Kursorische Evaluation von eTranslation

21. Januar 2025

Tim Hagen und Martin Potthast

Deep Semantic Learning
Universität Kassel und hessian.AI

Versuchsaufbau

Modelle

- *eTranslation* [europa.eu]
  Übersetzungsdienst der EU
- *DeepL* [deepl.com]
  State-of-the-art kommerzieller Übersetzungsdienst
- *Google Translate* [google.com]
  State-of-the-art kommerzieller Übersetzungsdienst
- *GPT-4o* [openai.com]
  State-of-the-art LLM, welches wir “zero shot” für machine translation prompten

Datensatz

- Datensatz: Books-Korpus des “OPUS”-Benchmark [nlpl.eu]
- Evaluation auf 1262 der 51 467 Textschnipsel
- Quelle des Datensatzes: Urheberrechtsfreie Literatur und ihre Übersetzungen

Experiment 1: Evaluation

Wie ähnlich sind die Übersetzungsdienste zu einer Referenzübersetzung?

Ähnlichkeitsmaße

1. *BLEU* [Post 2018]
   Anteil der n-Gramme der Übersetzung, die auch in der Referenz auftauchen
2. *BERTScore* [Zhang et al. 2019]
   Semantische Ähnlichkeit der Worte der Übersetzung zu Worten in der Referenz

Ergebnisse

1. Sehr ähnlich in beiden Maßen
2. eTranslation am “schlechtesten”

Dienst	**Books Corpus**	
	BLEU	BERTScore
ChatGPT	19,7 %	82,8 %
Google Translate	18,7 %	82,2 %
DeepL	20,4 %	82,5 %
eTranslation	17,5 %	82,1 %

Experiment 1: Evaluation

Wie ähnlich sind die Übersetzungsdienste zu einer Referenzübersetzung?

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BLEU**):

1. 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.

Experiment 1: Evaluation

Wie ähnlich sind die Übersetzungsdienste zu einer Referenzübersetzung?

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BLEU**):

1. 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.

2. 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.

Experiment 1: Evaluation

Wie ähnlich sind die Übersetzungsdienste zu einer Referenzübersetzung?

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BLEU**):

1. 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.

2. 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.

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BERTScore**):

1. 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.

Experiment 1: Evaluation

Wie ähnlich sind die Übersetzungsdienste zu einer Referenzübersetzung?

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BLEU**):

1. 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.

2. 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.

Auszug der Übersetzungen mit der geringsten Ähnlichkeit (**BERTScore**):

1. 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.

2. Again I paused; then bunglingly enounced—

   eTranslation: Wieder pausierte ich; dann bunglingly enounced—

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

Experiment 2: Vergleich

Wie ähnlich sind die Übersetzungsdienste unter einander?

- Vergleiche Modelle mit einander statt der Referenzübersetzung
- Ergebnisse:
  1. Alle Modelle sind einander recht ähnlich
  2. DeepL und Google Translate sind einander am ähnlichsten

Dienst	**ChatGPT**		**Google Translate**		**DeepL**	
	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore
Google Translate	49,7 %	90,8 %	—	—	—	—
DeepL	49,9 %	90,5 %	53,5 %	91,9 %	—	—
eTranslation	46,5 %	90,3 %	48,8 %	90,3 %	45,0 %	89,4 %

Automatische Metriken sind nicht alles

- Autom. Metriken sind *heuristiken* und können Qualitätsunterschiede stark unterschätzen [Marchisio et al. 2024]
- eTranslation unterscheidet amerikanisches und britisches Englisch nicht
- Nutzungsfreundlichkeit
  - eTranslation langsamer als DeepL und mehr “Klicks” notwendig
  - ChatGPT leicht zu verwenden und sehr flexibel
  - DeepL und ChatGPT unterstützen den iterativen Prozess des Schreibens

Zusammenfassung & Fazit

- eTranslation ist
  - im Interface rudimentär aber
  - in der Übersetzungsqualität vergleichbar zum State-of-the-art
- Hauptunterschied zu DeepL: Usability

- eTranslation
  - gut für gelegentliche Nutzung
- DeepL (und ChatGPT)
  - unterstützen den iterativen Schreibprozess besser
  - sind angenehmer zu verwenden