Universität Rostock

Traditio et Innovatio

Automatische Sprachübersetzung von LaTeX-Dokumenten

Name:	Hendrik Theede
Matrikelnummer:	221201256
Abgabedatum:	20251202
Betreuer und Gutachter:	Prof. Dr. rer. nat. habil. Clemens H. Cap Universität Rostock Fakultät für Elektrotechnik und Informatik

Bachelorarbeit am Lehrstuhl für Informations- und Kommunikationsdienste

Fakultät für Elektrotechnik und Informatik

Abstrakt

placeholder

Inhaltsverzeichnis

1 Einleitung

Wohingegen sich die Sprachübersetzung im Web schnell auf gängige Technologien wie DeepL oder Google's Gemini zurückführen lässt, zeigt sich eine ähnliche Übersetzung von TEX und LATEX Dokumenten nur in ernüchternder Weise verfolgt. Lösungsansätze zu diesem Problem existieren bereits, allerdings gehen diese oftmals Umwege und trennen die Fähigkeiten der TEX-Engine nicht in jedem Fall von den Technologien, welche verwendet werden sollen, um textliche Inhalte einer menschlichen Sprache in eine Andere zu übersetzen.
Wo eine naive Nutzung solcher Software bereits im Alltag schnell Schwierigkeiten aufzeigt, ist insbesondere in einem wissenschaftlichen und mathematischem Kontext eine gezielte Verwendung der dieser Technologien erstrebenswert, sodass nicht jegliche Texte unabhängig voneinander und kontextlos übersetzt werden. Andernfalls wäre es denkbar, dass das deutsche Wort "ungerade" seine Bedeutung gegenüber einer mathematischen Operation verliert (nach welcher eine Zahl modulo 2 in 1 resultiert) und als umgangssprachliches "schief" interpretiert wird und im Englischen respektiv als "odd", bzw. "crooked" übersetzt werden würde. Neben einer solchen Erhaltung von Kontexten ist auch eine selbstständige Erkennung der zu übersetzenden Sprache (Originalsprache eines Dokumentes) interessant, jedoch nicht zwingend erforderlich.
Weiterhin dürfen Übersetzungsprozesse selbstverständlich nicht darin enden, dass eine entstehende (bspw.) PDF entweder vollständig unlesbar wird. Daneben sollten allerdings auch keine unlesbaren Sektionen innerhalb der jeweiligen Dokumente entstehen, die aus von Layouting-Problemen resultieren, welche sich für die Übersetzung in einige Sprachen zeigen (jedoch in einigen Fällen unvermeidbar sind).
Wünschenswert ist neben vorigen Aspekten auch Möglichkeiten für den Endnutzer zu erlauben, sollte dieser spezielle Übersetzungen oder Kontexte für einige Wörter wünschen, welche jedoch nicht aus dem Dokument selbst hervorgehen. Außerdem sollte ein möglichst hoher Support für sowohl verschiedene menschliche Sprachen, aber auch verschiedene LATEX-Pakete gegeben sein, wobei Letzteres nur ein Bonus ist, sollten Systeme wie Ti*k*Z, bzw. `pgfplots` oder BibTEX innerhalb LATEX (zusammen mit TEX) nutzbar bleiben.

2 Problemfälle

Mittels TEX ist prinzipiell alles möglich (wie sich später zeigen wird), jedoch sollte man zunächst denkbare Schwierigkeiten für jegliche Sprachübersetzungen von TEX und LATEX Dokumenten nicht nur dahingegen schildern, dass sie auftreten könnten, sondern auch dahingegen klassifizieren, inwiefern sie häufig zu erwarten sind, geschweige denn sinnvoll oder gar unsinnig sein könnten (da sie z.B. eine zukünftige Bearbeitung eines Dokumentes erschweren könnten). Beruft man sich zunächst nur auf die reine LATEX-Syntax, werden vorerst wahrscheinlich nur simplere Probleme erkennbar, jedoch führt eine Näherung an die TEX-Engine (und insb. deren Primitiven) eine Vielzahl von komplexeren und speziellen Problemen mit sich.

2.1 Simple Probleme

Zeichenketten sind eine übliche Art und Weise, mit welcher man Wörter einer Sprache darstellen kann. Jedoch gehen die meisten Übersetzungs-Tools nicht davon aus, dass solche Zeichenketten Zeichen beinhalten, welche das folgende Wort zu einem Befehl (für eine Programmiersprache) machen. Zwar würde z.B. Google Translate für die Zeichenkette `Hello` korrekterweise das Deutsche `Hallo` liefern, aber bereits die Präambel von TEX-Dokumenten zeigt, wie `\title`, `\author` und `\date` respektiv zu `\Titel`, `\Autorin` und `\Datum` übersetzt werden würden (Stand: 01.10.2025). Benanntes Tool zeigt sich zudem inkonsistent. Beispielsweise wird `\section{saw}` zu `\Abschnitt{Säge}` übersetzt und `\section{Introduction}` zu `\section{Einführung}` übersetzt.

Whitespace sind eine herkömmliche Art verschiedene Wörter einer Sprache voneinander zu trennen (bspw. in den lateinischen oder kyrillischen Sprachen). Neben anfänglichen Schwierigkeiten, welche sich innerhalb von einzelnen Zeichenketten aufzeigen könnten, ist es genauso denkbar, dass einzelne Optionen in TEX oder LATEX innerhalb von eckigen oder geschwungenen Klammern nicht übersetzt werden dürften, ohne die Syntax zu brechen oder übersetzt werden müssten, damit ein gesamtes Dokument übersetzt wird. Denkbar sind hier direkt für das Erstere Farbdefinitionen, wie zum Beispiel `\definecolor{super light red}{rgb}{1,.5,.5}`. Sollte man versuchen **??**/Zeichenketten dadurch zu lösen, dass man einfach die Präambel in der Übersetzung ausschließt (also alles vor `\begin{document}`), so würde ein späteres Nutzen dieser Farbe `super light red` dafür sorgen, dass das Wort *light* alleine steht, und somit für nicht weit durchdachte Ansätze als ein zu übersetzendes Wort gelten würde.

Einbinden von anderen Dateien ist eine denkbare Art größere TEX-Dokumente in übersichtlichere kleinere Dateien zu strukturieren. Neben der Möglichkeit TEX-Dokumente selbst via `include` und `input` in ein übergreifendes Dokument einzufügen, ist es jedoch auch möglich verschiedene andere, bildliche Formate (bspw. PNG, PDF, ...) im Dokument zu integrieren. Insbesondere bei PDF kann es sehr interessant werden, ob und inwieweit textliche Inhalte erfasst und übersetzt werden, jedoch sind PDF ihrerseits wieder eine von LATEX und TEX abweichende Datei-/Dokumentenform und daher nicht weiter kritisch. Schade wäre es, wenn solche eingebundenen TEX-Dateien und deren textlichen Inhalte übersehen werden würden, andererseits unerwartet jedoch hervorragend, sollten sogar textliche Inhalte von PDF erkannt (und übersetzt) werden. Einer Erkennung von textlichen Inhalten auf einem Bild wird nicht weiter nachgesehen, da hierbei davon ausszugehen sein sollte, dass die Übersetzung von textuellen Inhalten in Bildern eine andere, gesonderte Disziplin in der Bildverarbeitung ist. (Evtl? Beispiele zeigen, dass hieran geforscht wird?)

2.2 Komplexere Probleme

Neben den zuvor geschilderten sehr einfachen Problemen, welche sich auch unabhängig von TEX (und LATEX) zeigen könnten (denn z.B. ein Übersetzen von Hashtags im Social Media sollte keine neue Idee sein) und gelöst sein müssen (da ansonsten sehr einfache und rudimentäre Werkzeuge für eine Dokumentenerstellung verloren gehen, da man sich ohne diese simplem Formatierungsoptionen wieder auf einfache Textdateien berufen könnte).

Makros sind eine Möglichkeit mehrere TEX-Befehle zusammenzufassen. Vor allem in LATEX sind eine Vielzahl dieser bereits vordefiniert, jedoch handelt es sich bei diesen meist um Wörter der englischen Sprache ("meist": manche dieser englischen Wörter treten auch in anderen Sprachen auf, bspw. *paragraph*↔"Paragraph"). Sollte es einem TEX-User leichter fallen in der z.B. französischen Sprache zu arbeiten, so könnte dieser beispielsweise neue, französische Makros mit
`\newcommand{\anglais}{This is some \textit{formatted} \texttt{english} \TeX{}-t}`
erzeugen. Das vorige Beispiel zeigt zudem auf, wie Texte innerhalb von TEX-Makros "verschwinden" können und wirft die Frage auf, wann und wie solche Texte übersetzt werden sollten. Am sinnvollsten erscheint zunächst nur Zeichenketten zu übersetzen, welche sich mit der prominentesten Sprache des gesamten Dokumentes decken, welche allerdings nicht ohne weiteres bekannt ist. Selbst wenn in dem gesamten Dokument größtenteils englische Wörter vorliegen, ist eigentlich nur interessant, in welcher Sprache die reinen Strings (welche auf der PDF lesbar erscheinen) geschrieben sind. Selbst diese Information alleine ist theoretisch gesehen noch keine Grundlage für eine Aussage darüber, welche Sprache in solche einem Fall übersetzt werden müsste, da man hier Kenntnis des eigentlichen, entgültigen Dokumentes bräuchte, denn es könnte auch von Interesse sein, innerhalb eines größtenteils z.B. deutschsprachigen Dokumentes nur vereinzelte, englische Sätze zu übersetzen. Hierauf wird in Abschnitt 2.4 näher eingegangen, da sich dieses Problem zunächst recht einfach durch eine Auswahlmöglichkeit der Ausgangssprache (= die zu Übersetzende) lösen ließe.
Gleiches ist zu berücksichtigen, sollte das Kommando `\renewcommand` verwendet werden, wobei dieses allerdings noch ein wenig mehr zulässt. Hiermit ist man auch dazu in der Lage existierende Befehle der LATEX-Syntax zu ändern, wodurch ein `\Abschnitt{Einleitung}` ebenfalls valide LATEX-Syntax werden könnte, welche ein TEX-Compiler als `\section{Einleitung}` richtig interpretieren könnte, aber ein übersetzendes Programm könnte dieses womoglich in `\section{example}` überführen. Dies scheint zunächst kein Problem zu sein, jedoch hätte zwischen einem `\renewcommand{\section}{\Abschnitt}` genauso ein `\newcommand{\section}{\frac{1+\sqrt{5}}{2}}` stattfinden können, wodurch `\section{example}` nicht in einem Abschnitt mit Titel "example", sondern in $\frac{1+\sqrt{5}}{2}$example resultieren würde.

Umgebungen sind, wie der Name es vermuten lässt, einzelne Bereiche im Dokument, welche gesondert behandelt werden und für welche sich jegliche Einstellungen, wie z.B. Textfarbe, Textgröße, Schriftart, Font und vieles Weitere nur für eine solche Umgebung anpassen lassen. Einerseits kann man über geschwungene Klammern `{}` eine Umgebung einmalig betreten oder verlassen, möchte jedoch auch die Möglichkeit erhalten diese erneut zu verwenden und ihr verschiedene Parameter zu übergeben. Eine Definition einer Umgebung in der Präambel lässt dies zu, wodurch sich neben den in **??** aufgezeigten Problemen nicht nur für etwaige Farboptionen und -einstellungen Strings aufzeigen, welche nicht übersetzt werden dürfen, sondern auch eigens (vom TEX-User) Ausgedachte (Hier: Verweis auf Anhang passend).

Pakete bieten eine TEX-Schnittstelle für die gesamte Welt! Zumindest rein theoretisch natürlich. Technisch gesehen bieten sie die Möglichkeit zuvor beschriebene Umgebungen und Makros in einer eigenen `.bst` zu bündeln, welche ihrerseits (vorrangig via) CTAN (jedoch auch auf jeglichem anderem Wege) zu

anderen TEX-Usern übertragen werden könnte. Verschiedene Pakete könnten hierbei eine Vielzahl individueller Probleme aufwerfen, zunächst ist jedoch mehr ein Fokus auf solche zu setzen, welche die Arbeit anderer Programme involvieren. Sie in Dokument mit einzubinden ist recht leicht und funktioniert nur auf eine begrenzte Anzahl an Methoden (`requirepackage` und `usepackage`) und sind ihrerseits, genauso wie **??**

Ti*k*Z ist zum Einen eine Möglichkeit in TEX zu malen, jedoch hauptsächlich dahingegen konszipiert in einem wissenschaftlichen Kontext verwendbare Diagramme mathematisch zu beschreiben oder auf Grundlage von Messwerten zu erzeugen. Die Syntax von Ti*k*Z und `pgfplots` kann innerhalb eines Dokumentes auch freistehende englische Wörter beinhalten, wie zum Beispiel in...

```
\begin{tikzpicture}[h!]
    \centering
    \begin{axis}[
        domain=-8:8
    ]
    \addplot{x};
    \end{axis}
\end{tikzpicture}
```

...bei welchem ein Übersetzen von "domain" Fehler produzieren würde, da Ti*k*, bzw. `pgf` von einem englischen Wort ausgeht.

BibTEX wird genutzt um Zitationen/Referenzen/Literaturverweise innerhalb eines Einzelnen oder mehreren Dokumenten zu nutzen und zu verwalten. Die BibTEX-Notation selbst beläuft sich auf eine einfache JavaScript Object Notation und trägt mit einer Ausnahme nur nicht zu übersetzende Inhalte, wie den Autor, den Titel des Werkes (welcher in der Originalsprache oder der durch den Autoren genehmigten übersetzten Titel), das Datum, einer URL, einer DOI, einer Angabe darüber, ob das zitierte Werk aus einem Buch, einer laufenden Reihe an wissenschaftlichen Publikationen (bspw. *nature*, *science*, *ACM Computating Surveys*,...) oder einer Konferenz (oder Ähnlichem) stammt. Neben diesen Angaben, welche allesamt nicht übersetzt werden brauchen, bleibt das Abstrakt eines zitierten Werkes interessant für einen Übersetzungsvorgang, sollte man davon ausgehen, dass man im Anschluss entstehende, übersetzte `.tex` Dateien an einen neuen Autoren übergeben möchte.

Mathematische Formeln selbst sind kein eigenes Paket, jedoch einer der praktischsten Use-Cases von TEX. Insbesondere für Menschen, welche sich eine handschriftliche Qualität und "Streichlust" (meint: das Durchstreichen auf dem Papier, sollte man sich verschrieben haben) mit der des Autoren (dieser Arbeit) teilen, sollte das digitale Medium TEX einiges an Aufwand ersparen und jegliche Herleitungen deutlicher und übersichtlicher machen. Hierzu gibt es wiederum mehrere denkbare Pakete, welche diesen bereits in TEX inhärent verankerten "*math mode*" erweitern oder vereinfachen können.

2.3 Spezielle Probleme

Höhere Vernestungsgrade

In Tabellen

In mathematischen Umgebungen

Quelltexte

Kommentare

Definitionen

Catcode und Unicode

2.4 Weitere Schwierigkeiten

Beabsichtigt ist dieser Abschnitt nicht in der Reihe von Problemen aufgefasst, sondern als Schwierigkeit(en) formuliert, da man sich hier von den Problemen abwenden würde, welche in der TEX-Syntax auftreten und bei sprachliche Hürden angelangt, welche sich für und zwischen verschiedenen Sprachen zeigen könnten.

Mehrdeutigkeiten innerhalb einer Sprache führen unter Umständen zu missverständlichen Übersetzungen.

Redewendungen sind eine Art und Weise...

Wirrer Sprachwechsel meint ein rapides Springen zwischen verschiedenen Menschensprachen innerhalb eines Dokumentes. Die Fragestellung hierbei ist, inwiefern ein sprachlicher Wechsel innerhalb eines Dokumentes erfasst wird, sollte eine automatische Spracherkennung der Ausgangssprache stattfinden. Dabei können verschiedenste (theoretisch: überabzählbar viele) Fälle auftreten, unter welchen z.B. Wechsel aus dem Deutschen in das Englische an beliebieger Stelle im Dokument, satzweisige Wechsel zwischen zwei und mehreren Sprachen, sowie ein nur kurzfristiger Wechsel in eine Sprache, innerhalb eines ansonstig monolingualen Dokumentes, welche allerdings Lexeme dieser beinhält (bspw.: ein norwegisches Dokument beinhält ein dänisches Zitat).

3 Eigenständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt und ohne fremde Hilfe verfasst habe. Dazu habe ich keine außer den von mir angegebenen Hilfsmitteln und Quellen verwendet und die den benutzten Werken inhaltlich und wörtlich entnommenen Stellen habe ich als solche kenntlich gemacht. Ich versichere, dass die eingereichte elektronische Fassung mit den gedruckten Exemplaren übereinstimmt.