich sein kann, wenn die Bedeutung eines Ausgangstextes von einem MÜ-System erschlossen wird, um schließlich diese im Zieltext wiederzugeben. Die Herausforderung eines sogenannten wissensbasierten Ansatzes (engl. *Knowledge Based Machine Translation*, abgekürzt KBMT) besteht darin, dass das System in der Lage sein soll, nicht nur das linguistische Wissen, z.B. morphologische und syntaktische Regeln, zu kodieren und interpretieren, sondern auch das allgemeine Weltwissen über einzelne Dinge, Sachverhalte, Ereignisse, Vorgehensweisen, das aus Enzyklopädien, Ontologien, Semantischen Netzen, Thesauri etc. gesammelt und erschlossen werden kann. Die wissensbasierte MÜ gilt als Spezialfall des Interlingua-Ansatzes (vgl. Ramlow 2009; Stein 2009). Ein solches System zu entwickeln, ist bis heute nur in einzelnen Experimentalprojekten und nur als Prototyp mit abgegrenzten Fachgebieten gelungen.

5. Resümee

Aus der biblischen Perspektive könnte die Sprachenvielfalt als eine Strafe Gottes für den Hochmut der Menschen angesehen werden (Keller 2004). Doch die Fortschritte in der Entwicklung der maschinellen Übersetzung haben

gezeigt, dass bereits heute viele MÜ-Systeme diese Strafe einigermaßen erträglicher machen können.

Am Anfang des Forschungsprozesses legten die Wissenschaftler/innen ihren Schwerpunkt auf die Entwicklung von Systemen, die auf ein komplex definiertes Regelwerk aufgebaut und durch linguistisches Wissen gesteuert worden sind. Doch schon bald kam es zur Ernüchterung, denn es wurde klar, dass es unrealistisch war, Regeln zu definieren, die alle denkbaren Fälle des Sprachgebrauchs berücksichtigen können. Hinzu kommt, dass menschliche Sprache lebendig ist und sich ständig verändert. Mit der Zeit kommen neue Wörter hinzu, die Wortbedeutungen ändern sich.

Durch die starke Zunahme an verfügbaren digitalen Texten in verschiedenen Sprachen und stetige Steigerung der Speicher- und Rechenleistung sind weitere Ansätze technisch ermöglicht worden, die keinerlei explizites linguistisches Wissen für die Realisierung eines MÜ-Systems erfordern. Die Ansätze beruhen auf Berechnungen der Wahrscheinlichkeiten einer Übersetzung. Ein großer Vorteil für den Erfolg der statistischen MÜ-Systeme besteht in dem – im Vergleich mit den etablierten regelbasierten Übersetzungssystemen – wesentlich geringeren Aufwand für die Hinzufügung weitere Sprachpaare. Der Verzicht auf die Nutzung linguistischen Wissens hat aber die MÜ mehr und mehr an ihre Grenzen gebracht.

Grundsätzlich gilt, dass bei regelbasierten Ansätzen die Qualität der Übersetzung in der Menge, Übergeneralisierbarkeit und Genauigkeit von dem in den Regeln kodierten linguistischen Wissen abhängt, bei statistischen Ansätzen hingegen von der Größe und der Qualität der parallelen und alignierten Textkorpora, die zum Lernen genutzt werden.

Betrachtet man die verschiedenen MÜ-Strategien, kommt man zu der naheliegenden Erkenntnis, dass sie in hybriden Systemen vereint werden könnten, um die Vorteile verschiedener Techniken weitgehend miteinander zu verbinden und die unerwünschten Nachteile möglichst zu vermeiden. Hieran wird derzeit verstärkt gearbeitet.

Obwohl viele Universitäten und viele Firmen daran arbeiten, die MÜ zu verbessern, bleibt die Qualität der regelbasierten, statistischen und hybriden MÜ-Systeme der von professionellen menschlichen Übersetzungen weit unterlegen. Derzeit fehlen noch Technologien, die ermöglichen, zahlreiche semantische, pragmatische und kontextuelle Aspekte sowie das Weltwissen und Kulturunterschiede zu erfassen, um nicht nur Gebrauchstexte, sondern auch literarische Prosa oder gar Poesie aus einer Ausgangssprache in eine beibiege Zielsprache zu übersetzen.

6. Literatur

AlAnsary, S. (2011): *Interlingua-based Machine Translation Systems: UNL versus other interlinguas. Proceedings of the 11th International Conference on Language Engineering, 2011.* Kairo.

Andries, P. (2008): *Unicode 5.0 en pratique: Codage des caractères et internationalisation des logiciels et des documents.* Paris: Dunod.

Bennet, S./Gerber, L. (2003): Inside commercial machine translation. In: Somers, H.L. (Hg.): *Computers and translation: a translator's Guide.* Amsterdam: John Benjamins Publishing, 176–190.

Brown, R.D. (1996): Example-based machine translation in the pangloss system. In: *Proceedings of the 16th Conference on Computational Linguistics, 1996.* Kopenhagen: Center for Sprokteknologie, 169–174.

Bär, J.A. (Hg.) (2003): *Von "aufmüpfig" bis "Teuro": die "Wörter der Jahre" 1971–2002.* Mannheim: Dudenverlag.

Carbonell, J./Klein, S./Miller, D./Steinbaum, M./Grassiany, T./Frey, J. (2006): *Context-Based Machine Translation. Proceedings of the Association for Machine Translation of the Americas, 2006.* Boston.

Carstensen, K.-U./Ebert, C./Ebert, C./Jekat, S./Langer, H./Klabunde, R. (Hg.) (2010): *Computerlinguistik und Sprachtechnologie: Eine Einführung.* 3. überarb. u. erw. Aufl. Heidelberg: Spektrum, Akademischer Verlag.

Carstensen, K.-U./Ebert, C./Endriss, C./Jekat, S./Klabunde, R./Langer, H. (Hg.) (2004): *Computerlinguistik und Sprachtechnologie: Eine Einführung.* 2. überarb. u. erw. Aufl. Heidelberg: Spektrum, Akademischer Verlag.

Eberle, K. (2008): Integration von regel- und statistikbasierten Methoden in der Maschinellen Übersetzung. In: *Journal for Language Technology and Computational Linguistics*, 37–70.

FUEN (Föderalistischen Union Europäischer Volksgruppen). Internet: www.fuen.org/de/europaeische-minderheiten/allgemein/.

Haspelmath, M. (2008): Vergleichende Erforschung der weltweiten Vielfalt des menschlichen Sprachbaus. In: Berlin-Brandenburgische Akademie der Wissenschaften (Hg.): *Jahrbuch 2007.* Berlin: Georg Thieme Verlag, 140–146.

Hutchins, W.J./Somers, H.L. (1992): *An introduction to machine translation.* London: Academic Press.

Isahara, H./Uchida, Y. (1995): Analysis, generation and semantic representation in CONTRAST – a context-based machine translation system. In: *Systems and Computers in Japan* 26 (14), 37–53.

Jurafsky, D./Martin, J.H. (2009): *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 2. Aufl. Upper Saddle River/NJ u.a.: Prentice Hall, Pearson Education International.

Kay, M./Röscheisen, M. (1993): Text-Translation Alignment. In: *Computational Linguistics* 19 (1), 121–142.

Keller, R. (2004): *Sprachwandel. BDÜ 2000: Faszination Sprache – Herausforderung Übersetzung.* Unveröffentlichter Aufsatz.

Kit, C./Pan, H./Webster, J.J. (2002): Example-based machine translation: a new paradigm. In: Sin-wai, C. (Hg.): *Translation and Information Technology.* Hong Kong: Chinese University of Hong Kong Press, 57–78.

Kjærsgaard, P.S. (1987): REFTEX – a context-based translation aid. In: *Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics, 1987.* Kopenhagen, 109–112.

Koehn, P. (2005): Europarl: a parallel corpus for statistical machine translation. In: *Proceedings of the Tenth Machine Translation Summit, 2005.* Phuket, 79–86.

Koehn, P. (2010): *Statistical machine translation.* Cambridge: Cambridge University Press.

Koehn, P./Och, F.J./Marcu, D. (2003): Statistical phrase-based translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003.* Stroudsburg/PA: Association for Computational Linguistics, 48–54.

Krenz, M./Ramlow, M. (2008): *Maschinelle Übersetzung und XML im Übersetzungsprozess: Prozess der Translation und Lokalisierung im Wandel.* Berlin: Frank & Timme.

Kuhlen, R./Semar, W./Strauch, D. (2013): *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis.* 6., völlig neu gef. Ausg. Berlin: Walter de Gruyter.

Mishra, R.B. (2010): *Artificial Intelligence.* Neu-Delhi: PHI Learning.

Oard, D.W./Och, F.J. (2003): Rapid-response machine translation for unexpected languages. In: *Proceedings of the MT Summit IX, 2003.* New Orleans: International Association for Machine Translation.

Pan, C. (2008): Einführung in die Volksgruppenfrage. In: Gamper, A./Pan, C. (Hg.): *Volksgruppen und regionale Selbstverwaltung in Europa.* (= Schriften zum internationalen und vergleichenden öffentlichen Recht 8). Wien: Facultas-WUV, 21–36.

Ramlow, M. (2009): *Die maschinelle Simulierbarkeit des Humanübersetzens: Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik.* Berlin: Frank & Timme.

Schubeft, K. (1995): *Zum gegenwärtigen Stand der maschinelten Übersetzung. Die 5. Jahrestagung der Gesellschaft für Interlinguistik, 1995.* Berlin.

Schulz, T. (2013): Lost in Translation. In: *Der Spiegel* 37, 78–79.

Schäfer, F. (2002): *Die maschinelle Übersetzung von Wirtschaftsfachtexten: Eine Evaluierung anhand des MÜ-Systems der EU-Kommission Systran im Sprachenpaar Französisch-Deutsch.* (= Europäische Hochschulschriften, Reihe 21, Linguistik). Frankfurt a.M. u.a.: Peter Lang.

Smith, J./Clark, S. (2009): EBMT for SMT: A new EBMT-SMT hybrid. In: Forcada, M.L./Way, A. (Hg.): *Proceedings of the 3rd International Workshop on Example-Based Machine Translation, 2009.* 3–10.

Somers, H. (2003a): An overview of EBMT. In: Carl, M./Way, A. (Hg.): *Recent advances in Example-Based Machine Translation.* Dordrecht: Kluwer, 3–57.

Somers, H.L. (2003b): Translation memory systems. In: Somers, H.L. (Hg.): *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing, 31–48.

Spiegel Online (2013): Längstes Wort der deutschen Sprache verschwindet. URL: www.spiegel.de/panorama/gesellschaft/laengstes-wort-der-deutschen-sprache-verschwindet-a-903370.html [Zuletzt abgerufen am 1.07.2014].

Stein, D. (2009): Maschinelle Übersetzung – ein Überblick. In: *Journal for Language Technology and Computational Linguistics* 24 (3), 5–18.

SYSTRAN Annual Financial Report 2009. URL: http://www.systransoft.com/download/annual-reports/systran-annual-report-2009.pdf [Zuletzt abgerufen am 1.07.2014].

Tripathi, S./Sarkhel, J.K. (2010): Approaches to machine translation. In: *Annals of Library and Information Studies* 57, 388–393.

Wahlster, W. (2000): *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin/Heidelberg u.a.: Springer.