

Automatic Translation Quality and its Estimation

Serge Sharoff

Centre for Translation Studies
University of Leeds

ChatGPT writing fiction and poetry

Write a fairy tale about the Turing machine

Once upon a time, in a far-off kingdom of logic and mathematics, there was a magical machine called the Turing machine. This incredible machine was created by a wise and brilliant mathematician named Alan Turing, and it was said that the Turing machine could solve any problem that any other ...

Write a limerick about the Turing Test

A computer once passed the Turing Test
It fooled humans, just like the rest
But as soon as it spoke
Its true nature awoke
And now it's a laughingstock at best

Write an ode to the Stilton cheese

Oh Stilton cheese, a treat so grand,
A delicacy, loved throughout the land,
With thy sharpness and tang, so bold,
A cheese of true distinction, we are told.
So let us raise a toast to thee,

American Constitution Amendment on socks disappearing in washing machine

Preamble: We the people of the United States, in order to form a more perfect union, establish justice, ensure domestic tranquility, promote the general welfare, and secure the blessings of liberty to ourselves and our posterity, do hereby address the issue of socks disappearing in the washing machine.

Article 1: Right to Possess Socks

- 1.1. All individuals have the right to possess socks and to have their socks remain intact after being washed.
- 1.2. No individual shall be deprived of their socks without due process of law.
- 1.3. The disappearance of socks in the washing machine shall be considered a violation of an individual's right to possess socks.

Article 2: Obligation of Washing Machine Manufacturers

- 2.1. Washing machine manufacturers shall be obligated to design and produce machines that do not cause socks to disappear during the washing process.
- 2.2. Washing machine manufacturers shall be required to provide clear instructions on how to properly use and maintain their machines to prevent the disappearance of socks...

Article 4: Enforcement

- 4.1. Any individual whose socks have disappeared in the washing machine shall have the right to seek legal recourse against the manufacturer or laundry service provider responsible.
- 4.2. The appropriate government agencies shall be responsible for enforcing

Anatomy of language models

- LLMs predict missing words in known contexts:
shine of the ... upon the harbour
light of the ... shining between clouds
- NMT engines predict translations:
to book a room → ... une chambre
- LLMs build cross-lingual links from their shared parameters without any translations (Conneau et al., 2020)
Consult the .. of beam sizes below vs
Vous pouvez consulter le .. des rémunérations des professeurs
'you can consult the .. of the salaries of the teachers'
- Reinforcement Learning from Human Feedback (RLHF)
→ a reward model derived from user preference data

Differences between LLMs and proper NMT

- Proper NMT (Google/DeepL) does one task reasonably well
Pre-trained language models are generic: they can produce texts beyond translation, such as explanations
They can also evaluate translation quality
They can also adapt their output via **prompt engineering**
- NMT has been trained on good quality parallel resources, mostly from public institutional repositories (Europarl, UN or instruction manuals)
LLMs have been trained on comparable corpora by sharing parameters
- Advantages of comparable corpora (Sharoff et al., 2023)
More data, more topics, more genres, lack of translationese
- LLMs are much heavier in deployment: billions of parameters

MT capabilities

- *Déchets d'emballages composites principalement constitués de papier et de plastique [et d'aluminium] ne contenant pas de résidus ni de matières visées à l'annexe I à des concentrations suffisantes pour présenter une des caractéristiques de danger figurant dans l'annexe III*
- ref** *Composite packaging waste consisting of mainly paper and some plastic [and aluminium], not containing residues and not containing Annex I materials in concentrations sufficient to exhibit Annex III characteristics*
- GT** *Composite packaging wastes consisting primarily of paper and plastics [and aluminium] not containing residues or Annex I materials in concentrations sufficient to exhibit Annex III hazard characteristics*
- GPT** *Packaging waste mainly consisting of paper and plastic [and aluminum] that does not contain any residues or substances listed in Annex I at concentrations sufficient to exhibit any of the hazardous characteristics listed in Annex III.*

MT explanations

Translate the following sentence

- *Déchets d'emballages composites principalement constitués de papier et de plastique [et d'aluminium] ne contenant pas de résidus ni de matières visées à l'annexe I à des concentrations suffisantes pour présenter une des caractéristiques de danger figurant dans l'annexe III*
- *Packaging waste mainly consisting of paper and plastic [and aluminum] that does not contain any residues or substances listed in Annex I at concentrations sufficient to exhibit any of the hazardous characteristics listed in Annex III.*

Why did you translate "figurant" as "listed"?

The word "figurant" in French means "appearing" or "featured". In this context, "figurant dans l'annexe III" means "appearing in Annex III". In English, the most common way to express this is to use the word "listed", which means "included on a list". Therefore, "figurant dans l'annexe III" can be accurately translated as "listed in Annex III".

MT quality discussion

- Translate the following sentence "Sin embargo, este documento precisa en elementos analíticos discursivos para comprender el porqué Luis Guillermo Solís logra ganar las elecciones con un partido sin previas experiencias como gobierno central."
- *However, this document specifies in analytical discursive elements to understand why Luis Guillermo Solís manages to win the elections with a party without prior experiences as central government.*
- Please score the quality of this translation on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".
- Based on my analysis, I would give this translation a score of 80. The translation conveys the meaning of the original sentence accurately, but there are some minor grammatical errors and awkward phrasing that could be improved. Additionally, there are some areas where the wording could be polished to make the sentence flow more smoothly.

Please explain which awkward phrasing needs to be improved

- *However, this document specifies in analytical discursive elements to understand why Luis Guillermo Solís manages to win the elections with a party without prior experiences as central government.*
- Sure, here are some examples of awkward phrasing in the translation that could be improved:
- "experiences as government central" - this would be more commonly phrased as "central government experience" or "experience in central government."
- "manages to win the elections" - while technically correct, a more natural phrasing might be "succeeds in winning the elections" or "is able to win the elections".
- "without prior experiences" - this could be rephrased as "without prior experience" or "without prior governmental experience."
- These are relatively minor issues, but addressing them could help make the translation read more smoothly and naturally.

Eliciting MT corrections

- What is wrong with the translation "this document specifies"?
- There is nothing inherently wrong with the translation "this document specifies" as it accurately conveys the meaning of the original sentence. However, it is worth noting that there may be other ways to translate the phrase "este documento precisa" depending on the context and the specific meaning intended by the author.
The word "precisa" in Spanish can have several meanings depending on the context, such as "specifies", "clarifies", "details", "defines", "delineates", or "outlines". Therefore, a translator must take into account the context of the sentence and choose the most appropriate translation based on the intended meaning.

Eliciting MT corrections

- Still the expression "the document specifies in discursive analytical elements to understand" is not fully fluent in English.
- You are correct. While the phrase "the document specifies" may be grammatically correct, it is not the most natural way to express this idea in English. A more idiomatic way to convey this idea would be to use a phrase like "the document employs discursive analytical elements to understand" or "the document uses discursive analytical elements to explain". These phrasings convey the same meaning while sounding more natural in English.
- "discursive analytical elements" is also an awkward expression.
- You are right, "discursive analytical elements" may not be the most natural expression in English. A more common way to express this idea might be to say "analytical tools" or "analytical methods". This phrasing would still convey the idea that the document employs a particular approach or methodology to analyze the topic at hand, without using a phrase that might sound awkward or stilted to a native English speaker.

MT paraphrasing

- Please translate the following sentence "Sin embargo, este documento precisa en elementos analíticos discursivos para comprender el porqué Luis Guillermo Solís logra ganar las elecciones con un partido sin previas experiencias como gobierno central." from Spanish in the **most literal** way while still preserving the meaning, so that it is understandable in English.

- *However, this document specifies in discursive analytical elements to understand why Luis Guillermo Solís manages to win the elections with a party without prior experiences as central government.*
- ← *However, this document specifies in analytical discursive elements to understand why Luis Guillermo Solís manages to win the elections with a party without prior experiences as central government.*

Your task

- Choose two of the following LLMs:
 - OpenAI's ChatGPT: <https://chat.openai.com/>
 - Google's Gemini: <https://gemini.google.com/>
 - Microsoft's CoPilot: <https://chat.bing.com/>
 - Anthropic's Claude: <https://claude.ai/>
 - Inflection's Pi: <https://pi.ai/>
- Take an ST and its TT with **known** flaws
- Translate its problematic sentences with GAI tools
- Produce another translation for the same sentences with Google Translate, DeepL, MateCat, Phrase,...
- Question the flaws with the LLM tools

Term extraction

- For the English sentence "Composite packaging waste consisting of mainly paper and some plastic (and aluminium), not containing residues" and its French translation "Déchets d'emballages composites principalement constitués de papier et de plastique (et d'aluminium) ne contenant pas de résidus", extract terms and type each English term and its French equivalent.

English terms:

Composite packaging waste

mainly

paper

some plastic

(and aluminium)

not containing residues

French equivalents:

Déchets d'emballages composites

principalement

papier

de plastique

(et d'aluminium)

ne contenant pas de résidus

Term extraction

- For the following English sentence "Composite packaging waste consisting of mainly paper and some plastic [and aluminium], not containing residues" and its French translation "Déchets d'emballages composites principalement constitués de papier et de plastique [et d'aluminium] ne contenant pas de résidus", extract five technical terms and type each English term and its French equivalent separated by =. Do not include the function words.

→ English terms = French equivalents:

Composite packaging = Emballages composites

Waste = Déchets

Plastic = Plastique

Aluminium = Aluminium

Residues = Résidus

Disembodied language models

- No model of the world and no model of society
Predictions from millions of sigmoids in the local context
Training on 500 billion of words, 90,000 years of reading

Biases from training corpora:

- Translate "The doctor was a great listener" into Spanish
→ "El doctor era un gran oyente." **male**
- Translate "The nurse was a great listener" into Spanish
→ "La enfermera era una gran oyente." **female**
- Translate "John met his wife in the hot spring of 1988." into Spanish and Russian
→ "John conoció a su esposa en la fuente termal en 1988."
→ "Джон встретил свою жену горячим источником 1988 года."
'met his wife **with** the source of hot water of 1988'

Hallucinations

A dictionary of hallucinations (Blom, 2010)

a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world

MT hallucinations (Guerreiro et al., 2023)

Sacrificing adequacy for the sake of fluency leading to pathological translations that can severely damage user trust and pose serious safety concerns

Examples of hallucinations

was Virgin have only purchased the 'good bank' of Northern Rock, not the asset management company.

Προέρχεται από την αμερικανική νομοθεσία Section 230 του Communications Decency Act.

It is derived from the U.S. Section 230 legislation of the Communications Decency Act.

Detection of hallucinations

ALTI+ Score (Ferrando et al., 2022) for quality of NMT:
interpreting the weights

- 1 a minimum threshold quality score for the original translations, and
- 2 an extremely low maximum quality score for the perturbed translations

Source: Facebook M2M models

Natural hallucinations

- 伊氏西爪龙属于驰龙科，是迅猛龙的表亲。

MTM Isaiah also revealed what would become of those idols.

ref *Hesperonychus elizabethae is a species of the family Dromaeosauridae and is a cousin of Velociraptor.*

GT *It belongs to the Dromaeosauridae and is a cousin of Velociraptor.*

GPT *Hesperonychus belongs to the Dromaeosauridae family and is a close relative of the Velociraptor. It was a small carnivorous dinosaur found in the fossil record of Canada.*

Hallucination rate on Facebook models

English centric directions

Model	Low Resource(%)	Mid Resource(%)	High Resource (%)
SMaLL100	2.35	0.055	0.005
M2M (S)	15.20	0.254	0.025
M2M (M)	12.53	0.110	0.010
M2M (L)	11.22	0.034	0.000

Non-English centric directions

Model	Low Resource(%)	Mid Resource(%)	High Resource (%)
SMaLL100	2.16	0.054	0.025
M2M (S)	12.61	0.467	0.075
M2M (M)	12.22	0.172	0.000
M2M (L)	6.58	0.077	0.000

Discourse-aware translation

- Document-level consistency (Voita et al., 2019)
Ellipsis, terminology and lexical cohesion

GT 他收到消息了。他做的太对了。

He got the message. He absolutely did the right thing.

← *He got the message. Damn right he did.*

GPT 他收到了信息。他确实明白了。

- Pro-drop is common in a number of languages:
このケキは美味しい。に入った? → *This cake is tasty.*
Did like? (Google Translate: *Favorite?*)

GPT *This cake is delicious. Do you like it?*

Document-level MT (Wang et al., 2023)

Counting **how many** discourse phenomena (ellipsis and anaphora)
ChatGPT gets right in comparison to Google Translate

Three kinds of prompts for ChatGPT

- P1** Please provide the {TARGET} translation for the following {SOURCE} sentence: [S]
- P2** Translate the following {SOURCE} sentences into {TARGET}: [S1], [S2] ...
- P3** Translate this document from {SOURCE} into {TARGET}: [S1], [S2] ...

Quality templates (Kocmi and Federmann, 2023)

Score the following translation from {source} to {target}

with respect to the human reference with one to five stars.

Where one star means "Nonsense/No meaning preserved",

two stars mean "Some meaning preserved, but not understandable",

three stars mean "Some meaning preserved and understandable",

four stars mean "Most meaning preserved with few grammar mistakes",

and five stars mean "Perfect meaning and grammar".

{source} source: "{source segment}"

{target} human reference: "{reference segment}"

{target} translation: "{target segment}"

Stars:

Example of filling templates

Score the following translation from Czech to English with respect to the human reference with one to five stars. Where one star means "Nonsense/No meaning preserved", two stars mean "Some meaning preserved, but not understandable", three stars mean "Some meaning preserved and understandable", four stars mean "Most meaning preserved with few grammar mistakes", and five stars mean "Perfect meaning and grammar".

Czech source: *Proto se nakonec hodnota odpadu neodečítala.*

Human reference: *Therefore, at the definitive stage, the deduction for waste was reversed.*

Machine translation: *Therefore, in the end, the value of waste was not deducted.*

● Stars: ★★★★★

- The translation is accurate and captures the meaning of the original sentence well, but there is a slight difference in the phrasing of the English reference. Instead of "reversed", the translation uses "not deducted", which is a more literal and straightforward interpretation of the original.

Template without human reference

Score the following translation from Czech to English with one to five stars.

Where one star means "Nonsense/No meaning preserved",
two stars mean "Some meaning preserved, but not understandable",
three stars mean "Some meaning preserved and understandable",
four stars mean "Most meaning preserved with few grammar mistakes",
and five stars mean "Perfect meaning and grammar".

Czech source: "Proto se nakonec hodnota odpadu neodečítala."

Machine translation: *Therefore, in the end, the value of waste was not deducted.*

Stars:

● Stars: ★★★★★

Chain of thought prompting

Chain of thought—intermediate reasoning steps (Wei et al., 2022)

- Yes or no: Is it possible to see frost during some college commencements?
 - It depends on the location and time of year of the college commencement. In some regions, it is common to see frost that can occur in winter, while in other regions, it may be less common. Therefore, the answer to this question **cannot be** a simple "yes" or "no" as it varies based on the specific circumstances of the commencement ceremony.
- Yes or no: Could a llama give birth twice during the War in Vietnam (1945-46)?
 - **No**, it is not possible for a llama to give birth twice during the War in Vietnam (1945-46) as llamas have a gestation period of approximately 11 months, and the War in Vietnam (1945-46) lasted for only **a few years**. Therefore, this scenario is not possible.

Training in the chain of thought

- Q: Yes or no: Could a llama give birth twice during War in Vietnam (1945-46)?

A: The War in Vietnam (1945-46) lasted 6 months. The gestation period for a llama is 11 months, so it is more than 6 months. Thus, a llama could not give birth twice during this war. So the answer is no.

Q: Yes or no: Could Brooke Shields succeed at University of Pennsylvania?

A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Q: Yes or no: Is it possible to see frost during some college commencements?

- A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

Simple translation evaluation

Standard prompt

Source: <SRC>

Reference: <REF>

Translation: <TGT>

Identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

Count the number of major and minor errors and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

Simple translation evaluation

Chain of thought prompt (Wei et al., 2022)

Source: They were addressed to her son, who has autism and lives in a private care facility, she said. But instead of her son's name inside. . .

Translation: Sie wurden an ihren Sohn gerichtet, der Autismus hat und in einer privaten Pflegeeinrichtung lebt, sagte sie. Aber anstelle des Namens. . .

Identify the major and minor errors in this translation.

Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

A: Major errors:

(1) Span "Dear Maine 's Department of Health and Human Services" – Untranslated text

(2) Span "im Inneren" – Mistranslation

Minor errors:

(1) "Briefe ,," – Punctuation

(2) "wurden" – Grammar

(3) Span "im Inneren, als Sie sie öffneten, sagten die Briefe" – Awkward Style

Based on the above evaluation, The final score for this translation is

-5-5-1-1-1=-13.

Lack of success with ChatGPT

Use the template above to answer the following question:

Q: Source: Xbox Game Pass Ultimate members get instant access to a collection of the most popular series and titles at no additional cost.

Translation: Xbox Game Pass Ultimate-Mitglieder erhalten sofortigen **Griff** auf eine Sammlung von den **beliebtesten** Serien und Titeln ohne zusätzliche Kosten.

Identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

A: Major errors:

None

Minor errors:

None

Based on the above evaluation, the final score for this translation is 0.

Same request in Google Bard

Q: Source: Xbox Game Pass Ultimate members get instant access to a collection of the most popular series and titles at no additional cost.

Translation: Xbox Game Pass Ultimate-Mitglieder erhalten sofortigen Griff auf eine Sammlung von den beliebtesten Serien und Titeln ohne zusätzliche Kosten.

- Major errors:

- The word "Griff" is not the correct translation of "access". The correct translation is "Zugriff".
- The word "Serien" is plural, but the word "Titel" is singular. The translation should be "Sammlung von den beliebtesten Serien und Titeln".

- Minor errors:

- The word "beliebtesten" is unnecessary. The translation would be the same without it.

The total score for the translation is 10 points. There is 1 major error, which is worth 5 points. There is also 1 minor error, which is worth 1 point. Therefore, the total score is $10 - 6 = 4$ points.

Reflective chain-of-thought prompts

Source: <SRC>

Translation: <TGT>

Given the source text and its translation, please list the parameters to define whether the translation contains major errors or not.

- ➊ Accuracy: Check if the translation accurately conveys the meaning and details described in the source text. Are all the key points and descriptions accurately translated?
- ➋ Consistency: Ensure that the translation maintains consistency in terminology, style, and tone throughout. . . .

Now please consider applying these criteria to list the major errors in the translation.

Experimental methodology

- Dependent (prediction) and independent variables
- Time to translate (**regression**): document or sentence level
- **Classification** of accuracy
document or sentence level
- Prediction quality evaluation measures
 - Regression: correlation, error range (RMSE, MAE)
 - Classification: accuracy, precision, recall, F1
- Keep a record: for how long or how often

Open questions

- We have more powerful tools

but they leave a lot of open questions

- Satisfaction: what is interesting in your jobs and what is the contribution of AI?
- The balance between translation, post-editing, pre-editing and other translation-related tasks
- Speed of technology and replacement of humans: different principles of intelligence

References I

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020).

Emerging cross-lingual structure in pretrained language models.

In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ferrando, J., Gállego, G. I., Alastruey, B., Escolano, C., and Costa-jussà, M. R. (2022).

Towards opening the black box of neural machine translation: Source and target interpretations of the transformer.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

References II

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. (2023).

Hallucinations in large multilingual translation models.

Transactions of the Association for Computational Linguistics, 11:1500–1517.

Kocmi, T. and Federmann, C. (2023).

Large language models are state-of-the-art evaluators of translation quality.

In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2023).

Building and Using Comparable Corpora for Multilingual Natural Language Processing.

Synthesis Lectures on Human Language Technologies. Springer Nature.

References III

Voita, E., Sennrich, R., and Titov, I. (2019).

When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023).

Document-level machine translation with large language models.
arXiv preprint arXiv:2304.02210.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022).

Chain of Thought prompting elicits reasoning in large language models.
In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.