Mittelstand-Digital
Zentrum
Franken

Digitalisierungsprojekt

EXPLORATIVE ANALYSE VON METHODEN UND ALGORITHMEN ZUR QUALITÄTSBEWERTUNG MASCHINELL ÜBERSETZTER TEXTE

Mittelstand-
Digital

Gefördert durch:
Bundesministerium
für Wirtschaft
und Klimaschutz
aufgrund eines Beschlusses
des Deutschen Bundestages

Inhalt

Abbildungsverzeichnis

1. Einleitung

In der heutigen globalisierten Welt spielen maschinelle Übersetzungssysteme eine zentrale Rolle bei der internationalen Kommunikation und diese hat in den letzten Jahren bemerkenswerte Fortschritte gemacht, vor allem durch die Weiterentwicklung und Anwendung moderner Algorithmen des maschinellen Lernens und der künstlichen Intelligenz. Trotz dieser Fortschritte bleibt die Qualität maschinell übersetzter Texte ein zentrales Anliegen, insbesondere in Bereichen, in denen präzise und verständliche Übersetzungen von entscheidender Bedeutung sind. Die Überprüfung und Bewertung der Qualität dieser Übersetzungen ist daher ein wesentlicher Bestandteil der Weiterentwicklung und Optimierung von Übersetzungsalgorithmen.

Abbildung 1: Analyse von Texten (AI generiert mit Adobe Firefly)

Bereits seit langer Zeit beschäftigt sich die Linguistik mit Methoden zur Evaluierung von Übersetzungen. Diese Methoden, die auf soliden theoretischen und empirischen Grundlagen basieren, haben sich als äußerst effektiv erwiesen, um die Genauigkeit, Konsistenz und inhaltliche Übereinstimmung von Übersetzungen zu bewerten. Im Zuge der technologischen Entwicklung und der zunehmenden Bedeutung maschineller Übersetzungen stellt sich nun die Aufgabe, diese bewährten linguistischen Methoden in den maschinellen Bereich zu übertragen, um eine effiziente und zuverlässige Qualitätskontrolle zu gewährleisten. Durch den Einsatz moderner Techniken der natürlichen Sprachverarbeitung (Natural Language Processing, NLP) und speziell trainierter Modelle wie BERT können tiefgehende Analysen der Übersetzungsqualität durchgeführt werden. Diese Studie konzentriert sich auf die Analyse englischer Texte, die ins Deutsche übersetzt wurden, und untersucht, inwieweit diese Übersetzungen inhaltlich konsistent und dem Originaltext ähnlich sind.

2. Ziele und Aufgabenstellung

Wie die Graphik veranschaulicht, werden Texte eher in der lokalen Sprache gelesen als in Englischer Sprache. Die Texte werden in Englisch verfasst und daraufhin in 38 verschiedene Sprachen zur internationelen Kommunikation übersetzt. Um das Vertrauen in den Inhalt der Texte zu stärken und Missverständnissen vorzubeugen ist daher eine hochwertige Übersetzung unerlässlich.

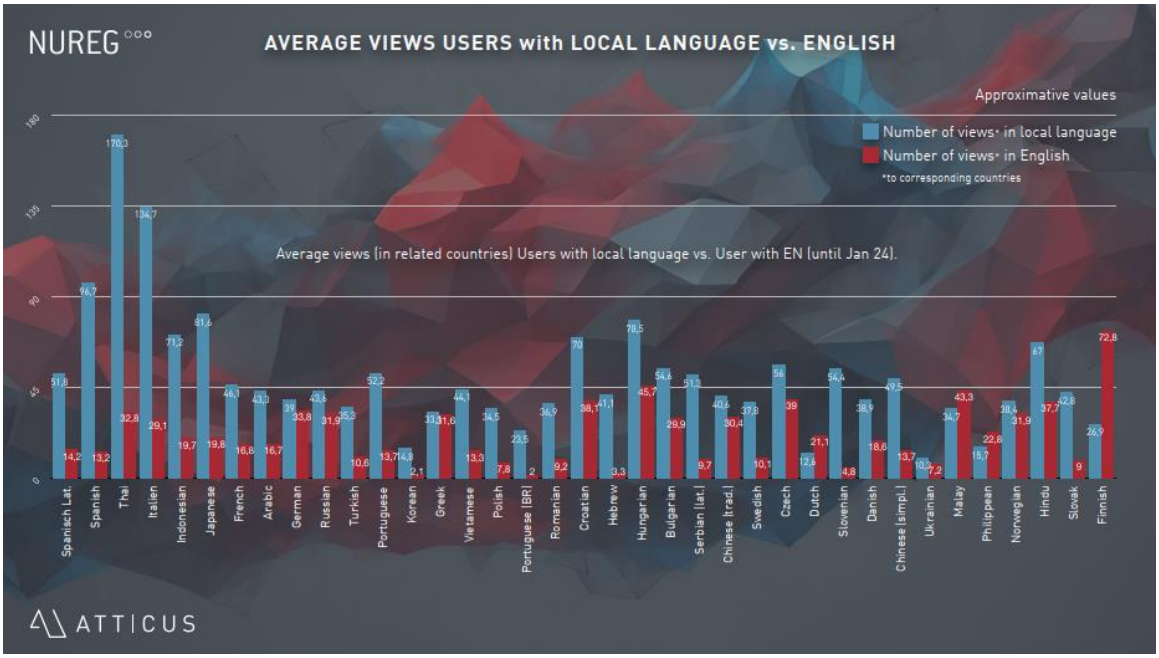

Abbildung 2: Ansichten der Texte in der Ursprungs- vs. Übersetzen Sprache (Nureg GmbH, 2023)

Im Erstgespräch mit der Nureg GmbH ergaben sich folgende Fragestellungen:

1. Gibt es Evaluierungsmöglichkeiten, um die Qualität der Übersetzungen zu prüfen?
2. Welche beweiskräftigen Kennzahlen oder wahrgenommene Bewertungen der Lernzufriedenheit gibt es?
3. Welche Sprachen sind gut und welche können nicht verwendet werden?
4. Welche Sprachen werden am meisten benötigt, weshalb ein größerer Fokus daraufgelegt werden muss?
5. Lassen sich Muster in den Sprachen erkennen, welche systematisch immer wieder nicht gut genug übersetzt werden und wie konsistent sind die Übersetzungen?

Das Hauptziel dieses Digitalisierungsprojekts ist es, die Effektivität verschiedener Methoden und Algorithmen zur Qualitätsbewertung maschinell übersetzter Texte zu untersuchen und zu analysieren. Durch eine explorative Analyse sollen die Stärken und Schwächen der verschiedenen Ansätze identifiziert und verstanden werden.

Zieldefinition:

„Explorative Analyse von Methoden und Algorithmen zur Qualitätsbewertung maschinell übersetzter Texte“

Im Einzelnen verfolgt das Projekt folgende spezifische Ziele:

1. **Literaturrecherche**: Eine Erhebung bestehender Methoden und Algorithmen zur Qualitätsbewertung von maschinellen Übersetzungen.
2. **Auswahl und Implementierung**: Die Auswahl der relevantesten Methoden sowie deren Implementierung, um praktische Tests durchführen zu können.
3. **Test und Vergleich**: Durchführung von Tests anhand eines spezifischen Datensatzes maschinell übersetzter Texte und Vergleich der verschiedenen Methoden hinsichtlich ihrer Effektivität.
4. **Ergebnisanalyse**: Analyse und Gegenüberstellung der Ergebnisse zur Identifikation der Stärken und Schwächen der einzelnen Methoden.

Die Ergebnisse dieses Projekts sollen ein fundiertes Verständnis der Effektivität unterschiedlicher Bewertungsmethoden bieten und aufzeigen, welche Ansätze für die Qualitätsbewertung maschineller Übersetzungen besonders geeignet sind. Alle Schritte und Ergebnisse des Projekts wurden systematisch in einem GitHub-Repository dokumentiert, um die Nachvollziehbarkeit und Reproduzierbarkeit der Forschung sicherzustellen.

3. Organisation und Ablauf des Projekts

Um die erfolgreiche Umsetzung dieses Projekts zu gewährleisten, ist eine klare und strukturierte Vorgehensweise notwendig. In diesem Abschnitt werden die organisatorischen Aspekte des Projekts vorgestellt, um ein Verständnis für den Ablauf und die beteiligten Akteure und die Organisation zu schaffen.

3.1. Beteiligte

Ein wesentlicher Faktor für den Projekterfolg ist die Zusammenarbeit der beteiligten Akteure. In diesem Abschnitt werden die wichtigsten Beteiligten des Projekts vorgestellt:

Alexander Roth

Nureg GmbH

IT-Projektmanager

Projektorganisation und -überwachung. Bindeglied zwischen der Organisatorischen und fachlichen Bereichen. Vorstellung des Istzustandes sowie der Anforderungen und Ziele des Projekts.

Jan Welslau

Nureg GmbH

Intermodal Transport Control System

Fachlicher Spezialist für die Softwareentwicklung der Atticus 2.1 zur „Advances Text und Translation Handling“ (Firmeninternen Softwarelösung zur Übersetzung)

Sandra Nuißl

Mittelstand Digitalzentrum Franken

AI Strategist und Research Asisstant

Wissenschaftliche Mitarbeiterin mit dem Schwerpunkt LLMs. Zuständig für die Recherche und Ausarbeitung der Evaluierung von Methoden und Algorithmen im Rahmen des Projekts.

3.2. Roadmap

Die Projekt-Roadmap diente als Fahrplan für die vergangenen Monate und gibt einen Überblick über die Meilensteine und Arbeitspakete. Sie strukturiert den zeitlichen Ablauf und stellt sicher, dass alle Beteiligten über den Fortschritt informiert sind:

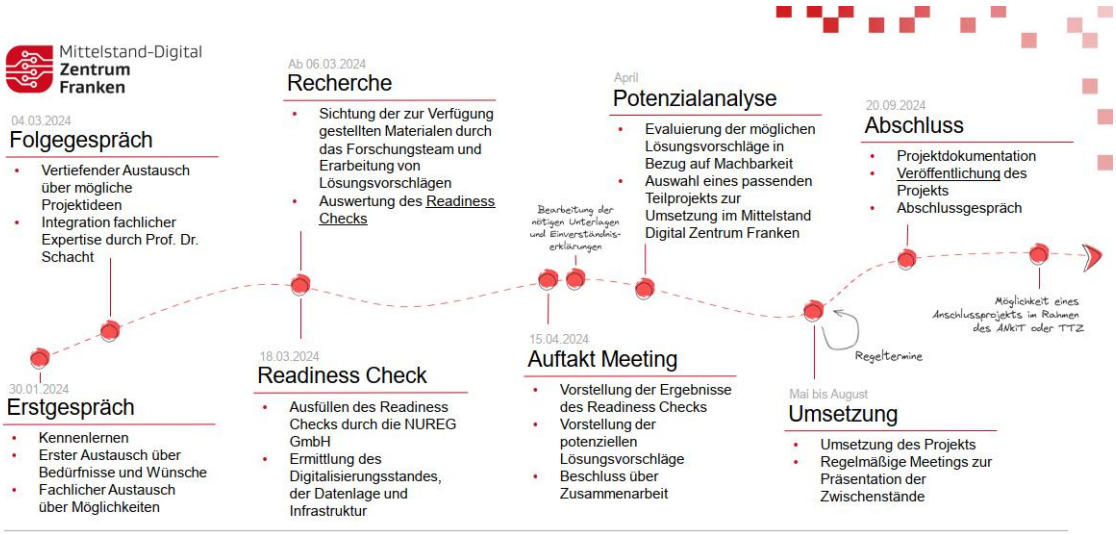

Abbildung 3: Roadmap für den Projektablauf (eigene Darstellung)

Das Projekt startete am 30. Januar 2024 mit einem Erstgespräch, bei dem sich alle Beteiligten kennenlernten und über ihre Bedürfnisse und Wünsche austauschten. In diesem ersten Treffen wurden außerdem mögliche Ansätze und Wege für eine Zusammenarbeit erörtert. Am 4. März 2024 folgte ein weiteres Gespräch, in welchem die zuvor besprochenen Projektideen vertieft wurden. Dabei wird fachspezifische Expertise eingebunden, unter anderem durch Prof. Dr. Sigurd Schacht, um die Ideen weiter zu konkretisieren und den Projektrahmen zu schärfen. Ab dem 6. März 2024 begann die Recherchephase. Hier sichtete das Forschungsteam die bereitgestellten Materialien und entwickelt basierend auf den Ergebnissen des sogenannten „Readiness Checks“ konkrete Lösungsvorschläge. Der Readiness Check selbst fand am 18. März 2024 statt und wurde von der NUREG GmbH durchgeführt. In diesem Test wird der aktuelle Stand der Digitalisierung des Projektpartners analysiert, wobei besonders die vorhandene Datenlage und Infrastruktur im Fokus standen. Am 15. April 2024 wurde im Auftaktmeeting die Ergebnisse des Readiness Checks besprochen. und eine Entscheidung über die weitere Zusammenarbeit getroffen. Die Umsetzung dieses Projekts erfolgte in den Monaten von Mai bis August 2024. In dieser Zeit fanden regelmäßige Meetings statt, um den Fortschritt zu besprechen und Zwischenergebnisse zu präsentieren.

Das Projekt wurde schließlich am 20. September 2024 mit der Abschlussphase beendet. Diese umfasst die Dokumentation und Veröffentlichung der Projektergebnisse sowie ein abschließendes Gespräch. Es besteht zudem die Möglichkeit, das Projekt im Rahmen des AMKT oder TTZ weiterzuführen.

3.3. Readiness Check

Um die Effektivität der implementierten Methoden frühzeitig zu bewerten, wurde ein initialer Readiness Check durchgeführt. Dieser Test dient dazu, potenzielle Probleme frühzeitig zu identifizieren und notwendige Anpassungen vorzunehmen. Der Test wurde speziell für mittelständische Unternehmen entwickelt und ist eine gute Grundlage, um weitere Angebote des Mittelstand Digital Zentrums Franken gezielt auszuwählen. Im Test werden fünf Dimensionen betrachtet, die für die erfolgreiche Umsetzung des digitalen Transformationsprozesses relevant sind:

Abbildung 4: Fünf Dimensionen des Reifegrads (eigene Darstelleung in Anlehnung an (Hellge, 2024))

Anhand von 25 Fragen und allgemeinen Angaben zum Unternehmen wurde folgender Reifegrad ermittelt, welcher in der Abbildung 5 visualisiert wurde:

Es ist zu erkennen, dass der Bereich „Strategie" den höchsten Reifegrad aufweist, gefolgt von den Bereichen „Organisation und Prozesse", „Technologien" und „Mitarbeiter". Der Bereich „Produkt und Dienstleistung" des Unternehmens hingegen bietet noch große Potenziale für Digitalisierung, Automatisierung und künstliche Intelligenz, bei welchem wir im Rahmen ansetzen werden und neue Möglichkeiten zur maschinellen Prüfung der Übersetzungsqualität evaluieren werden.

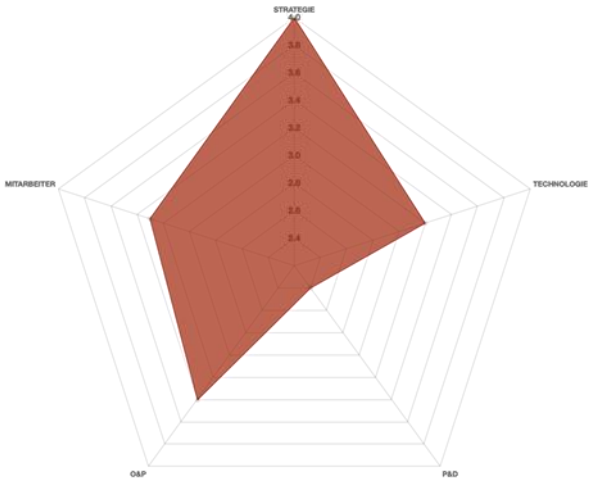

Abbildung 5: Ergebnisse des Readinesschecks für die Nureg GmbH

4. Rahmenbedingungen

Die Grundlage für die Bewertung der maschinell übersetzten Texte bildet eine sorgfältige Datenvorbereitung. In diesem Kapitel wird der Prozess der Datenvorbereitung beschrieben, der entscheidend dafür ist, dass alle erforderlichen Informationen korrekt und vollständig vorliegen. Zunächst wird die Struktur und der Inhalt der eingehenden JSON-Datei analysiert, die die ursprünglichen Texte sowie deren Übersetzungen in verschiedenen Zielsprachen enthält. Durch die Transformation dieser Daten in ein strukturiertes Format wird sichergestellt, dass sie für die nachfolgenden Analysen optimal nutzbar sind. Im Anschluss werden eine Reihe von Qualitätskriterien definiert, anhand derer im Folgenden die Qualitätsprüfung durchgeführt wird:

- **Kohärenz**: Wie gut die Bedeutung des ursprünglichen Textes beibehalten wurde.
- **Lesbarkeit**: Ob der übersetzte Text flüssig und natürlich klingt.
- **Grammatik**: Satzbau, Zeichensetzung und der Gleichen.
- **Kontext**: Ob der Kontext des Textes richtig interpretiert wurde.
- **Sinnhaftigkeit**: Ob der übersetzte Text inhaltlich Sinn ergibt.

4.1. Datenstruktur verstehen

In der Datenvorbereitung wurde die JSON-Datei mit den Ausgangstexten und ihren Übersetzungen in mehrere Sprachen mithilfe eines Python-Skripts eingelesen. Der Inhalt wurde anschließend in ein Python-Objekt und schließlich in ein DataFrame-Format überführt, um die Daten besser handhabbar zu machen. Durch die Nutzung der `pandas`-Bibliothek lagen die Daten nun in tabellarischer Form vor, was die weitere Analyse und Bewertung der Übersetzungsqualität erleichterte. Nachdem die Datei erfolgreich geladen war, bestand der nächste Schritt darin, die Datenstruktur zu analysieren. Es wurde geprüft, wie die Ausgangstexte und deren Übersetzungen organisiert sind. Die Ergebnisse dieser Analyse lassen sich in dem Notebook „Evaluation_Dataset“ nachlesen. Die JSON-Datei enthielt für jeden Ausgangstext eine Liste von Übersetzungen, wobei jede Übersetzung durch die Zielsprache und den übersetzten Text gekennzeichnet war.

Die JSON-Datei ist ein Array von Objekten, wobei jedes Objekt folgende Felder enthält:

- SourceLanguage: Die Ausgangssprache des Textes (z. B. "en-US" für Englisch, USA).
- Text: Der Originaltext in der Ausgangssprache.
- Translations: Ein Array von Objekten, die jeweils eine Übersetzung in eine bestimmte Zielsprache enthalten. Jedes Objekt im Translations-Array umfasst:
    - Language: Die Zielsprache der Übersetzung (z. B. "de-DE" für Deutsch, Deutschland).
    - Text: Den übersetzten Text in der Zielsprache.

4.2. Qualitätskriterien definieren

In diesem Abschnitt werden die verschiedenen Qualitätskriterien definiert, die für die Bewertung der maschinell übersetzten Texte herangezogen wurden. Die fünf Säulen der Textbewertung – Genauigkeit, Klarheit, Kohärenz, Konsistenz und Angemessenheit – stammen aus der Linguistik und dienen als bewährte Kriterien zur Beurteilung der Qualität von Texten. Diese Konzepte sind tief in der Sprachwissenschaft verwurzelt und ermöglichen eine systematische Analyse der sprachlichen und strukturellen Aspekte von Texten. Indem wir diese Säulen auf die maschinelle Übersetzung anwenden, wird eine interdisziplinäre Brücke zwischen der traditionellen Linguistik und der modernen Technologie geschlagen. (Kuhn, 2023)

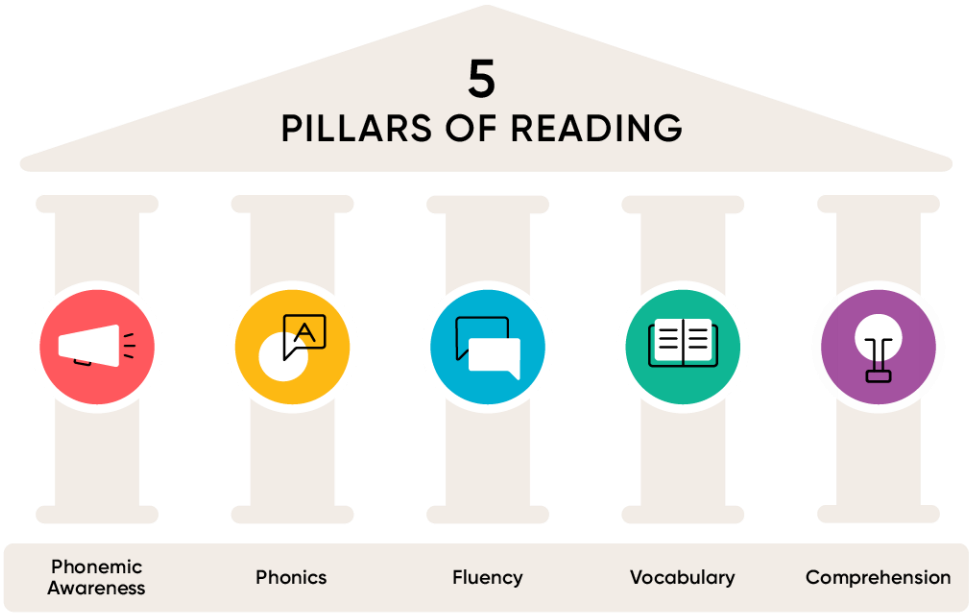

Abbildung 6: 5 Säulen des Lesens (Kuhn, 2023)

In der maschinellen Evaluierung von Übersetzungstexten bieten diese Kriterien einen robusten Rahmen, um die Qualität maschinell erzeugter Texte objektiv zu beurteilen. Dadurch wird es möglich, sowohl die sprachliche Korrektheit als auch die kommunikative Effektivität von maschinellen Übersetzungen fundiert zu analysieren und weiterzuentwickeln. Für den weiteren Projektverlauf wurden die folgenden Aspekte der fünf Säulen ausgewählt:

- Similarity Score
- Readability Score
- Grammatikalische Prüfung
- Schwierige Wörter und Eigennamen

4.2.1. Similarity Score

Die Kohärenz ist entscheidend für eine hochwertige Übersetzung, da sie sicherstellt, dass der Text inhaltlich logisch bleibt. Die Semantik bewertet, wie gut die Bedeutung des Originaltextes in der Übersetzung erhalten wurde, einschließlich der korrekten Wiedergabe von Begriffen. Um die semantische Ähnlichkeit maschinell zu prüfen, kommen Techniken wie der Cosinus-Similarity-Score zum Einsatz, die die inhaltliche Nähe zwischen Original und Übersetzung messen. Diese Methode hilft sicherzustellen, dass die Übersetzung nicht nur wörtlich, sondern auch inhaltlich präzise ist. Eine sinnvolle Übersetzung muss den Textzusammenhang bewahren, um logische und kohärente Aussagen zu gewährleisten. Hierbei unterstützen NLP-Algorithmen, die Kontext und Kohärenz in der Übersetzung analysieren und optimieren.

4.2.2. Readability

Die Lesbarkeit eines Textes ist ein entscheidender Faktor für seine Gesamtqualität. Besonders bei übersetzten Texten ist es wichtig, dass sie flüssig und natürlich klingen, damit sie für Muttersprachler mühelos verständlich sind. Hierbei kommen Lesbarkeitsmetriken wie der Flesch-Reading-Ease-Score zum Einsatz. Diese Metriken ermöglichen eine quantitative Bewertung der Lesbarkeit, indem sie Faktoren wie Satzlänge und Wortkomplexität berücksichtigen. Der Flesch-Reading-Ease-Score etwa liefert einen objektiven Wert, der angibt, wie einfach oder schwierig ein Text zu lesen ist. Eine gute Übersetzung sollte demnach nicht nur inhaltlich korrekt sein, sondern auch angenehm und leicht verständlich, um dem Leser ein optimales Leseerlebnis zu bieten.

4.2.3. Grammatikalische Prüfung

Grammatikalische Fehler in der Übersetzung können auftreten, wenn Wörter oder Sätze nicht korrekt übersetzt werden und dadurch die ursprüngliche Bedeutung verändert oder verzerrt wird. Solche Fehler können die Verständlichkeit und Glaubwürdigkeit des Textes erheblich beeinträchtigen. Zur Identifikation dieser Übersetzungsfehler stehen verschiedene Methoden zur Verfügung, darunter auch automatisierte Fehlererkennungstools wie LanguageTool. Das Hauptziel besteht darin, Übersetzungen zu liefern, die sowohl grammatikalisch als auch lexikalisch fehlerfrei sind, um die Qualität und Genauigkeit des Textes zu gewährleisten.

4.2.4. Schwierige Wörter und Eigennamen

Schwierige Wörter und Eigennamen stellen oft eine besondere Herausforderung in der Übersetzung dar, da sie spezifische kulturelle oder kontextuelle Bedeutungen tragen, die nicht immer direkt übertragbar sind. Eine präzise Übersetzung erfordert daher eine sorgfältige Recherche und oft auch Anpassungen, um sicherzustellen, dass diese Begriffe im Zieltext korrekt und verständlich wiedergegeben werden.

5. Methoden und Algorithmen

Im Rahmen dieses Projekts wurden verschiedene Methoden und Algorithmen zur Bewertung der Qualität von maschinell übersetzten Texten untersucht. Dabei wurden insbesondere zwei Hauptansätze betrachtet: stringbasierte Metriken und Deep Learning Metriken. Alle Forschungsergebnisse sind ausführlich in den entsprechenden Notebooks im Research-Ordner des GitHub-Repositorys dokumentiert.

Zu diesem Zweck wurden die folgenden Metriken analysiert und anhand des Datensatzes getestet:

- BLEU (Bilingual Evaluation Understudy)
- chrF (Character F-score)
- BERTScore (Bidirectional Encoder Representations from Transformers)
- mBERT (multilingual BERT)
- COMET (Crosslingual Optimized Metric for Evaluation of Translation)
- GPT (Generative Pre-trained Transformer)

Stringbasierte Metriken, wie BLEU und chrF, bewerten Übersetzungen anhand des genauen Übereinstimmens von Wörtern oder Phrasen zwischen der generierten Übersetzung und Referenztexten. Diese Metriken zerlegen Texte in Tokens oder n-Gramme (Wortpaare oder -gruppen) und messen die Präzision sowie die Überlappung dieser Einheiten. Obwohl sie effizient und einfach zu berechnen sind, erfassen sie nicht immer die semantische Bedeutung oder den Kontext der Texte.

Im Gegensatz dazu nutzen Deep Learning Metriken wie BERTScore und mBERT die von vortrainierten Modellen wie BERT erzeugten Embeddings. Diese Modelle analysieren die semantische Ähnlichkeit zwischen Texten, indem sie die Bedeutung und den Kontext der Wörter berücksichtigen. Durch die Bewertung auf einer semantischen Ebene bieten sie oft genauere Einsichten in die Qualität von Übersetzungen, insbesondere in mehrsprachigen oder komplexen Kontexten.

Ein weiterer Unterschied liegt in der Flexibilität: Während stringbasierte Metriken oft für spezifische Aufgaben und Sprachen optimiert sind, können Deep Learning Metriken durch Transferlernen auf verschiedene Sprachen und Aufgaben angewendet werden. Allerdings erfordern sie häufig mehr Rechenleistung und Ressourcen für Training und Ausführung im Vergleich zu stringbasierten Ansätzen. Insgesamt bieten beide Ansätze unterschiedliche Stärken und Einsatzmöglichkeiten, abhängig von den spezifischen Anforderungen und Zielen der Evaluierung natürlicher Sprachverarbeitungssysteme. Die Wahl der geeigneten Metrik hängt von Faktoren wie Genauigkeit, Skalierbarkeit und verfügbaren Ressourcen ab.

5.1. BLEU (Bilingual Evaluation Understudy)

Der BLEU-Score, 2002 von Papineni et al. entwickelt, ist eine Metrik zur Bewertung maschineller Übersetzungen. Er misst die Qualität, indem er maschinelle Übersetzungen mit Referenzübersetzungen vergleicht und n-gram Übereinstimmungen analysiert. Die Formel lautet:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Hierbei steht "BP" für den Brevity Penalty, der kurze Übersetzungen bestraft, um eine ausgewogenere Bewertung zu gewährleisten. Der Score reflektiert, wie gut die maschinelle Übersetzung mit den Referenzübersetzungen übereinstimmt, wobei 1 eine perfekte Übereinstimmung und 0 keine Übereinstimmung bedeutet (Kishore Papineni, 2022).

In der Praxis werden Übersetzungen in tokenisierte Sätze zerlegt. Tools wie NLTK in Python bieten Funktionen wie sentence_bleu und corpus_bleu, um diese Token in numerische BLEU-Scores umzuwandeln. Der BLEU-Score bietet eine schnelle, quantitative Bewertung von Übersetzungen und ermöglicht den Vergleich unterschiedlicher Systeme. Jedoch konzentriert er sich auf die Wortübereinstimmung und kann komplexe semantische und grammatikalische Aspekte einer Übersetzung vernachlässigen. Auch ist eine menschliche Referenz für jede maschinelle Übersetzung erforderlich, was seine Anwendbarkeit einschränken kann (Vashee, 2019).

Die Vor- und Nachteile des BLEU Scores wurden in nachfolgender Tabelle zusammengetragen:

Vorteile	Nachteile
Einfache Implementierung: BLEU kann einfach implementiert werden und bietet schnell interpretierbare Ergebnisse.	Begrenzte Bewertung: BLEU bewertet hauptsächlich die Übereinstimmung von Wortsequenzen und vernachlässigt semantische oder grammatikalische Aspekte.
Quantitative Bewertung: Durch die numerische Bewertung zwischen 0 und 1 ermöglicht BLEU einen direkten Vergleich der Übersetzungsqualität verschiedener Systeme.	Fokus auf Wörter: Der Fokus liegt stark auf der Übereinstimmung von Wörtern und berücksichtigt nicht die semantische Tiefe oder syntaktische Genauigkeit der Übersetzung.
Breite Anwendung: Aufgrund seiner Einfachheit und Verfügbarkeit wird BLEU häufig in der Forschung und Evaluierung maschineller Übersetzungssysteme verwendet.	Einschränkung auf Referenzübersetzungen: BLEU setzt voraus, dass für jede zu bewertende Übersetzung mindestens eine menschliche Referenz verfügbar ist.
Standardisiertes Verfahren: Da BLEU weit verbreitet ist, ermöglicht es einen Standardisierungsrahmen für die Vergleichbarkeit von Übersetzungssystemen.	Brevity Penalty: Die Brevity Penalty kann zu unrealistischen Bewertungen führen, wenn die maschinelle Übersetzung deutlich länger ist als die Referenzübersetzung.

5.2. ChrF++

Der chrF-Score ist eine Metrik zur Bewertung maschineller Übersetzungen, die auf Zeichen-n-Grammen basiert und das harmonische Mittel aus Präzision und Recall verwendet. Im Gegensatz zu wortbasierten Metriken arbeitet chrF++ mit kurzen Zeichenfolgen, was seine Robustheit gegenüber Satztokenisierungen erhöht und eine teilweise Belohnung für falsch geschriebene Wörter ermöglicht. Die Formel lautet:

$$\mathrm{CHRF}\beta = (1+\beta^2)\frac{\mathrm{CHRP}\cdot\mathrm{CHRR}}{\beta^2\cdot\mathrm{CHRP}+\mathrm{CHRR}}$$

Dieser Score bewertet die Ähnlichkeit zwischen der maschinellen Übersetzung und den Referenzübersetzungen durch Zeichen-n-Gramme. Wesentliche Parameter wie char_order (Gewichtung der Zeichenreihenfolge), word_order (Gewichtung der Wortreihenfolge) und beta (Gewichtung der n-Gramme) spielen dabei eine zentrale Rolle. Ein höherer Beta-Wert legt mehr Gewicht auf größere n-Gramme, was nützlich ist, um bestimmte Aspekte einer Übersetzung besonders hervorzuheben (Popović, 2015). chrF wird üblicherweise mit spezialisierten Bibliotheken oder eigenen Skripten implementiert, die auf den Prinzipien der Zeichen-n-Gramm-Metriken basieren. Zur Berechnung des Scores werden zunächst die maschinellen Übersetzungen (Hypothesen) und die Referenzübersetzungen in Zeichenfolgen umgewandelt und tokenisiert. Anschließend wird das chrF Modell oder die entsprechende Bibliothek geladen und konfiguriert, um die spezifischen Parameter wie char_order, word_order und beta zu berücksichtigen. Nachdem die Daten vorbereitet sind, wird der Score berechnet, der üblicherweise in Prozent angegeben wird. Ein höherer chrF-Score deutet auf eine größere Ähnlichkeit und damit möglicherweise auf eine bessere Übersetzungsqualität hin (huggingface, kein Datum).

chrF bietet den Vorteil einer robusten Leistung gegenüber Tokenisierungsfehlern und belohnt teilweise falsch geschriebene Wörter. Es ist jedoch weniger empfindlich gegenüber semantischen und syntaktischen Aspekten als wortbasierte Metriken wie BLEU und kann durch die Abhängigkeit von spezifischen Referenzübersetzungen in unterschiedlichen Domänen eingeschränkt werden.

Die Vor- und Nachteile des chrF Scores wurden in nachfolgender Tabelle zusammengetragen:

Vorteile	Nachteile
Robustheit gegenüber Tokenisierungsfehlern	Begrenzte Berücksichtigung semantischer und syntaktischer Aspekte
Teilweise Belohnung für falsch geschriebene Wörter	Abhängigkeit von spezifischen Referenzübersetzungen
Einfache Integration in bestehende Arbeitsabläufe	Begrenzte Berücksichtigung semantischer und syntaktischer Aspekte

5.3. Bert Score

Der BERTScore, eingeführt im Jahr 2020, verwendet die kontextabhängigen Einbettungen des BERT-Modells, um die Qualität maschineller Übersetzungen präziser zu bewerten. Im Gegensatz zu traditionellen Metriken wie BLEU berücksichtigt BERTScore die semantische Ähnlichkeit und den Kontext der Texte durch die Berechnung der Kosinusähnlichkeit der Einbettungen. Dies ermöglicht eine genauere Beurteilung der Übersetzungsqualität und ist besonders wertvoll in der maschinellen Übersetzung und anderen Anwendungen der natürlichen Sprachverarbeitung (Sellam, Das, & Parikh, 2020).

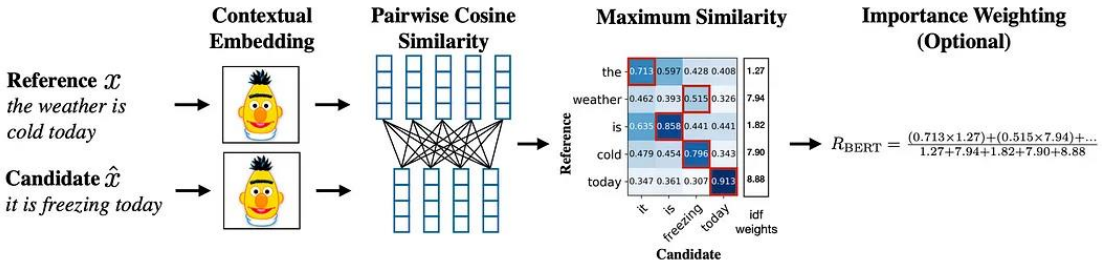

Abbildung 7: BERTScore (Zhang, Kishore, Wu, & Weinberger, 2020)

Der BERTScore nutzt die kontextuellen Einbettungen eines vorab trainierten BERT-Modells zur Bewertung von Übersetzungen. Die Ähnlichkeit zwischen den maschinellen Übersetzungen und den Referenztexten wird durch Kosinusähnlichkeit, gegebenenfalls gewichtet mit inversen Dokumenthäufigkeitswerten, berechnet. Die Metrik liefert drei Hauptkennzahlen zur Bewertung der Übersetzungsqualität:

- Precision (Präzision): Das Verhältnis der korrekten Übersetzungen zu allen vorgenommenen Übersetzungen. Eine hohe Präzision zeigt, dass die meisten Übersetzungen korrekt waren.
- Recall (Erkennungsrate): Das Verhältnis der korrekten Übersetzungen zu allen möglichen korrekten Übersetzungen. Ein hoher Recall bedeutet, dass die meisten korrekten Übersetzungen erkannt wurden.
- F1-Score: Das harmonische Mittel von Precision und Recall, das eine kombinierte Bewertung der Übersetzungsqualität bietet. Ein hoher F1-Score weist auf hohe Werte sowohl für Präzision als auch für Recall hin (Zhang, Kishore, Wu, & Weinberger, 2020).

Die Implementierung des BERTScores erfolgt typischerweise über spezialisierte Bibliotheken oder eigene Skripte, die auf BERT-Modellen basieren. Eine dieser Library ist die sogenannte Textstat Library, welche sich vor allem mit der Readability eines Textes befasst. Zunächst werden die zu bewertenden Übersetzungen und die Referenztexte vorbereitet und in ein BERT-Modell eingegeben, das dann die Metrik berechnet. Das Ergebnis wird üblicherweise als prozentualer Score ausgegeben, der die Ähnlichkeit der Übersetzungen zu den Referenztexten angibt.

Der BERTScore bietet eine moderne und präzise Methode zur Bewertung maschineller Übersetzungen, indem er die Kontextualisierung der Wörter durch BERT nutzt. Er eignet sich besonders gut für Szenarien, in denen semantische Genauigkeit und Kontextualisierung von entscheidender Bedeutung sind. Allerdings erfordert die Nutzung von BERTScore erhebliche Rechenressourcen und kann für einfachere Anwendungen überdimensioniert sein.

Die Vor- und Nachteile des BERTScores wurden in nachfolgender Tabelle zusammengetragen:

Vorteile	Nachteile
Verwendung von kontextbezogenen Einbettungen	Erfordert umfangreiche Rechenressourcen
Berücksichtigung von semantischen Zusammenhängen	Komplexere Implementierung im Vergleich zu einfachen Wort- oder Zeichen-basierten Metriken
Gute Performance bei komplexen Übersetzungen	Kann anfällig für Rauschen oder irrelevante Details in den Embeddings sein
Flexibilität durch parametrisierte Gewichtung	

5.4. mBert

mBERT (Multilingual BERT) ist eine Weiterentwicklung von BERT, die speziell für die Verarbeitung und das Verständnis mehrsprachiger Texte entwickelt wurde. Im Gegensatz zu sprachspezifischen Modellen kann mBERT mehrere Sprachen ohne separate Trainingsphasen oder Modelle bewältigen. Auf Basis der Transformer-Architektur nutzt mBERT bidirektionale Encoder, um den Kontext vor und nach jedem Wort zu erfassen, was ein tieferes Verständnis des Textes ermöglicht. Das Modell wird auf multilingualen Daten trainiert und verwendet kontextbezogene Einbettungen, um Texte in numerische Vektoren umzuwandeln. Die semantische Ähnlichkeit zwischen Texten wird durch die Kosinusähnlichkeit dieser Vektoren gemessen. Zur Nutzung von mBERT für die Textähnlichkeitsmessung wird zunächst das Modell zusammen mit dem Tokenizer aus der Transformers-Bibliothek geladen. Die Testdaten werden tokenisiert und in das Modell eingespeist, um die Embeddings zu berechnen. Der Cosine Similarity Score zwischen den Embeddings der Texte gibt die semantische Ähnlichkeit an. Ein höherer Score deutet auf eine größere Ähnlichkeit hin und reflektiert die Qualität der Übersetzung oder den Grad der Textübereinstimmung (Liu, et al., 2019).

Die Vor- und Nachteile des mBERT Scores wurden in nachfolgender Tabelle zusammengetragen:

Vorteile	Nachteile
Multilinguales Training ermöglicht Nutzung in verschiedenen Sprachen ohne separate Modelle.	Möglicher Verlust an Präzision für spezifische Sprachstrukturen.
Bidirektionales Encoding erfasst umfassenden Kontext vor und nach jedem Wort.	Ressourcenintensiv in Bezug auf Speicher und Rechenleistung.
Transferlernen ermöglicht Anwendung in verschiedenen multilingualen Umgebungen.	Kann Schwierigkeiten haben, seltene Sprachen oder spezifische Domänen gut zu unterstützen.

5.5. COMET

COMET (Cross-lingual Optimized Metric for Evaluation of Translation) ist eine Metrik zur Bewertung der Qualität maschineller Übersetzungen, die 2020 von Rei et al. entwickelt wurde. Die Methode zielt darauf ab, die Bewertung der Übersetzungsqualität kontextsensitiver zu gestalten, indem es die semantische Ähnlichkeit zwischen der maschinellen und der Referenzübersetzung erfasst. Dabei werden tiefergehende Sprachverständnisfähigkeiten genutzt, um die Qualität der Übersetzung besser zu erfassen. COMET verwendet sogenannte "Transformationen" auf der Basis von vortrainierten Sprachmodellen wie BERT oder RoBERTa, um kontextuelle Informationen zu extrahieren. Diese Informationen werden verwendet, um einen numerischen Score zu berechnen, der die Übersetzungsqualität zwischen 0 und 1 darstellt, wobei 1 eine perfekte und 0 eine unzulängliche Übersetzung bedeutet (Rei et al., 2020).

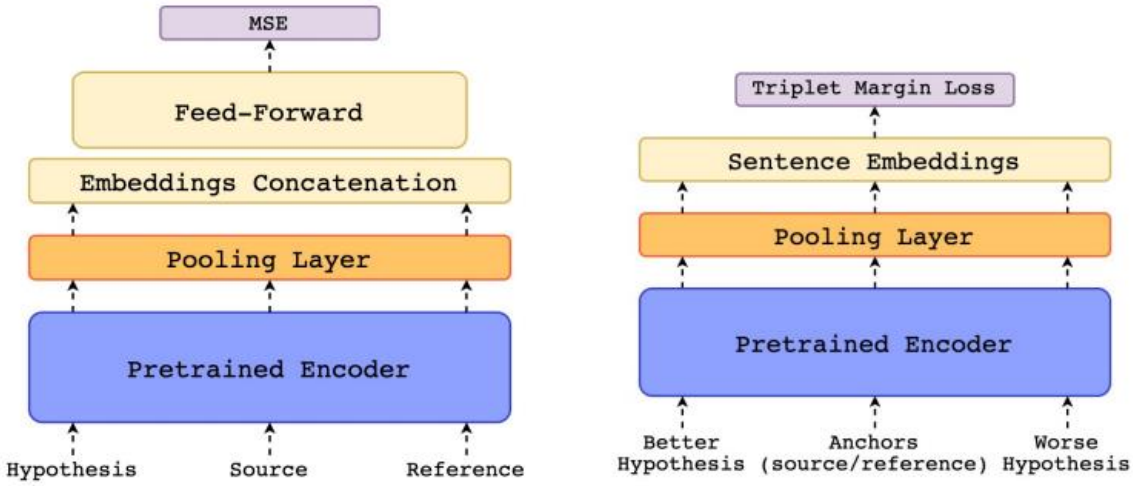

Abbildung 8: Schätzmodell optimiert MSE; Ranking-Modell nutzt Triplet Margin Loss (Rei, Stewart, Farinha, & Lavie, 2020)

In der Praxis wird COMET häufig in Kombination mit anderen Metriken eingesetzt, um eine umfassendere Bewertung von Übersetzungen zu erhalten. Die Implementierung kann über vorgefertigte Pakete und Bibliotheken erfolgen, die eine einfache Integration in bestehende Übersetzungssysteme ermöglichen (Rei, Stewart, Farinha, & Lavie, 2020).

Die Vor- und Nachteile von COMET werden in der folgenden Tabelle zusammengefasst:

Vorteile	Nachteile
Tiefe semantische Analyse durch neuronale Modelle, für semantische Qualitätsbewertung	Die Berechnung des COMET-Scores ist sehr rechenintensiv
Die Metrik berücksichtigt den Kontext der Wörter und Sätze, was zu präziseren Bewertungen führt.	COMET benötigt vortrainierte Sprachmodelle, die nicht immer für alle Sprachen verfügbar sind.
COMET kann für verschiedene Sprachpaare und Übersetzungsdomänen angepasst werden.	Die Resultate sind weniger intuitiv interpretierbar als Wortübereinstimmungsmetriken wie BLEU.

5.6. GPT-basierte Bewertungsmethoden

GPT (Generative Pre-trained Transformer), insbesondere die neueren Versionen wie GPT-3 und GPT-4, bieten eine neuartige Herangehensweise zur Bewertung der Qualität von maschinellen Übersetzungen. Diese Modelle nutzen fortschrittliche Ansätze des natürlichen Sprachverständnisses, um die Qualität von Übersetzungen zu bewerten, indem sie kontextuelle und semantische Informationen berücksichtigen und so eine differenziertere Bewertung liefern. Für die Analyse der Übersetzungsqualität wird, wie bei BERT, die semantische Ähnlichkeit mithilfe des Similarity Scores gemessen. Dabei kommt jedoch das Modell "text-embedding-ada-002" von OpenAI zum Einsatz. Dieses Modell ist speziell darauf trainiert, Texte in numerische Vektoren (Embeddings) zu transformieren. Diese Embeddings repräsentieren den semantischen Inhalt eines Textes und ermöglichen es, die Bedeutung von Texten zu vergleichen und zu analysieren (Open AI, 2022).

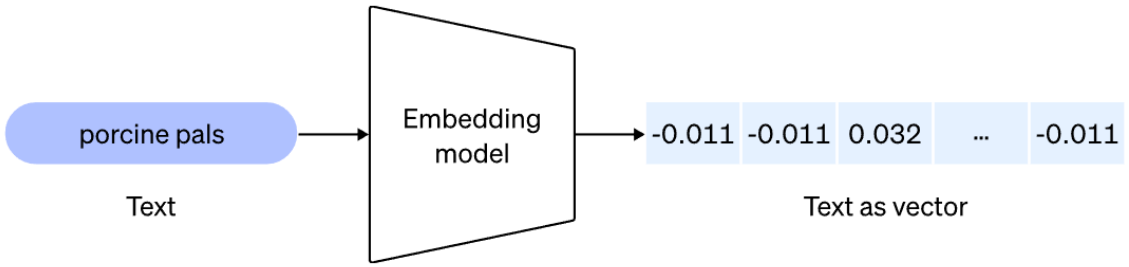

Abbildung 9: Embedding Model von Open AI (Open AI, 2022)

GPT-Modelle bewerten Übersetzungen direkt, indem sie einen Score vergeben, der die Qualität im Vergleich zu typischen Texten der Zielsprache angibt. Alternativ können sie die Übersetzung im Vergleich zu Referenzübersetzungen bewerten und dabei semantische Ähnlichkeiten und Unterschiede erkennen.

Die Vor- und Nachteile von GPT werden in der folgenden Tabelle zusammengefasst:

Vorteile	Nachteile
OpenAI's text-embedding-ada-002 Modell liefert hochwertige Embeddings, die eine präzise Berechnung der semantischen Ähnlichkeit ermöglichen	Obwohl der Code flexibel ist, könnte es schwierig sein, spezifische Anpassungen vorzunehmen, die über das hinausgehen, was die OpenAI API standardmäßig bietet.
Der Code ist einfach und direkt. Er nutzt die OpenAI API und die cosine_similarity Funktion aus scikit-learn	Der Code ist auf die Verfügbarkeit und Zuverlässigkeit der OpenAI API angewiesen.
Der Code ist leicht anpassbar für die Integration zusätzliche Textvorverarbeitungen	Texte müssen an die OpenAI-Server gesendet werden, was Datenschutzbedenken hervorrufen könnte, insbesondere bei sensiblen Daten
	Die Verwendung der OpenAI API ist kostenpflichtig.

6. Ergebnisse

Die umfassende Recherche zur Evaluierung der Qualität maschinell übersetzter Texte führte zu der Übersicht in Abbildung 10. Diese verdeutlicht, dass die Qualitätsprüfung in drei Hauptbereiche unterteilt werden kann: „Lesbarkeit“, „Ähnlichkeit“ und „Grammatik“. Für jeden dieser Bereiche wurden spezifische Evaluierungsalgorithmen entwickelt und anhand des Datensatzes getestet.

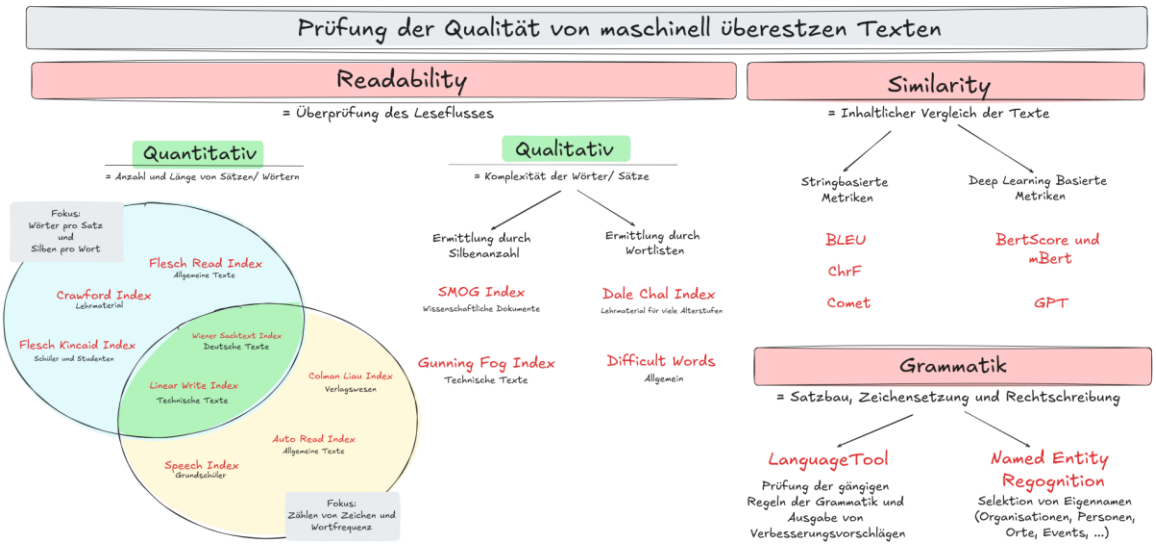

Abbildung 10: Übersicht der Recherche Ergebnisse in Bezug auf die Evaluierung der Übersetzungsqualität (eigene Darstellung)

Basierend auf der Recherche und Evaluierung wurden die am besten geeigneten Metriken und Algorithmen in dem Jupyter Notebook „Qualitätsprüfung“ zusammengefasst. Dazu gehören folgende Metriken:

- Konsistenz der Übersetzungen
- Similarity Score
- Readability Score
- Grammatikalische Prüfung
- Named Entity Recognition
- Reading Time

Der verwendete Datensatz umfasste 617.769 Texte, von denen 16.356 auf Deutsch vorlagen. Nach der Entfernung von Duplikaten blieben 7.630 einzigartige Texte übrig. Der Schwerpunkt lag auf den deutschen Übersetzungen der englischen Originaltexte, um die Ergebnisse menschlich überprüfen zu können und die Qualität der Metriken zu evaluieren.

6.1. Konsistenz der Übersetzungen

Ein Bestandteil dieser Arbeit war die Überprüfung der Konsistenz der Übersetzungen. Hierbei wurde untersucht, ob identische englische Texte auch immer identisch ins Deutsche übersetzt wurden. Hierfür wurden die Duplikate in den Originaltexten identifiziert und deren entsprechende Übersetzungen überprüft. Es stellte sich heraus, dass nicht alle Duplikate im Quelltext auch konsistente Übersetzungen im Deutschen aufwiesen (Siehe Abbildung 11).

Out[5]:

	Source-Text	Translation	Language
1	Have a question or suggestion? Ask below!	Haben Sie eine Frage oder Anregung? Fragen Sie...	German
4	Have a question or suggestion? Ask below!	Have a question or suggestion? Ask below!	German
99	Have a question or suggestion? Ask below!	Haben Sie eine Frage oder einen Vorschlag? Fra...	German
151	Have a question or suggestion? Ask below!	Hast du eine Frage oder Anregung? Frage uns un...	German

Abbildung 11: Inkonsistenz in Übersetzungen (Github)

Solche Inkonsistenzen können zu Verständnisschwierigkeiten führen und die allgemeine Qualität der Übersetzungen beeinträchtigen. Der ursprüngliche DataFrame, welcher die Duplikate enthielt, umfasst 9.022 Zeilen. Innerhalb dieser Daten wurden 296 einzigartige englische Texte und 337 einzigartige deutsche Texte identifiziert. Dies bedeutet, dass 41 Texte nicht konsistent übersetzt wurden. Angesichts der Großen Datenmenge ist diese Ungenauigkeit 0,45% in einem akzeptablen Bereich.

6.2. Similarity Score

Die Similarity beschreibt die inhaltliche Übereinstimmung zwischen der Ausgangs- und der Zielsprache, insbesondere die semantische Ähnlichkeit zwischen Originaltexten und deren Übersetzungen. Ein zentraler Aspekt dieser Studie war die Untersuchung der inhaltlichen Übereinstimmung zwischen den Textpaaren, wobei ein vortrainiertes BERT-Modell eingesetzt wurde, das die Bedeutung von Texten tiefgreifend analysieren kann. Die Ähnlichkeit der Texte wurde durch den Vergleich der von BERT generierten Vektorrepräsentationen ermittelt. Diese Methode ermöglicht es, semantische Zusammenhänge zu erfassen, die über rein lexikalische Übereinstimmungen hinausgehen. Die Ergebnisse zeigten, in welchen Fällen die Übersetzungen möglicherweise nicht die vollständige Bedeutung des Originaltextes wiedergaben, was zu Fehlinterpretationen führen könnte. Solche Analysen sind besonders wertvoll, um die Qualität maschineller Übersetzungen auf einem tiefen semantischen Niveau zu bewerten.

Bei dieser Analyse ist zu beachten, dass bei längeren Texten die maximale Token-Sequenz überschritten werden kann, die auf 512 Token begrenzt ist. Insgesamt lagen 161 Texte über diesem Maximalwert und konnten daher nicht in die Auswertung einbezogen werden.

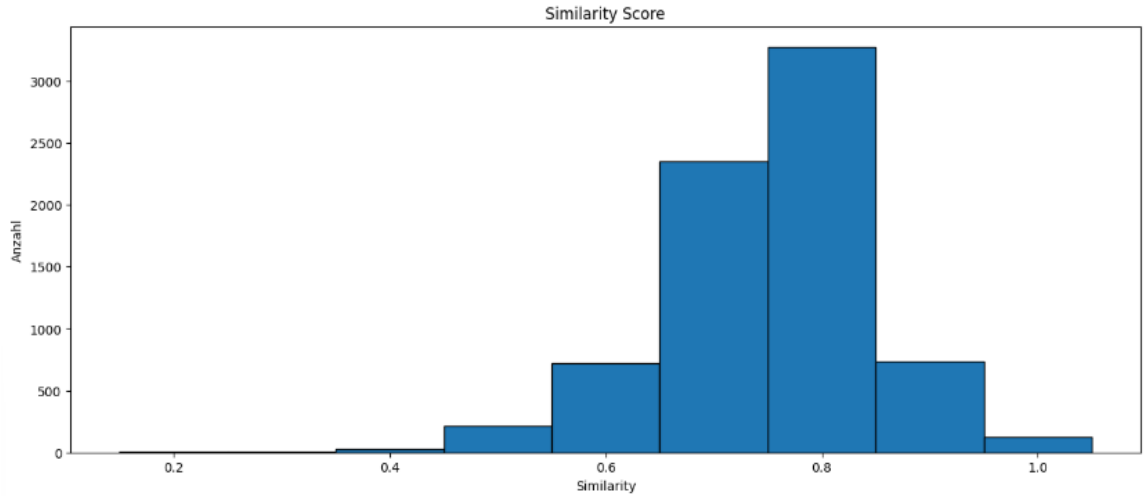

Abbildung 13: Verteilung der Similarity Scores über den Beispieldatensatz (Github)

Die Analyse hat gezeigt, dass die meisten Übersetzungen eine hohe (0,6–0,8) bis sehr hohe (0,8–1,0) Ähnlichkeit aufweisen. Dies spricht dafür, dass das verwendete Übersetzungstool bei der Übertragung vom Englischen ins Deutsche sehr gute Arbeit leistet und weiterhin verwendet werden kann. Der Code zur Überprüfung der Ähnlichkeit kann auch auf andere Sprachen angewendet werden. Dabei muss jedoch darauf geachtet werden, das jeweils passende Embedding zu nutzen und gegebenenfalls ein sprachspezifisches statt eines multilingualen Modells einzusetzen. Auch nach der Rückübersetzung der Texte vom Deutschen ins Englische ergaben sich vergleichbare Similarity Scores. Nur in einigen wenigen Fällen wich der Score um etwa 0,1 Punkte ab.

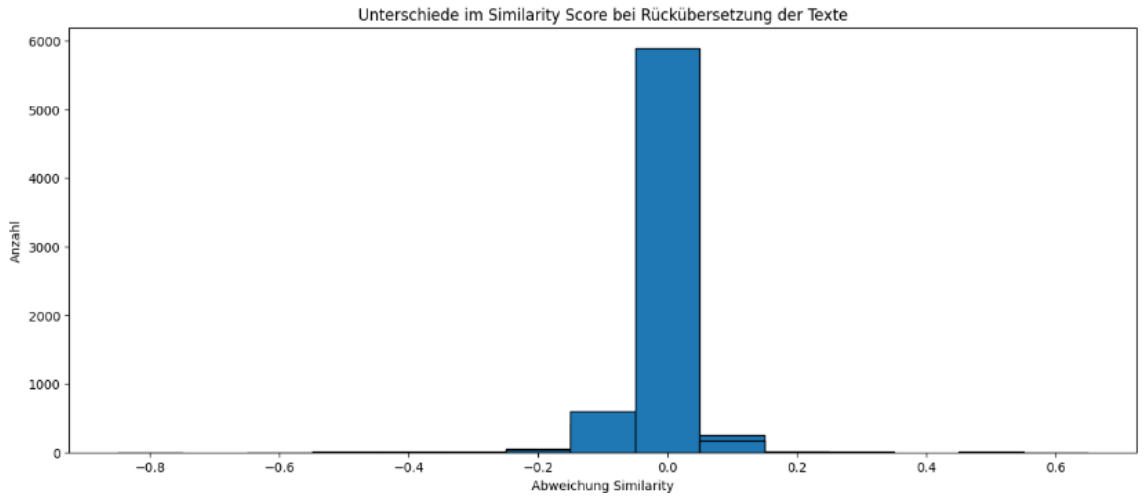

Abbildung 12: Abweichung der Similarity Scores nach Rückübersetzung der Texte (Github)

6.3. Grammatikalische Prüfung

Neben der inhaltlichen Richtigkeit eines Textes ist die Grammatik von entscheidender Bedeutung für dessen Akzeptanz und Professionalität. Die „LanguageTool“-Bibliothek von Python bietet eine einfache Möglichkeit, Texte auf grammatikalische Regeln wie Rechtschreibung, Satzbau und Zeichensetzung zu überprüfen. Sie erkennt fehlerhafte Passagen, markiert sie und schlägt Verbesserungen vor. Der Score zeigt an, wie viele Fehler pro Text gefunden wurden und gibt damit Aufschluss über die grammatikalische Korrektheit der Übersetzung.

Eine erste Analyse zeigt, dass bestimmte Texte mehr grammatikalische Fehler aufweisen als andere. Dies könnte auf eine weniger sorgfältige Übersetzung oder auf Herausforderungen bei der Übersetzung komplexer Sätze hindeuten. Insgesamt wurden in zwei Dritteln der Texte des Datensatzes Fehler gefunden, während ein Drittel fehlerfrei ist.

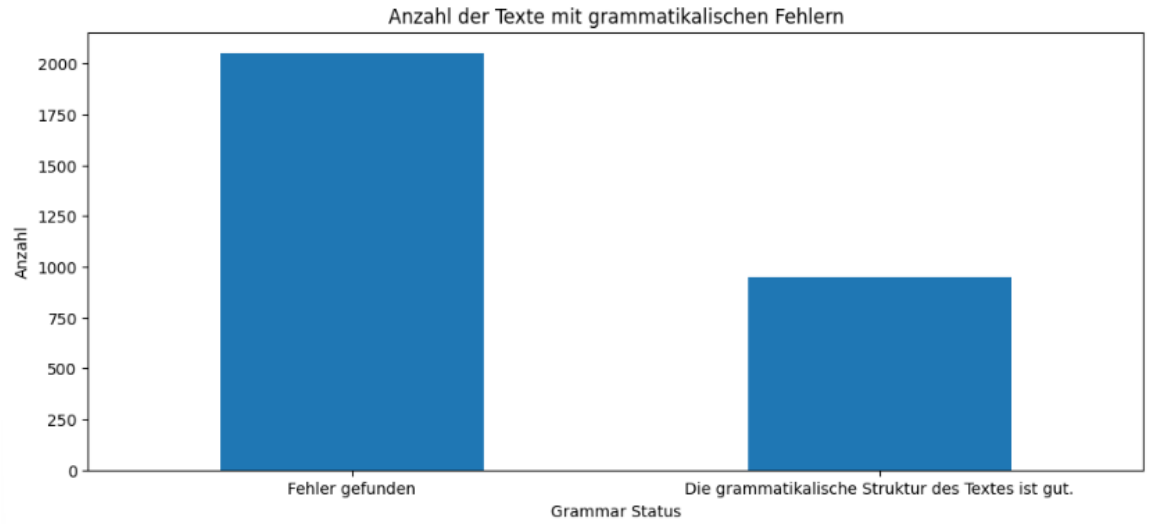

Abbildung 14: Anteil der Texte des Datensatzes mit grammatikalischen Fehlern (Github)

Es ist wichtig zu beachten, dass ein einzelner Text mehrere unterschiedliche grammatikalische Fehler enthalten kann. Insgesamt wurden 69 verschiedene Fehlermeldungen in diesem Datensatz identifiziert. Bei genauerer Betrachtung zeigt sich, dass der größte Anteil der Fehler bei der Verwendung von Satzzeichen und der Wortstellung auftritt, was auf systematische Schwächen in der Übersetzungsqualität hinweisen könnte und genauer untersucht werden muss. Insgesamt wurden 13.836 Fehlermeldungen bezüglich der Regel „GERMAN_SPELLER_RULE“ registriert. Deutlich seltener traten Fehler in den Bereichen „DE_CASE“, „WHITESPACE_RULE“ und „EINHEIT_LEERZEICHEN“ auf. Für jede dieser Kategorien werden in den Fehlernachrichten mögliche Ursachen angegeben. Dabei ist zu beachten, dass es pro Fehlermeldung unterschiedliche Formulierungen geben kann. Dies liegt daran, dass die Fehlernachrichten häufig auf ein spezifisches Wort im Text Bezug nehmen, wodurch aus einer einzigen Fehlermeldung verschiedene Varianten der Fehlernachricht entstehen. Aus der Tabelle geht hervor, dass die

„GERMAN_SPELLER_RULE“ mögliche Rechtschreibfehler identifiziert. Dies ist unter anderem darauf zurückzuführen, dass einige Texte vollständig in Großbuchstaben verfasst wurden. Es ist zudem wichtig zu berücksichtigen, dass das Übersetzungstool und das Tool zur grammatikalischen Prüfung möglicherweise mit unterschiedlichen Rechtschreibregeln arbeiten. Wenn diese Faktoren berücksichtigt werden, lässt sich feststellen, dass bei etwa 14.000 Rechtschreibfehlern in einem Datensatz von 7.000 Texten (sowohl langen als auch kurzen) durchschnittlich etwa 2 Rechtschreibfehler pro Text auftreten.

Fehler	Nachrichten	Anzahl Nachricht
GERMAN_SPELLER_RULE	Möglicher Tippfehler gefunden.	13835
DE_CASE	Außer am Satzanfang werden nur Nomen und Eigennamen großgeschrieben.	1714
WHITESPACE_RULE	Möglicher Tippfehler: mehr als ein Leerzeichen hintereinander	1496
EINHEIT_LEERZEICHEN	Vor Einheitenzeichen sollte ein Leerzeichen gesetzt werden.	1265

Abbildung 15: Häufigste Fehlermeldungen inkl. Fehlernachricht (Github)

Außerdem fällt auf, dass grammatikalische Fehler nur selten auf Satzbaufehler zurückzuführen sind. Dies ist ein positives Zeichen für die Qualität und Glaubwürdigkeit des Textes. Die zweithäufigste Fehlerquelle sind Probleme bei der Setzung von Leerzeichen sowie die Verwendung doppelter Leerzeichen. Mit deutlichem Abstand folgen einige Fehler in der Groß- und Kleinschreibung. Auch hier sollte geprüft werden, ob diese Fehler möglicherweise auf die Beschaffenheit bestimmter Textbausteine zurückzuführen sind, die absichtlich großgeschrieben wurden, beispielsweise bei Buttontexten.

6.4. Named Entity Recognition

Die Named Entity Recognition (NER) identifiziert und klassifiziert bestimmte Entitäten im Text, wie z.B. Personen, Organisationen oder Orte. Der Score gibt an, wie viele und welche Entitäten korrekt erkannt wurden. Dies könnte besonders interessant sein, da Eigennamen in der Ausgangssprache vom Übersetzungstool korrekt erkannt und entweder unverändert in die Zielsprache übernommen oder richtig übersetzt werden. Zur Implementierung des Scores gibt es eine Library von Spacy, welche sowohl die englische als auch die deutsche Sprache mithilfe des geeigneten Embeddings unterstütz. Bei Anwendung auf den Datensatz gehet jedoch hervor, dass der NER-Score bei der Erkennung von Entitäten je nach Textinhalt variiert. In Übersetzungen

können z.B. Entitäten wie Namen oder Organisationen anders behandelt werden als in der Ausgangssprache, was zu unterschiedlichen Erkennungsraten führt. Während beispielsweise in der Ausgangssprache nur „Organisationen“ erkannt wurden, traten in den deutschen Übersetzungen mehrere Kategorien auf.

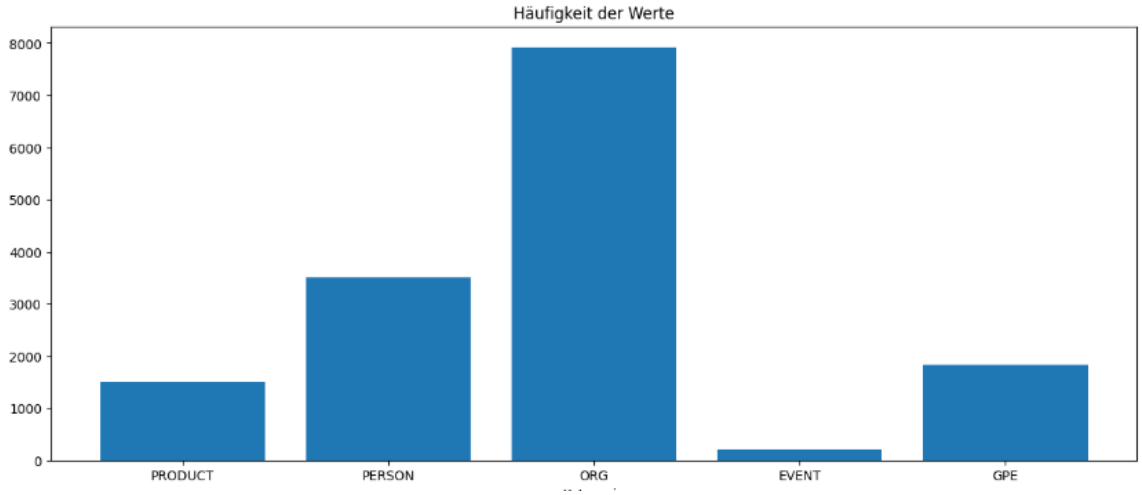

Abbildung 16: Kategorien der Eigennamen in den deutschen Übersetzungen (Github)

Mit deutlichem Abstand wurden Organisationen am häufigsten als Kategorie erkannt, jedoch traten auch Produkte und Personen vermehrt auf. Um die Kategorisierung weiter zu verbessern und eine Liste von Produkten und Eigennamen zu erstellen, die regelmäßig nicht korrekt erkannt werden, könnte ein geeigneteres Embedding verwendet werden. Es wäre sinnvoll, ein Embedding zu nutzen, das nicht nur auf eine bestimmte Sprache trainiert ist, sondern auch auf spezielle Fälle wie Eigennamen und den Marketingbereich optimiert wurde.

6.5. Readability Score

Der Readability-Score bewertet, wie leicht ein Text zu lesen und zu verstehen ist. Dafür werden verschiedene Metriken verwendet, der auf Faktoren wie Satzlänge, Silbenanzahl, Wortfrequenz und Wortkomplexität basiert. Die als geeignet identifizierten Metriken der Library Textstat wurden auf den gesamten Datensatz angewendet, die Ergebnisse visualisiert und verglichen. Jede der ausgewählten Metriken befasst sich mit einem anderen Fokus der Lesbarkeit:

- Flesch Reading Ease Index (Fokus: Silben und Wörter)
- Wiener Sachtextformel (Fokus: Deutsche Sprache)
- Gunning Fog Index (Fokus: Satzlänge und Silben)
- Dale Call Deutsch (Fokus: Wortliste schwieriger Wörter)

Die genauen Berechnungen der Metriken können in dem Notebook „Textstat“ nachvollzogen werden.

Flesch Reading Ease Index

Der Flesch Reading Ease Index ist ein Maß zur Beurteilung der Lesbarkeit von Texten mit dem Fokus auf der Länge der Wörter und Sätze: Je kürzer die Wörter und Sätze, desto höher ist die Lesbarkeit. Der Index gibt einen Wert auf einer Skala von 0 bis 100 an, wobei höhere Werte leichter verständliche Texte anzeigen. Ein Wert von 60 bis 70 gilt als gut lesbar für ein breites Publikum, während Werte unter 30 auf schwierige Texte hinweisen. Der Index hilft dabei, Texte für eine allgemeine Zielgruppe verständlicher zu gestalten.

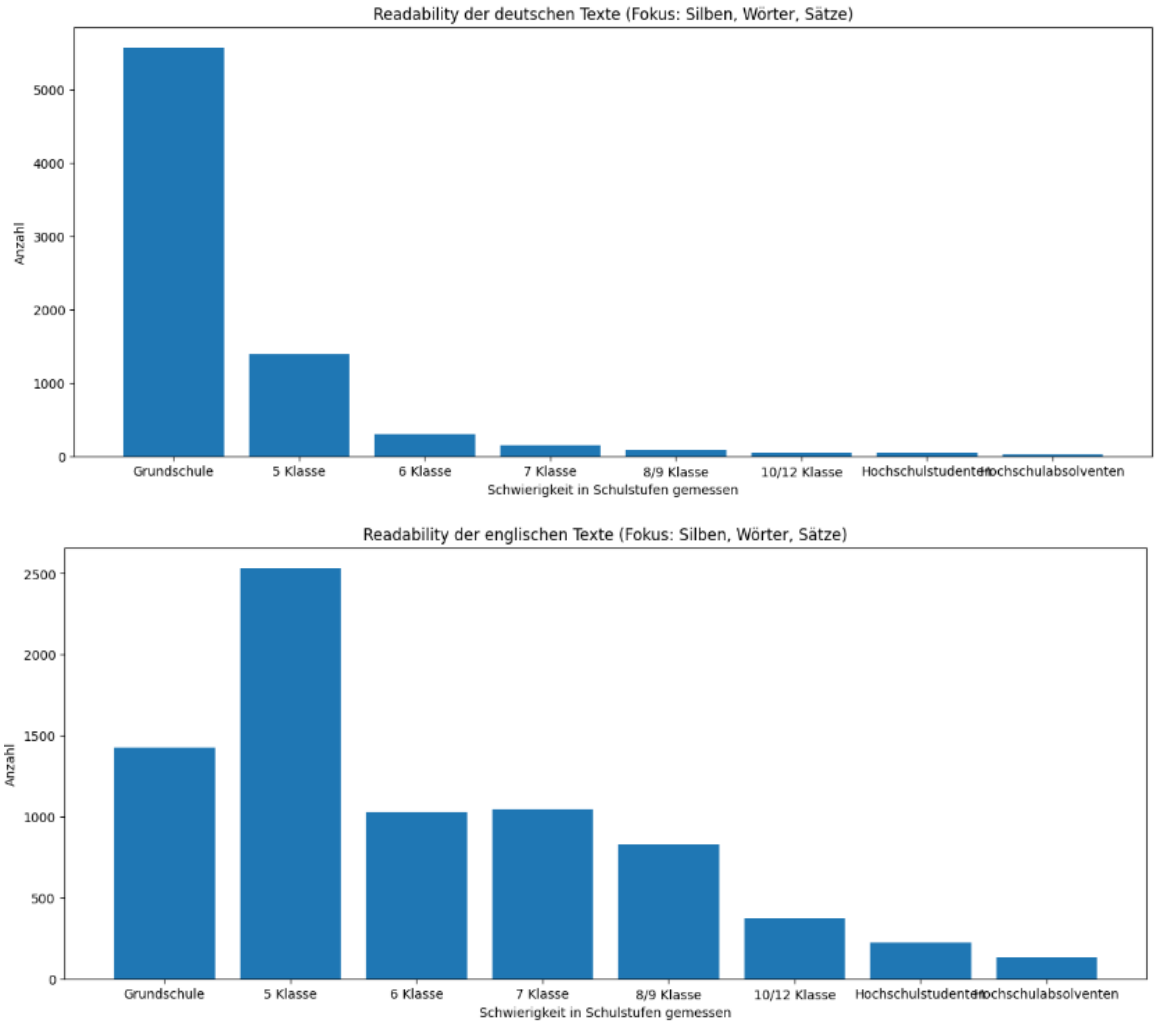

Abbildung 17: Verlgiech des Flesch Reading Ease Index Englisch – Deutsch (Github)

Beim Vergleich des Flesch Reading Ease Index der englischen Ausgangstexte mit den deutschen Übersetzungen fällt auf, dass die deutschen Texte als leichter verständlich bewertet werden als die englischen. Dies könnte darauf hindeuten, dass die englischen Texte tendenziell ein komplexeres Vokabular mit mehr Silben verwenden und längere Sätze enthalten als ihre deutschen Übersetzungen.

Wiener Sachtextformel

Die Wiener Sachtextformel ist ein Modell zur Erstellung verständlicher und klar strukturierter Sachtexte. Sie basiert auf der Annahme, dass Texte effektiv durch eine klare Gliederung und präzise Sprache überzeugen und ist speziell für die deutsche Sprache entwickelt worden. Sie umfasst verschiedene Kennzahlen wie Satzlänge, Wortlänge, Silbenanzahl und deren Verhältnisse zueinander, um zu bestimmen, wie anspruchsvoll oder zugänglich ein Text ist.

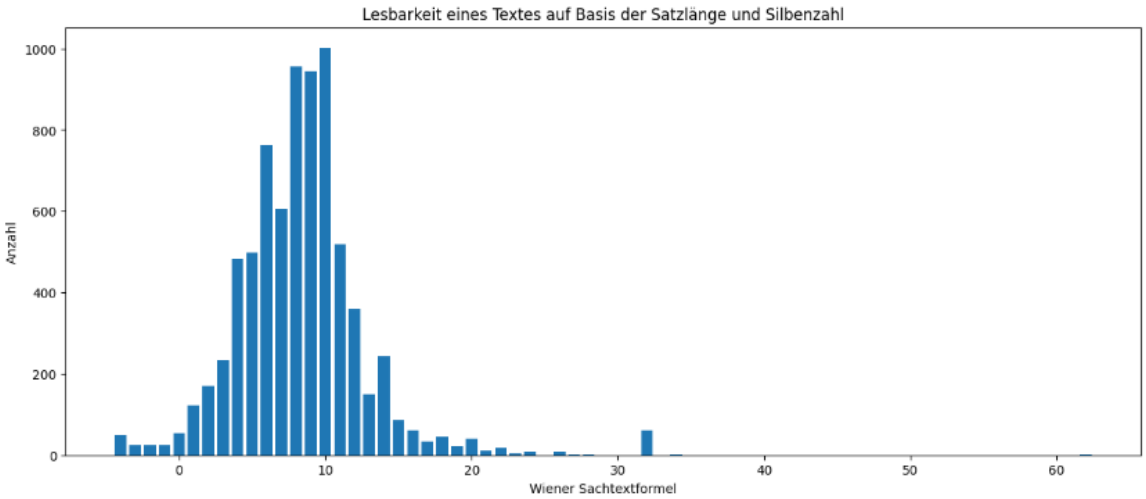

Abbildung 18: Bewertung der Lesbarkeit der deutschen Übersetzung mithilfe der Wiener Sachtextformel (Github)

Die Ergebnisse des Scores werden in den Jahren der Schulbildung angegeben, die erforderlich sind, um den Text vollständig zu verstehen. Im Vergleich zum Flesch Reading Ease Index zeigt sich, dass die Texte tendenziell komplexer sind, da für die meisten Texte eine Bildung von 6 bis 10 Jahren notwendig ist. Im Gegensatz dazu benötigte man laut der vorherigen Auswertung lediglich 4 bis 5 Jahre. Diese Diskrepanz kann darauf zurückzuführen sein, dass bei der Bewertung höhere Maßstäbe für die Textschwierigkeit angelegt werden, insbesondere in Bezug auf Wortlänge, Tokenanzahl, Silbenanzahl und deren Verhältnisse. Es ist in weiteren Untersuchungen zu prüfen, wie genau die Parameter für die Berechnung der durchschnittlichen Silbenanzahl und Satzlänge gesetzt wurden.

Gunning Fog Index

Der Gunning Fog Index ist eine Lesbarkeitsformel, welcher ebenfalls berechnet, wie viele Jahre Schulbildung im Durchschnitt erforderlich sind, um den Text leicht zu verstehen. Die Formel beschäftigt sich mit der Komplexität von Wörtern und berücksichtigt die Anzahl der Wörter, Sätze und die Anzahl der Wörter mit mehr als zwei Silben. Dies bedeutet je länger ein Wort ist, desto schwieriger wird dieses eingestuft. Ein höherer Indexwert deutet auf einen komplexeren Text hin,

während ein niedrigerer Wert auf eine leichtere Lesbarkeit hinweist. Die Methode dient oft zur Einschätzung der Zugänglichkeit von schriftlichen Materialien für technische Texte.

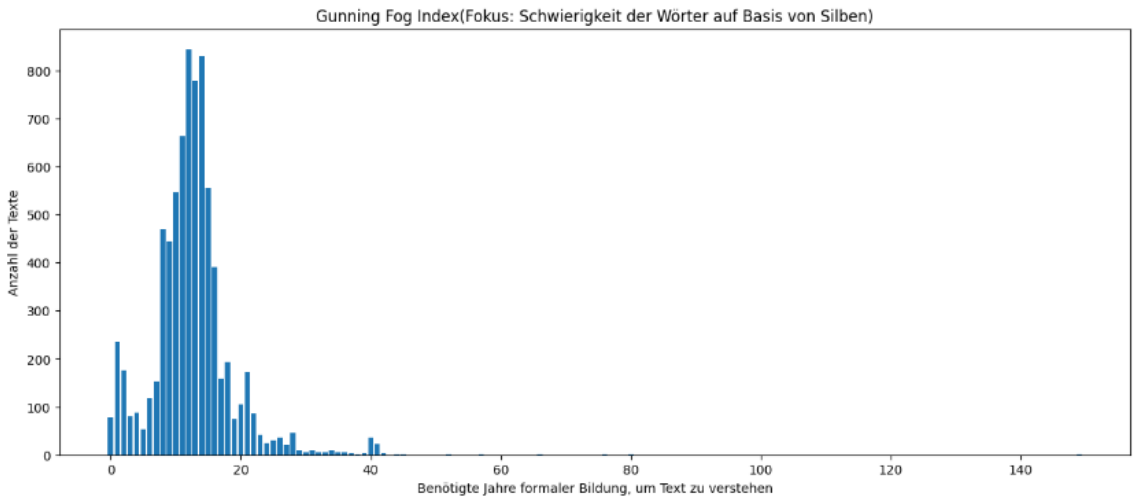

Abbildung 19: Bewertung der schwierigen Wörter des Datensatzes mithilfe des Gunning Fog Index (Github)

Die Ergebnisse der Anwendung auf den Datensatz stimmen überein mit denen der Wiener Sachtextformel, was darauf hindeutet, dass beide Metriken die Schwierigkeit der Wörter auf Grundlage der Silbenanzahl ähnlich bewerten.

Dale Call Deutsch

Der Dale-Call Reading Difficulty Index, misst die Lesbarkeit eines Textes, indem er die Schwierigen Wörter in einem Text identifiziert. Hierfür gibt es eine Liste von 3000 „einfachen Wörtern“, welche mit dem Datensatz abgeglichen wird. Er wurde entwickelt, um die Verständlichkeit von Texten für unterschiedliche Bildungsniveaus zu bewerten. Der Index gibt an, welches Alter nötig ist, um den Text leicht zu verstehen. Dabei werden längere Sätze und kompliziertere Wörter als Zeichen für höhere Schwierigkeit gewertet. Der Dale-Call Index ist nützlich, um den Schwierigkeitsgrad von Texten für Bildungs- und Kommunikationszwecke einzuschätzen.

```
# Berechnung des Dale-Chall-Scores
percent_difficult_words = (num_difficult_words / num_words) * 100
score = 0.1579 * percent_difficult_words + 0.0496 * (num_words / num_sentences)

if score >= 5:
    score += 3.636
```

Abbildung 20: Berechnung des Dall Cale Index in Python (Github)

Die Bibliothek Textstat bietet diesen Index derzeit nur für die englische Sprache an. Dennoch kann die Berechnung des Dale-Carr Index auch auf andere Sprachen wie Deutsch übertragen

werden. Dazu muss die zugrunde liegende Rechenlogik in eine spezifische Funktion für die jeweilige Sprache integriert und die Wortliste entsprechend übersetzt werden. Wird diese Funktion auf den Beispieldatensatz angewendet, zeigt sich ein etwas anderes Bild im Vergleich zu den bisher verwendeten Scores. In diesem Fall sind die meisten Texte so gestaltet, dass sie von Personen im Alter von 17 Jahren (entsprechend etwa 10 Jahren Schulbildung oder der 10. Klassenstufe) verstanden werden können. Die Wortliste kann jedoch je nach Zielgruppe angepasst werden, indem einfacher verständliche Wörter hinzugefügt werden, um ein realistischeres Bild der Textverständlichkeit zu erhalten.

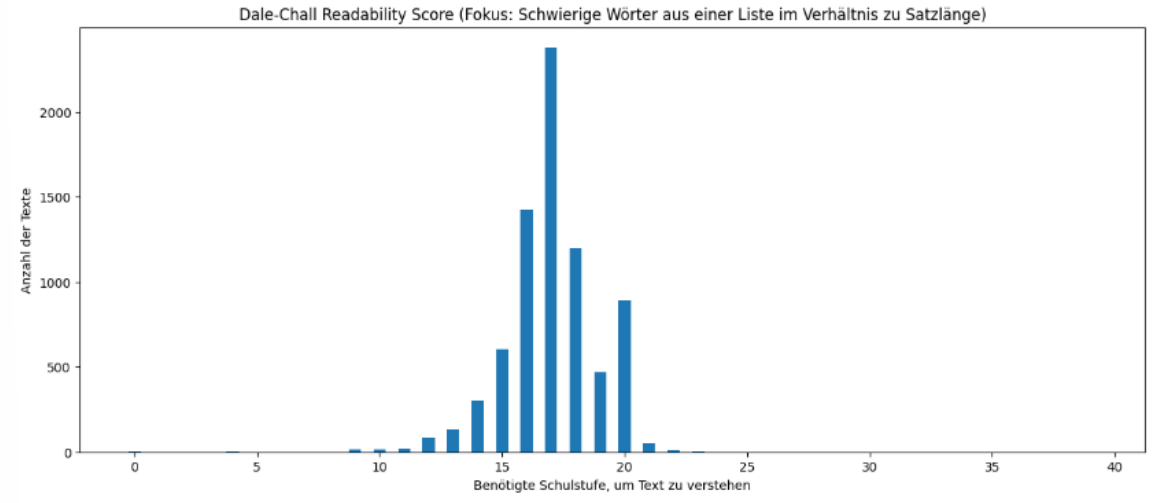

Abbildung 21: Bewertung der Lesbarkeit mithilfe des Dale Call Index (Github)

Zusammenfassend deuten die Ergebnisse der Readability darauf hin, dass die Lesbarkeit der Übersetzungen variiert. Texte mit komplexen Satzstrukturen oder Fachbegriffen sind in der Regel schwerer zu lesen. Dies könnte daran liegen, dass technische oder spezialisierte Texte eine geringere Lesbarkeit aufweisen, was auf die Verwendung von Fachterminologie und langen Wörtern zurückzuführen ist. Dies könnte die Verständlichkeit für ein allgemeines Publikum erschweren. Um die Lesbarkeit zu verbessern, könnten alternative Übersetzungen für komplexe Sätze angeboten werden. Zusätzlich könnte eine Anpassung des Vokabulars an das Zielpublikum erfolgen, um die Verständlichkeit zu erhöhen.

6.6. Reading Time

Ergänzend zu den bisherigen Analysen wurde zur verfolständigung des Bildes die Benötigte Lesezeit der Texte unter die Lupe genommen. Der Reading Time-Score gibt an, wie lange es dauert, einen Text zu lesen. Dies wird basierend auf der durchschnittlichen Lesegeschwindigkeit einer Person berechnet. Die Lesezeiten variieren je nach Textlänge und Komplexität. Texte mit längeren Sätzen und anspruchsvollerem Vokabular erfordern in der Regel mehr Zeit. Angesichts der Tatsache, dass die Texte im Datensatz normalerweise kurz sind, ist es überraschend, dass für

manche Texte eine Lesezeit von bis zu 20 oder 30 Minuten erforderlich ist, vergleichbar mit einem längeren Blogartikel. Dies könnte darauf hindeuten, dass die Textabschnitte aufgrund ihrer Länge oder Komplexität unverhältnismäßig viel Zeit in Anspruch nehmen, was die Leserfreundlichkeit beeinträchtigen könnte.

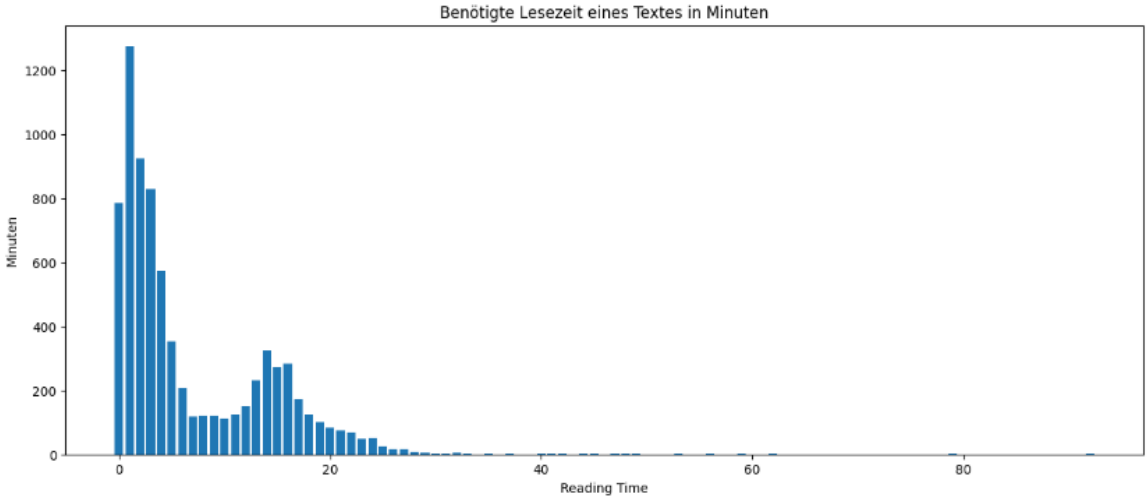

Abbildung 22: Benötigte Lesezeit der Texte des Beispieldatensatzes (Github)

In einer weiteren Untersuchung könnte es sinnvoll sein, die Texte mit einem hohen Reading Time Score auszuwählen und deren Readability zu analysieren. Texte, die sowohl eine lange Lesezeit als auch einen hohen Readability aufweisen, könnten auf eine übermäßige Komplexität hinweisen. Im Gegensatz dazu könnten Texte mit einer niedrigen Lesbarkeit durch ihre Länge die hohe Lesezeit erklären. Diese Analyse könnte helfen, die Balance zwischen Textkomplexität und -länge besser zu verstehen und die Leserfreundlichkeit zu optimieren.

7. Diskussion und Ausblick

Die durchgeführte Analyse hat bedeutende Erkenntnisse zur Evaluierung der Qualität maschinell übersetzter Texte geliefert. Besonders hervorzuheben ist die Fähigkeit des entwickelten Codes, inkonsistente Übersetzungen zu identifizieren und die semantische Übereinstimmung umfassend zu analysieren. Diese Analyse geht weit über einfache Wortvergleiche hinaus und ermöglicht eine tiefere Bewertung der Übersetzungsqualität.

Jedoch gibt es auch signifikante Herausforderungen. Die Rechenintensität bei der Anwendung des BERT-Modells für den Similarity Score führt zu langen Laufzeiten, insbesondere bei großen Datenmengen. Zudem könnte die Analyse auf Satzebene die Entdeckung von Inkonsistenzen auf Dokumenten- oder Abschnittsebene erschweren, und kulturelle Nuancen in Übersetzungen bleiben möglicherweise unberücksichtigt.

Für zukünftige Arbeiten könnten folgende Verbesserungen und Erweiterungen von Bedeutung sein:

- **Erweiterung auf weitere Sprachen**: Die Übertragung der Bewertungsmethoden auf andere Sprachen könnte durch Anpassung der Embeddings und Übersetzung der Wortlisten erfolgen.
- **Optimierung der Rechenleistung**: Die Implementierung von Optimierungen zur Beschleunigung der Ähnlichkeitsberechnungen, wie der Einsatz spezialisierter Hardware oder weniger rechenintensiver prädiktiver Modelle, könnte die Effizienz verbessern und so die Analysen im großen Stil ermöglichen.
- **Fehlerkategorien detaillierter untersuchen**: Eine genauere Analyse von Fehlerkategorien wie Leerzeichen, Rechtschreibung und Groß-/Kleinschreibung könnte dazu beitragen, spezifische Fehlerquellen besser zu verstehen und zu beheben.
- **Berücksichtigung von Lesbarkeit**: Die Einbeziehung von Lesbarkeitsmetriken könnte zusätzliche Einblicke in die Verständlichkeit der Texte liefern und die Nutzerfreundlichkeit verbessern.
- **Entwicklung einer Strategie für Eigennamen**: Die Erstellung einer Liste zur Behandlung von Eigennamen könnte die Übersetzungsqualität weiter optimieren und konsistenter gestalten.

Zusammenfassend bieten die bisherigen Methoden eine solide Grundlage, um die Übersetzungsqualität zu bewerten. Künftige Forschungen sollten hybride Ansätze in Betracht ziehen, die die Stärken verschiedener Methoden kombinieren, und die Auswirkungen unterschiedlicher Trainingsdaten auf die Bewertungsgenauigkeit weiter untersuchen. Diese Schritte könnten dazu beitragen, die Qualität der maschinellen Übersetzung und ihre praktische Anwendbarkeit zu verbessern.

Literaturverzeichnis

Hellge, D. V. (2024). *READINESS-CHECK*. Retrieved from https://digitalzentrum-kaiserslautern.de/unser-angebot/self-service/readiness-check

huggingface. (n.d.). *Metric: chr_f*. Retrieved from huggingface: https://huggingface.co/spaces/evaluate-metric/chrf

Kishore Papineni, S. R.-J. (2022). *Bleu: a Method for Automatic Evaluation of Machine Translation*. Retrieved from https://aclanthology.org/P02-1040/

Kuhn, R. (2023, 05 23). *10 Science of Reading Strategies for Reading Instruction*. Retrieved from hmhco.com: https://www.hmhco.com/blog/science-of-reading-strategies-for-reading-instruction

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, 07 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Retrieved from https://arxiv.org/pdf/1907.11692

Nureg GmbH. (2023). *Idea and Vision draft.*

Open AI. (2022, 12 15). *New and improved embedding model*. Retrieved from https://openai.com/index/new-and-improved-embedding-model/

Popović, M. (2015). *chrF: character n-gram F-score for automatic MT evaluation*. Retrieved from https://aclanthology.org/W15-3049/

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, 10 19). *COMET: A Neural Framework for MT Evaluation*. Retrieved from https://arxiv.org/pdf/2009.09025

Sellam, P. T., Das, D., & Parikh, A. (2020). *BLEURT: Learning Robust Metrics for Text Generation*. Retrieved from https://arxiv.org/pdf/2004.04696

Vashee, K. (2019, 04 19). *Understanding MT Quality: BLEU Scores*. Retrieved from https://kvashee.medium.com/understanding-mt-quality-bleu-scores-9a19ed20526d

Zhang, T., Kishore, V., Wu, F., & Weinberger, K. Q. (2020, 02 24). *BERTSCORE: EVALUATING TEXT GENERATION WITH*. Retrieved from https://arxiv.org/pdf/1904.09675