Kursorischer Vergleich der Übersetzungsdienste ChatGPT, Google Translate, DeepL und eTranslation

Tim Hagen Martin Potthast
Universität Kassel und hessian.AI

Um die Qualität der Übersetzungen der Dienste ChatGPT, Google Translate, DeepL und eTranslation zu vergleichen, haben wir zwei Experimente durchgeführt: Experiment 1 vergleicht verschiedene Übersetzungsverfahren auf Basis eines Benchmarks und Experiment 2 analysiert die Übereinstimmung der Übersetzungen der Dienste untereinander.

Wir stellen fest, dass die Qualität der Übersetzungen der verglichenen Dienste in der Regel ähnlich gut ist. Dennoch übersetzen ChatGPT, Google Translate und DeepL messbar besser als eTranslation. Für den praktischen Gebrauch macht die Benutzerfreundlichkeit (Geschwindigkeit, Integration, Zusatzfunktionen) der Dienste den größten Unterschied.

Stand der Technik

Google Translate und DeepL werden in der Forschung zur maschinellen Übersetzung als repräsentative Baselines verwendet [2, 3, 6]. Seit der Einführung von GPT-4 ist auch ChatGPT konkurrenzfähig mit dem Stand der Technik, insbesondere bei der Übersetzung von so genannten „ressourcenstarken Sprachen“, d. h. Sprachpaaren, für die viele Daten im Internet verfügbar sind, wie z. B. Deutsch und Englisch. Die in der Forschung gemessenen relativen Qualitätsunterschiede zwischen diesen Diensten sind konsistent mit unseren Messungen.

Experiment 1: Vergleich auf einem Benchmark.

Daten Wir verwenden das weitläufig eingesetzte „OPUS“-Benchmark.[1] OPUS sammelt zahlreiche sogenannte parallele Textkorpora für 744 Sprachen. Jedes Korpus enthält satzweise sprachübergreifend zugeordnete Übersetzungen von Texten in zwei oder mehr Sprachen. Als Grundlage für unseren kursorischen Vergleich verwenden wir das Books-Korpus.[2] Es besteht aus urheberrechtsfreier Literatur bekannter Autoren. Aus diesem 51.67 Sätze für das Sprachpaar Englisch–Deutsch umfassenden Datensatz ziehen wir eine Stichprobe von 1.262 Sätzen, die als Referenz für die Erfolgsmessung dient.

Vorgehen Jeder der 1.262 englischen Sätze wird jeweils den Diensten ChatGPT (in der Variante GPT-4o), Google Translate, DeepL und eTranslation übergeben und die deutsche Übersetzung aufgezeichnet. Im Fall von ChatGPT verwenden wir den folgenden Systemprompt:

```
You are an english-german translation service. When the user writes you a
message, you must respond only with the german translation of that message.
```

[1] https://opus.nlpl.eu/
[2] https://opus.nlpl.eu/Books/corpus/version/Books

Dienst	Books Corpus		ChatGPT		Google Translate		DeepL	
	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore
ChatGPT	19,7 %	82,8 %	—	—	—	—	—	—
Google Translate	18,7 %	82,2 %	49,7 %	90,8 %	—	—	—	—
DeepL	20,4 %	82,5 %	49,9 %	90,5 %	53,5 %	91,9 %	—	—
eTranslation	17,5 %	82,1 %	46,5 %	90,3 %	48,8 %	90,3 %	45,0 %	89,4 %

Tabelle 1: Ergebnisse der Experimente 1 (Spalte „Books Corpus“) und 2 (Spalten „ChatGPT“, „Google Translate“ und „DeepL“), bei denen die jeweils bezeichneten Datenquellen als Referenzübersetzung dienten.

Insbesondere ist der Prompt an ChatGPT „zero-shot“ (also ohne Beispiele) und wir haben kein Prompt-Engineering betrieben um die Ausgaben von ChatGPT weiter zu optimieren. Das soll eine alltägliche Nutzung durch unbedarfte Nutzer:innen nachbilden. Die aufgezeichneten Übersetzungen werden anschließend mit der jeweiligen Referenzübersetzung des Books-Corpus verglichen, um die Güte der Übersetzung festzustellen.

Erfolgsmessung Die Güte jeder Übersetzung wird mittels zweier Erfolgsmaße quantifiziert: Das erste Maß heißt „Bilingual Evaluation Understudy“ [5] (BLEU, ausgesprochen wie das französische Wort für ‘blau’). Es bemisst die Qualität einer Übersetzung anhand von überlappenden Phrasen einer Kandidatenübersetzung mit einer Referenzübersetzung. Je mehr Phrasen gleich sind und je länger die Phrasen werden, desto höher der gemessene Wert. BLEU errechnet im Extremfall einen minimalen Wert von 0, wenn kein einziges Wort übereinstimmt, bzw. einen maximalen Wert von 1, wenn die Kandidaten- und die Referenzübersetzung exakt gleich sind. Der errechnete Wert wird üblicherweise mit 100 multipliziert und als Prozentzahl angegeben.

Das zweite Maß heißt BERTScore [7] und verwendet ein modernes neuronales Netz, das „Bidirectional Encoder Representations from Transformers“ (BERT) heißt [1], um die inhaltliche Ähnlichkeit zweier Sätze zu bestimmen, unabhängig davon, ob einzelne Wörter oder Phrasen überlappen. Mit diesem Maß werden auch Kandidatenübersetzungen berücksichtigt, die die Referenzübersetzung paraphrasieren. Als Paraphrase wird ein Satz bezeichnet, der den Inhalt eines anderen Satzes sinngemäß mit anderen Worten wiedergibt. Auch hier wird ein minimaler Wert von 0 berechnet, wenn zwei Sätze keine inhaltliche Ähnlichkeit aufweisen, bzw. eine maximaler Wert von 1, wenn die Bedeutung der Sätze exakt übereinstimmt.

Während das erste Maß fordert, dass die Referenzübersetzung möglichst wortgetreu getroffen wird, berücksichtigt das zweite Maß, dass es zahlreiche alternative Übersetzungen geben kann. Dieses Maß fordert daher nur eine näherungsweise akkurate Übersetzung.

Ergebnisse Tabelle 1 (Spalte „Books Corpus“) fasst die Messergebnisse des Experiments zusammen. ChatGPT und DeepL erzielen sehr ähnliche BLEU-Werte von rund 20 %. eTranslation rangiert mit 17,5 % auf dem letzten Platz. Die BERTScores der vier Dienste sind nahezu identisch.

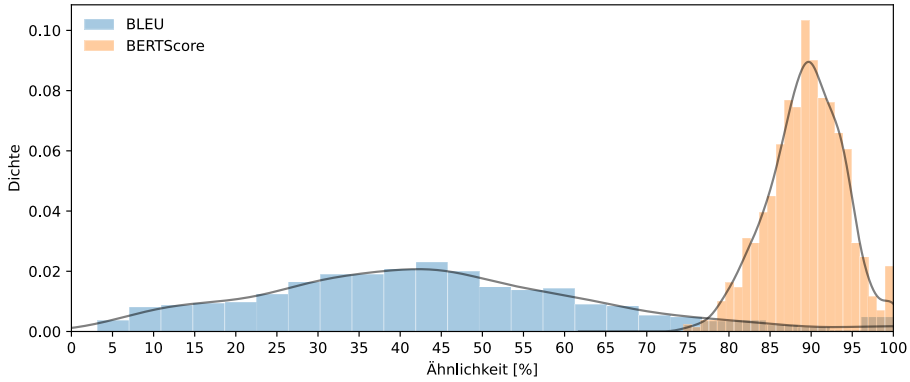

Abbildung 1: Die Verteilungen der Ähnlichkeitswerte (BLEU und BERTScore) für die Evaluation von eTranslation mit DeepL als Referenzübersetzung.

Experiment 2: Relativer Vergleich der Dienste.

Daten Die in Experiment 1 gesammelten deutschen Übersetzungen der 1.262 englischen Sätze von den vier verschiedenen Diensten dienen als Grundlage für dieses Experiment. Für jeden englischen Satz liegen also je vier deutsche Kandidatenübersetzungen vor.

Vorgehen Unter Verwendung der oben genannten Erfolgsmaße BLEU und BERTScore wird jede Kandidatenübersetzung eines Satzes mit den Kandidatenübersetzungen der jeweils anderen Dienste für denselben englischen Satz verglichen, um relative Unterschiede zu messen. Zum Beispiel wird eine Übersetzung DeepLs mit der eTranslations verglichen, um die Unterschiede zwischen den Diensten zu quantifizieren.

Ergebnisse Tabelle 1 (Spalten „ChatGPT“, „Google Translate“ und „DeepL“) fasst die Messergebnisse zusammen. DeepL und Google Translate sind sich in ihren Übersetzungen ähnlicher als die anderen Dienste. Der Vergleich von DeepL mit eTranslation ergibt einen BLEU-Wert von 45,0 %, der Vergleich von DeepL mit ChatGPT einen BLEU-Wert von 49,9 % und der Vergleich von eTranslation mit ChatGPT 46,5 %. Die jeweiligen BERTScores dieser Vergleiche belaufen sich in allen Fällen auf rund 90 %. Insgesamt sind die Übersetzungen der Dienste also recht ähnlich.

Der Anschaulichkeit halber, stellt Abbildung 1 zudem die Verteilung beider Ähnlichkeitswerte dar, wenn eTranslation mit DeepL als Referenz evaluiert wird. Es wird auch hier deutlich, dass die Übersetzungen von eTranslation und DeepL semantisch sehr ähnlich sind (BERTScore streut gering um $\sim 90\,\%$). Außerdem gibt es keine Übersetzung, die im BERTScore schlechter als 75 % der Referenzübersetzung ist und in beiden Metriken gibt es einen kleinen Spike nahe der 100 % Ähnlichkeit, die zeigt, dass ca. 2 % von eTranslations Übersetzungen quasi identisch sind zu DeepLs.

Diskussion der Ergebnisse

Der zahlenmäßige Vergleich suggeriert, dass die Übersetzungen von eTranslation denen von DeepL und ChatGPT leicht unterlegen sind. Der Unterschied erscheint allerdings nicht dramatisch; die meisten Übersetzungen sind nahezu identisch bzw. gleich gut. Der Großteil der Übersetzungen sind annähernd bis ununterscheidbar gleich gut wie die von DeepL und ChatGPT.

Automatische Erfolgsmessungen für die Verarbeitung natürlicher Sprache sind aber lediglich grobe Heuristiken und können tatsächliche Qualitätsunterschiede zum Teil stark unterschätzen [4].

Zur Veranschaulichung der gemessenen Unterschiede und von den möglicherweise auftretenden Qualitätseinbußen bei der Verwendung von eTranslation können die folgenden sechs Beispiele dienen, hier im Vergleich zu DeepL. Zunächst die drei Beispiele von eTranslation, deren BLEU-Werte im Vergleich zu denen DeepLs am geringsten sind:

Englisch: „Be seated somewhere; and until you can speak pleasantly, remain silent."

eTranslation: „Irgendwo sitzen; Und bis du angenehm reden kannst, schweige."

DeepL: „Setzen Sie sich irgendwo hin, und bis Sie angenehm sprechen können, schweigen Sie."

Englisch: „Who blames me? Many, no doubt; and I shall be called discontented."

eTranslation: „Wer gibt mir die Schuld? Viele, ohne Zweifel; Und ich werde unzufrieden genannt werden."

DeepL: „Wer tadelt mich? Zweifellos viele, und man wird mich unzufrieden nennen."

Englisch: „That eye of hers, that voice stirred every antipathy I had."

eTranslation: „Dieses Auge von ihr, diese Stimme erregte jede Antipathie, die ich hatte."

DeepL: „Ihr Blick und ihre Stimme weckten meine ganze Antipathie."

Anschließend die drei Beispiele von eTranslation, deren BERTScores im Vergleich zu denen DeepLs am geringsten sind:

English: „Not at all–it bears the most gracious message in the world: for the rest, you are not my conscience-keeper, so don't make yourself uneasy."

eTranslation: „Überhaupt nicht - es trägt die gnädigste Botschaft der Welt: Für den Rest, du bist nicht mein Gewissen-Hüter, so machen Sie sich nicht unbehaglich."

DeepL: „Im Übrigen sind Sie nicht mein Gewissenswächter, also machen Sie sich keine Sorgen."

English: „Again I paused; then bunglingly enounced–"

eTranslation: „Wieder pausierte ich; dann bunglingly enounced–"

DeepL: „Wieder hielt ich inne, dann sagte ich stümperhaft."

English: „Here, leaning over the banister, I cried out suddenly, and without at all deliberating on my words–"

eTranslation: „Hier, über das Geländer gelehnt, schrie ich plötzlich und ohne über meine Worte nachzudenken..."

DeepL: „Hier lehnte ich mich über das Geländer und rief plötzlich, ohne zu überlegen, was ich sagen sollte."

Es zeigt sich, dass der messbare Qualitätsunterschied eTranslations zu DeepL sich in falscher bzw. nicht geläufiger Grammatik und Wortwahl äußert. Darüber hinaus kennt eTranslation eher wenig verwendete Wörter wie 'bunglingly' nicht. Eine weitere Beobachtung ist, dass eTranslation nicht systematisch zwischen britischem Englisch und amerikanischem Englisch unterscheidet. Zum Beispiel wird 'Kofferraum' mit 'trunk' und 'Taschenlampe' mit 'flashlight' wie im amerikanischen Englisch übersetzt, aber 'Meter' mit 'Metre' wie im britischen Englisch. Wir konnten jedoch nicht beobachten, dass eTranslation amerikanisches und britisches Englisch innerhalb derselben Übersetzung vermischt.

Nutzungsfreundlichkeit Ein wesentlicher Unterschied zwischen den drei Diensten besteht in ihrer Nutzungsfreundlichkeit für die tägliche Arbeit.

Die Schnittstelle von DeepL (sowohl die auf der Webseite als auch die lokale zu installierende Software) sowie die mögliche Integration des Dienstes in gängige Office-Programme sind bedeutend anwenderfreundlicher als die der anderen Dienste. DeepL bietet als einziger der vier Dienste Formulierungsalternativen für jedes Wort auf Knopfdruck an, erlaubt ein schnelles Umschalten der Übersetzungsrichtung und bietet gegen Entgelt eine personalisierbare Übersetzungsglossarfunktion. Dies beschleunigt die Arbeit und Optimierung von Übersetzungen sowie Texten, selbst innerhalb der eigenen Muttersprache erheblich.

Im Vergleich ist ChatGPT ist zwar auch leicht zu verwenden, man muss allerdings jedes Mal einen Prompt zum Übersetzen formulieren. Ansonsten ist ChatGPT als generisches Sprachmodell ein Werkzeug für sehr viele alltägliche Aufgaben gleichzeitig, wohingegen DeepL hoch spezialisiert auf Übersetzungsaufgaben und sekundär die Optimierung von Formulierungen ist.

Die Nutzungsschnittstelle von eTranslation ist dagegen vergleichsweise rudimentär. Der Dienst antwortet bedeutend langsamer als die anderen Dienste, was die Arbeit verlangsamt. Die Schnittstelle bietet darüber hinaus keine der vorgenannten Funktionen oder Integrationen.

eTranslation eignet sich unserer Ansicht nach daher für die gelegentliche Nutzung. DeepL, Google Translate und sekundär ChatGPT, eignen sich hingegen für das professionelle Schreiben im fachlichen Kontext bei häufiger Nutzung. Es hängt also stark vom Anspruch an die erwartbare Ergebnisqualität und der Häufigkeit der Nutzung ab, welcher Dienst die Nase vorn hat.

Limitierungen Einschränkend ist zu sagen, dass wir nur eine verhältnismäßig kleine Anzahl von Sätzen in den Experimenten heranziehen konnten. Das liegt daran, dass die jeweiligen Dienste nur geringe Freikontingente für den automatisierten Zugriff anbieten und ansonsten teils kostenpflicht sind. Dennoch denken wir, dass selbst mit einer deutlich größeren Zahl von Beispielsätzen keine dramatisch anderen Ergebnisse zu erwarten wären.

Darüber hinaus deckt das Books Corpus nur das Genre der Literatur ab, nicht jedoch die zahlreichen anderen denkbaren Textgenres, die in der praktischen Verwendung in Wissenschaft und Verwaltung häufiger zu erwarten sind. Was Fachterminologie anbelangt, ist jedoch zu erwarten, dass eTranslation, da es von und für EU-Institutionen entwickelt wurde, es mit den kommerziellen Diensten aufnehmen kann. Darüber sind die Referenzübersetzungen des Books Corpus nur teilweise exakt, da literarische Text, die professionell übersetzt wurde häufig auch das Satzgefüge verändern. In der Evaluation betrifft dieses Problem aber alle Dienste gleichermaßen und somit sind relative Qualitätsunterschiede in der Übersetzung denoch verlässlich messbar.

Zuletzt ist zu sagen, dass die kommerziellen Dienste wahrscheinlich auf den öffentlich verfügbaren Daten des OPUS-Benchmarks und damit insbesondere auch auf denen des Books Corpus trainiert wurden. Eventuell sind die gemessenen Ergebnisse also leicht überschätzt.

Literatur

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[2] Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. A paradigm shift: The future of machine translation lies with large language models. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1339–1352. ELRA and ICCL, 2024.

[3] Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt? In Philipp Koehn, Barry Haddon, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 224–245. Association for Computational Linguistics, 2023.

[4] Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. How does quantization affect multilingual llms? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 15928–15947. Association for Computational Linguistics, 2024.

[5] Matt Post. A call for clarity in reporting BLEU scores. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics, 2018.

[6] Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. Benchmarking and improving long-text translation with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7175–7187. Association for Computational Linguistics, 2024.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.