

1 Einheit 1 - Vorstellung/Translationsrelevante Technologien

1.1 Organisatorisches

Das gesamte Material, also Folien, werden stets hochgeladen, ist jedoch für die Prüfung nicht ausreichend und es werden wahrscheinlich auch Notizen vonnöten sein um die Prüfung zu bestehen.

Die Vorlesung ist Vortragsbasiert, also gibt es keine Bewertung während des Kurses, jedoch werden im Jänner die Vorlesungen online sein, wodurch man die Werkzeuge gleich am PC testen kann.

Der Text von Sharon O'Brien über Fehler in der Maschinentranslation ist 16 Seiten lang und ist bis zum 14.12 zu lesen. Dieser ist auch relevant für die Prüfung. Die Prüfung selbst ist in Präsenz und Single-Choice. Er wird Praxisorientiert sein.

Es gibt insgesamt 6 Einheiten, wobei 3 vor Weihnachten in Präsenz stattfinden werden. Nach Weihnachten gibt es drei weitere Einheiten, welche online stattfinden werden.

Der Text ist einem längeren Buch entnommen, welches Open-Access ist weshalb man überall frei darauf zugreifen kann.

1.2 Was für Übersetzungstechnologien gibt es?

Heutzutage gibt es einige Technologien mit denen Übersetzt werden kann, oder die einem bei der Übersetzung helfen. Einige Werkzeuge sind DeepL oder Google Translate, welche automatisch übersetzen. Werkzeuge die bei der Übersetzung helfen sind Wörterbücher wie Linguee, Reverso aber auch reine Wörterbücher wie Pons oder der Duden.

1.3 Wie teilt man Translationstechnologie ein?

Die Reichweite von Übersetzungswerkzeugen ist außerordentlich groß. Werkzeuge zur Übersetzung sind zwar die beliebtesten, wo es mehrere dutzend unterschiedliche Anbieter gibt, jedoch gibt es auch eine größere Zahl an Anbietern für unter anderem Maschinenlernstechnologie. Übersetzungstechnologien können in zwei Kategorien unterteilt werden:

- Computergestützte Technologien (CAT)
 - Diese kann man weiter einteilen:
 - * Nach Grad der Automatisierung: Sind Menschen mehr oder weniger im Prozess des Übersetzens involviert? Es reicht von Vollautomatischen Übersetzern bis zur traditionellen Humantranslation, wo man keine Hilfsmittel verwendet. In der Mitte dieser liegen die *Computer-assisted translation technologies* Machine-aided human translations und human-aided machine translations, abhängig davon welcher Anteil größer ist. Im Feld des literarischen Übersetzens wird immer noch größtenteils ohne Hilfsmittel übersetzt, da die Übersetzung sehr vom Kontext abhängig ist. Man kann diese Werte als Tabelle darstellen, indem zwischen *Human Translation*, *Computer-aided Translation*, *Machine Translations* unterschieden wird. Den größten Teil macht hierbei das CAT aus, da es einen Übersetzer bei der Tätigkeit unterstützt, was meist die beste Kombination aus Zeit und Effizienz ist.
 - * Nach Verwendung im Übersetzungsprozess: *Before-Translation*, *During-Translation*, *After-Translation*
 - * Nach der Beziehung zum Übersetzungsprozess: Wie Allgemein oder Spezifisch ist die Technologie? Ist es Allgemein wie z.B. Microsoft Word oder ist es spezifischer wie PONS oder Linguee. An unterster Stelle liegen Programme spezifisch für professionelle Übersetzer wie Trados oder Antconc
 - * Nach der Dimension der Translation: Werden die Werkzeuge für die Lehre, professionelle Übersetzer oder im Bereich der Forschung des Übersetzens verwendet?

1.4 Geschichte der Translationstechnologie

Maschinelle Übersetzung besteht theoretisch schon seit den 30er Jahren als die Ersten PCs entwickelt wurden. Die Idee des automatischen Übersetzens ist jedoch bedeutend älter wobei Leibniz und Descartes bereits Überlegungen darüber gemacht hatten. **George Artsrouni** patentierte 1930 das *Mechanical Brain*. In diesem ersten Versuch wurden spezifische

Phrasen abgespeichert, wodurch es nur einen limitierten Anwendungsbereich hatte, jedoch die Basis für spätere Forschung bildete. Gleichzeitig überlegte der Russe **Petr Petrovich Smirnov-Troyanskij** über eine Übersetzungsmaschine, welche Wörter konjugieren kann, konnte es jedoch nie realisieren.

1.4.1 Das Weaver Memorandum

Warren Weaver, bekannt vom Warren Weaver Kommunikationsmodell, überlegte auch was eine Übersetzung ausmacht. Er hatte mehrere Überlegungen:

1. Eine Übersetzung muss Kontextbasiert sein, also muss es eine Bank für Geld von einer Bank zum sitzen unterscheiden können
2. Sie muss logische Komponenten der Sprache abbilden können
3. Da zu dem Zeitpunkt noch Kodewechsel populär war, dachte er auch über universelle Bedeutungselemente nach, welche in jeder Sprache das selbe bedeuten. Wenn man diese Elemente nun findet könnte eine Maschine jede Sprache in jede Sprache übersetzen. Das hat sich später als fehlerhaft herausgestellt, wodurch diese Überlegung hinfällig wurde
4. Es gibt Sprachuniversalia. Dass es Konzepte gibt, welche in jeder Sprache vorhanden sind. Wenn man also die Universalia definiert sollte man diese Worte einfacher übersetzen können.

1.4.2 Erste Schritte

In den 40er Jahren wurden die ersten Überlegungen von maschineller Übersetzung getätigt, wie eben das Weaver Memorandum. Zu diesem Zeit wurde auch Übersetzung als Kodewechsel gesehen.

1954 wurde das erste System zur Übersetzung vom Englischen zum Russischen vorgestellt, größtenteils um in der Zeit des Kalten Krieges schneller an Informationen zu kommen. Zu diesem Zeitpunkt war man der Ansicht, dass es bald zu einem Durchbruch in der maschinellen Übersetzung kommen würde. Nachdem dies nicht so schnell passierte, wurde das ALPAC(Automatic Language Processing Advisory Committee) gegründet. Dieses Komitee kam zu dem Schluss, dass vollautomatische Übersetzung nicht so schnell erreichbar ist, und man stattdessen in Menschliche Hilfswerkzeuge zur Übersetzung investieren sollte. Dadurch wurde ein Großteil der Forschungsgelder in den USA von maschineller Übersetzung abgezogen, welches zu diesem Zeitpunkt Vorreiter in der Technologie war. Andere Länder wie Japan führten diese Forschung jedoch weiter.

In den USA wurde aber sehr wohl die Translation Memory entwickelt, welche als Glossar für Übersetzer dienen sollte.

In Europa gab es auch Forschung für Terminologiedatenbanken wie EURODICAUTOM, welcher Vorgänger des IATE war. Da diese Forschung jedoch in alle Sprachen der Mitgliedsstaaten übersetzt werden musste, war der Prozess etwas verlangsamt. Eine weitere für Rechtstexte von Kanda ist TERMIUM.

Das IATE ist heute noch öffentlich zugänglich und kann verwendet werden um spezifische Fachtermini in allen Sprachen der EU nachzusehen. So kann man nachsehen, was zum Beispiel die GDPR in verschiedenen Sprachen heißt, wie gut dieser bewertet wurde und woher er stammt.

In den 80er Jahren tauchten kommerzielle Übersetzungstools auf und TRADOS, welches seit 1984 besteht, ist heute noch ein populäres Werkzeug. Zwar hat es vor 40 Jahren sehr anders ausgesehen, doch ist die Kernfunktionalität noch immer die selbe. So versucht das Programm bereits getätigte Übersetzungen zu speichern und den Benutzer darauf hinzuweisen wenn die momentane Übersetzung mit einer früheren ähnlich ist. Es weiß auch, wenn nur ein Teil der Übersetzung ähnlich ist.

Ab den 1990er Jahren gab es immer mehr Anbieter solcher Technologie, wobei diese Tools auch mehr Funktionalitäten erhielten. Zwar ist die Speicherung von Übersetzungen die Grundfunktionalität, doch kamen mit der Zeit immer mehr unterstützende Funktionen hinzu. Eines davon ist die Alignierung, mit welchem man externe Texte in das Tool einpflegen kann um diese auch innerhalb des Programms verwenden kann.

Mit der Zeit wurden diese Tools immer benutzerfreundlicher und es wurde versucht die Tools an den Workflow der Übersetzer anzupassen, anstatt zu erwarten, dass die Menschen sich der Maschine anpassen. Auch wurden Cloud-Funktionen und Integration in andere Programme realisiert. Gleichzeitig wurden offene Standards entwickelt um Terminologiedatenbanken mit anderen Leuten als Datei zu Teilen. Maschinelle Übersetzung erfuhr durch den ALPAC-report einen Dämpfer in den USA. In Japan wurde diese Technologie jedoch weiterentwickelt, und in den 1990er Jahren wurden die ersten statistischen Übersetzungstechnologien veröffentlicht. Google stieg erst 2007 auf statistische Übersetzungstechnologie um und 2016 auf neuronale Übersetzungstechnologien. Dass 2016 neuronale Übersetzer besser geworden waren, wurde durch das maschinelle Qualitätssicherungsverfahren BLEU

2 Information, Wissen, Technologie und Translation

Ein jeder Übersetzungsauftrag erweitert das Wissen eines jeden Übersetzers, da man sich stets sprachlich und kulturell weiterbilden muss. Da man bei der Übersetzung nicht nur Wörter, sondern den Sinn übersetzt, muss man den Text zuerst verstehen, bevor er übersetzt werden kann.

2.1 Wissensmanagement

Wissensmanagement ist die Ansammlung von Methoden um effektiv übersetzen zu können. Hierbei gibt es einige Grundbegriffe:

- Daten
 - Zeichen aus einem Datenvorrat, welche durch Syntax kombiniert werden
- Information
 - Die Syntax um diese Zeichen miteinander zu kombinieren.
- Wissen
 - Der Kontext, welcher die Zeichen und die Syntax im richtigen Weg anwendet
 - Geteilt in wiederum zwei Kategorien:
 - * Explizites Wissen: Eindeutig erlerntes Wissen, wie es in einem Lehrbuch gefunden werden kann
 - * Implizites Wissen: Nicht offensichtliches Wissen welches indirekt weitergegeben wird.

Wissensmanagement wird auf unterschiedlichen Ebenen angewandt. Wissensmanagement innerhalb einer Organisation beispielsweise wird größtmöglich effektiv untereinander geteilt um die Organisation weiterzubringen? Persönliches Wissensmanagement hingegen ist das Wissen, welches eine Person benötigt um erfolgreich zu sein. Zum Beispiel wenn man selbstständig ist.

Es wird stets danach gestrebt, Wissen zu personalisieren. Implizites Wissen ist ein Prozess, welcher es vermittelt, jedoch findet man dieses oft nicht in einem Buch in einer Bibliothek.

Es gibt verschieden Arten von Wissen:

- Sprachliche, linguistische und übersetzerische Kenntnisse
 - Kann in kodifiziertes
 - * Kann in Lehrbüchern gefunden werden und stellt das Basiswissen dar, welches für Übersetzung relevant ist.
 - und nicht-kodifiziertes Wissen eingeteilt werden:
 - * Kontextverständnis oder Bedürfnisse der ZIELLESERSCHAFT. Diese können nicht explizit mit einem Buch erlernt werden und müssen mit der Zeit erarbeitet werden.

- Landes- und Kulturwissen
 - Kodifiziert:
 - * Variationen zwischen Ländern wie British und American English
 - Nicht Kodifizierbar:
 - * Moralische Einstellungen beziehungsweise das politische Umfeld. Übersetzungsverbände und kulturelle Veranstaltungen können verwendet werden um sich dieses Wissen anzueignen.
- Allgemeines Sachwissen
 - Kodifiziert:
 - * Industrienormen zwischen Ländern oder spezifisches Fachwissen. Diese können in Ontologien dargestellt werden. Diese zeigen Verhältnisse zwischen Gegenstandsbereichen und zeigen aktive und passive Prozesse.
 - Nicht kodifiziert:
 - * Hausverstand und Erfahrung, welches durch bspw. Plattformen
- Kunden und firmenspezifisches Wissen

In diesen Studien wird festgelegt, welche Kompetenzen Übersetzer benötigen um übersetzen zu können. In Europa besteht der EMT Kompetenzrahmen, welcher in Zusammenarbeit mit Übersetzungsorganisationen entstanden ist.

Studien die sich diesem Kompetenzrahmen verpflichten, lehren anhand dieses Rahmens, welcher darüber schreibt, dass ein Übersetzer bspw. wissen sollte wie maschinelles Übersetzen funktioniert.

2.2 Datenbestände

2.2.1 Lexikalische Datenbestände

Lexikalische Datenbestände basieren auf der Lexikographie, welche sich mit Wörterbüchern beschäftigt. Computergestützte Lexikographie besteht schon seit den 40er Jahren, haben sich in den letzten 80 Jahren jedoch bedeutend weiterentwickelt. So haben sich einige Standards with TEI P5 durchgesetzt, mit welchem man Wörterbucheinträge standardisiert formatieren kann.

Bei Elektronischen Wörterbüchern kann man zwischen einigen Sachen unterscheiden:

Ob mit den Wörterbüchern Menschen oder Computer angesprochen werden sollen. Zusätzlich unterscheidet man zwischen Retrogradisierten oder neu-digitalisierten Wörterbüchern. Ob das Wörterbuch digitalisiert wurde oder komplett neu als digitales Wörterbuch konzipiert wurde.

Besteht das Wörterbuch aus freiwilligen Nutzern oder wird diese redaktionell betreut.

Im Gegensatz zu Wörterbüchern gibt es auch Terminologische Ressourcen, welche oft riesige Korpora darstellen und auch oft öffentlich einsehbar sind. Innerhalb der Terminologie unterscheidet man zwischen Gemein- und Fachsprache, welche zwischen Termini und Wörter geht. Wenn eine Fachsprache eine spezielle Definition für ein Gemeinsprachliches Wort hat, ist das ein Termini.

Man muss auch zwischen Begriff, Benennung und Gegenstand unterscheiden. Ein Gegenstand ist ein beliebiger Ausschnitt aus der wahrnehmbaren Welt: Ein Haus kann gleich ein Gegenstand sein wie das Konzept der Freundschaft. Diese Gegenstände werden durch Begriffe definiert. Diese Begriffe existieren nur in den Gedanken der betroffenen Personen. Erst wenn die Benennung ins Spiel kommt wird mit dem Begriff ein spezifisches Wort verbunden. Hier werden Terminologie und Lexikographie unterschieden. Bei der Terminologie steht die Begriff an erster Stelle, während in der Lexicographie die Benennung wichtiger ist.

Also beginnt die Lexikographie mit der Bezeichnung und leitet von dort die Benennung ab, während es bei der Terminologie umgekehrt ist.

Terminologiedatenbanken haben auch ihren eigenen Standard, sodass das Termbase Exchange Format (TBX) universell verwendbar ist.

Mittels diesem Standard kodieren Unternehmen ihre Datenbanken manchmal und machen diese öffentlich zugänglich.

2.3 Translation Memory Tools

Während CAT tools und Translation Memory Tools oft synonym verwendet werden, ist CAT der Überbegriff und wenn Leute von CAT Tools sprechen, meinen sie oft TMTs. Diese Tools funktionieren so, dass man bereits getätigte Übersetzungen abspeichern kann und das Programm, wenn ein gleicher oder zumindest ähnlicher Satz übersetzt werden muss, diesen einem anzeigt. Solche Datensätze bestehen stets aus Metadaten und einem Übersetzungspaar.

Die bekanntesten Tools für diesen Zweck sind *Déjà Vu*, *MemoQ* oder *Memsources*.

In der GUI wird das Interface in zwei Segmente geteilt. Auf einer Seite hat man den Originaltext und ein Eingabefeld zur Übersetzung, während der andere die Datenbank durchsucht und eventuelle Treffer anzeigt.

Man nennt diese Tools zwar immer noch Translation Memory Tools, doch haben solche Programme heutzutage viel weitere Möglichkeiten.

Solche Translationtools bestehen bereits seit den 80er Jahren mit TRADOS, welches auch heute noch Marktführer ist.

TMTs bestehen aus verschiedenen Komponenten:

- Übersetzungsspeicher
 - Können lokal oder in der Cloud gespeichert werden. So kann jeder die aktualisierte Version abrufen.
- Terminologiedatenbank
- Alignierungsprogramm
 - Wenn man bereits einen großen Corpus hat, dieser aber nicht Teil des Programms ist, gibt es sogenannte Alignierung um solche Korpora zu importieren. Traditionell wird der Zeiltext zuerst Segmentiert und dann diese Segmente in die Datenbank importiert.
- Maschinelle Übersetzung

Die TMTs speichert den Korpus in kleineren Segmente, da das die Abspeicherung vereinfacht.

Wenn das TMT die Datenbank durchsucht unterscheidet man zwischen 100% Matches und fuzzy matches. Bei einem 100% Match wurde in der Vergangenheit der genau gleiche Satz übersetzt, während es bei einem fuzzy match nicht unbedingt der Fall sein muss. Man kann es individuell einstellen, wann man einen fuzzy match haben will.

Es gibt mehrere Wege um so ein TMT verwenden zu können. So gibt es ein Add-On zu Microsoft Word namens WordFast, aber auch alleinstehende tools wie die vorhin genannten MemoQ.

Diese tools müssen nicht unbedingt auf einem PC installiert werden und man kann gewisse tools, welche über das Intranet einer Organisation oder Web/Cloud basiert ist.

2.4 Anwendungsbeispiele

TMTs sind sehr nützlich, wenn man immer wieder mit dem selben Kunden zu tun hat oder oft einen ähnlichen Texttypen aufweisen, da man dann direkt auf die Datenbank zugreifen kann.