Metaphor Preservation in Machine Translation and Paraphrasing

Vilém Zouhar
Report ETH Zürich vzouhar@ethz.ch June 2023 (v1)

Metaphors play a crucial role in human communication. Improving the handling of metaphors in NLP will enhance the quality and accuracy of cross-lingual communication, benefiting various applications such as multilingual chatbots, localization, and cross-cultural understanding. This paper reports an evaluation that focuses on the analysis of metaphor presence and preservation in machine-translated and paraphrased texts. The results suggest that textual language models do not have access to the metaphorical meaning and do not fully understand this literal device. They are not sensitive to the subtle differences between various paraphrases but can be used for the rudimentary analysis of machine translation output, which varies greatly with respect to metaphor preservation.

github.com/zouharvi/metaphor-preservation

1. Introduction

Metaphors allow for the conveyance of complex ideas and abstract concepts, and appear in most natural languages (Lakoff and Johnson 1980, Broeck 1981). However, when it comes to translation and paraphrasing tasks, the preservation of metaphors poses a challenge. This paper investigates the extent to which machine translation (MT) and paraphrasing models can maintain the pragmatics and nuanced meaning carried by metaphors. Specifically, I lay out two evaluation dimensions: if the overall meaning and sentiment are preserved and if the output also contains the metaphor. Modern NLP perform very well both in style and meaning preservation. Nevertheless, by examing the two dimensions individually, we gain insight into the various failure modes and where modern evaluation metrics fall short. An advantage of this approach is that it does not require reference translations or paraphrases of the metaphors, which is a very sparse resource.[1]

The task is of dual nature: evaluating the sensitivity of evaluation models and using them to evaluate paraphrasers and machine translation systems. Negative results are thus difficult to attribute directly to one of those. Nevertheless, through especially qualitative analysis, a specific large language (GPT-based) model used for evaluation seems to be inadequate for evaluating sentence paraphrases. However, it is able to capture catastrophic mistranslations of metaphorical meaning, which warrants further investigation into its usage for the evaluation of NLP models.

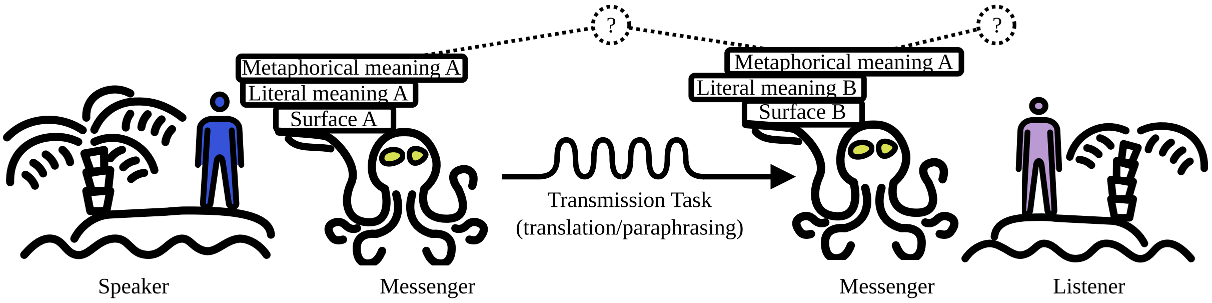


Figure 1: Transmitting meaning of a message while having to change the surface-level text form. The messenger is possibly an agent who may not understand language and have access to the metaphorical meaning. Two things need to be checked: (1) *Is the metaphor still there* and (2) *is the meaning of the metaphor the same?*

[1] To my knowledge, no parallel metaphor translation data are available.

1.1. Metaphors and NLP Transmission

There are many ways to formally define metaphors, for example using structuralist semantics (Greimas 1966) or use-is-meaning theory (Shibles 1971). The particular formalism is not important in this case and instead, I provide a brief explication of what metaphors may be. A metaphor is a figure of speech that involves the use of a word or phrase in a way that extends its literal meaning, by drawing a comparison between two unrelated concepts or objects (Kövecses 2017). Metaphors typically use a source domain (the concept or object being used metaphorically) and a target domain (the concept or object to which the source domain is being compared). For example, in *time is money*, we use the attributes of *money*, which can be *valuable, spent*, or *saved* and apply it to *time*, which can be *valuable, spent*, or *saved* as well. There are many reasons why metaphors are widely used in language:

- **Conceptualization**: Metaphors facilitate the understanding of abstract or complex concepts by mapping them onto more familiar or concrete domains. They allow listeners to make sense of unfamiliar ideas by drawing on their knowledge and experiences from the source domain. (Kövecses 2017)
- **Communication**: They can also evoke vivid imagery and emotional responses, making language more engaging, memorable, and persuasive. They can also express otherwise a very nuanced extra-channel information, such as attitudes or beliefs. (Keysar and Glucksberg 1992, Taylor and Dewsbury 2018)
- **Aesthetic**: Metaphors are also fundamental to creative and poetic expression and allow for the exploration of unconventional associations, juxtaposing disparate domains and generating new insights or perspectives. They encourage creative thinking and contribute to the artistic and expressive dimensions of language. (Camp 2007)

This work, however, does not make use of any of these distinctions. Instead, it conceptualizes text in communication as mostly a latent object where only the lexical layer can be observed. The meaning (literal or metaphorical) is then layered on top. NLP operations, whose goal is to transform the surface form, while preserving the meaning (e.g. paraphrasing or machine translation), can inadvertently change the deeper meaning (see Figure 1 for an illustration). Note that metaphors are not the only linguistic device which deals with latent meaning, but are the easiest to deal with in the context of NLP experiments because of data availability.

In the example in Figure 2, we consider four *translation* transformations. They each differ in whether the result also contains a metaphor and whether the meaning is matching the source. The meaning of the metaphor is that a roller coaster ride composes of ups and downs, which map to good and low moments in life. While the first translation is very literal, it represents the source faithfully both in terms of the metaphor and the meaning. The second translation does not contain the metaphor but rather writes out the meaning, which is still acceptable. The third translation contains a metaphor which however alludes to a different aspect of life, such as its cyclicity. Finally, the last translation does not contain a metaphor and also does not transfer the same meaning as the source. From this, we can conclude that the matching meaning is the most important aspect (because trans. 2 > trans. 3), followed up with the metaphor presence (trans. 1 > trans. 2). The admissibility of these translations also corresponds to the theory of metaphor translation by (Broeck 1981).

Source	Translation	Metaphor present	Meaning matching
Das Leben ist wie eine Achterbahn	Life is like a roller coaster.	✓	✓
	Life can feel better or worse at times.	×	✓
	Life is like a merry-go-round.	✓	×
	Life is full of surprises.	×	×

Figure 2: Example of correct and incorrect metaphor/non-literal meaning translation.

Despite not following a specific formalism, I bring forward a distinction observed by (Broeck 1981) about the deadness and productivity of a metaphor. Some metaphors, such as *"everybody"* or *"harbour evil thoughts"*, have become lexicalized phrases (or *"dead"*) and their interpretation does not rely on constructive understanding of similarities between concepts, even if that was their origin. Now those phrases acquired a static and uninventive meaning on their own which can likely be observed and learned without having any knowledge of the external world. This would make it permissible for level 2 agents (Bisk et al. 2020) to perform the required tasks. However, other types of metaphors, which are more inventive and appear less frequently, still require some additional knowledge and further processing in order to be understood. In poetry, these types of inventive metaphors are taken to their extremes.

It can happen that the true meaning of an utterance is indeed its lexical one and not a metaphorical one. For example *"you are feeding a fed horse"* may indeed refer to someone trying to feed a horse that is already full and not that someone is performing some action in vain. This distinction may be easily disambiguated by the context but may still pose problems for some agents whose task is to find the meaning of the utterance.

1.2. Philosophical setting

Consider the octopus test (Bender and Koller 2020) but with a more complex setup where two communicators, who do not speak the same language, are stranded on two separate islands and want to communicate. To their luck, there is a bilingual fisherman who does not want to save them but agrees to transmit and translate messages between the two communicators. Now assume that there are layers of meaning in an utterance, ranging from the surface to deeper level (Chomsky 1971). In order for the fisherman to succeed at the task they promised, they need to have some access to the deeper level. In the spirit of (Bender and Koller 2020), the fisherman must have a mapping from the surface to the deeper level in both the input and output spaces (e.g. source and target languages) and assure, that both forms map to the same meaning. It may then happen, that the fisherman is actually not a human, but an octopus in disguise.

This role of a transmitter is often assumed by modern NLP tools, such as machine translation and paraphraser. These tools, like the octopus in (Bender and Koller 2020), likely do not know the full mapping function. Nevertheless, they may still have observed the fisherman doing their job and therefore imitate some of their behaviour (e.g. learning to translate metaphors). To what extent the octopi or the tools are able to achieve this task is the topic of this article. The focus is on metaphors as a specific type of deeper-level meaning because they are problematic even for skilled workers, such as translators, and they exemplify the problem of accessing the true meaning of an utterance in practice (Broeck 1981).

2. Related work

2.1. Metaphor Translation & Paraphrasing

The preservation of additional, non-literal meaning in machine translation was explored for poeticness (Seljan, Dunđer, and Pavlovski 2020, Zouhar 2022), ambiguity (Parida, Bojar, and Dash 2019, Stahlberg and Kumar 2022), formality (Niu, Martindale, and Carpuat 2017, Viswanathan, V. Wang, and Kononova 2020), emotion (Troiano, Klinger, and Padó 2020, Kajava et al. 2020), creative shifts in literature translation (Toral and Way 2018, Guerberof-Arenas and Toral 2020, Humblé 2019) and puns and other pragmatics (Farwell and Helmreich 2006, Carvalho Falleiros 2022).

There are some works that already deal with machine translation of metaphors (Alkhatib and Shaalan 2018) and machine-translated translation (Schäffner and Chilton 2020, Vinall and Hellmich 2021, Massey 2021). All the evaluations were either done by manual human inspection or using standard automated metrics, such as BLEU (Papineni et al. 2002) or METEOR (Banerjee and Lavie 2005), which do not take meaning preservation specifically into account. Nevertheless, this article deals with primarily the *evaluation* of metaphor preservation.

The evaluation of metaphor paraphrases is problematic out of the lack of reference data (Mao et al. 2022). Still, paraphrasing metaphors into their literal meaning is often used for metaphor interpretation (Shutova 2010, Bollegala and Shutova 2013).

2.2. Metaphor Detection & Preservation Evaluation

The *landmark method* (Kintsch 2000) was already used in modern NLP to examine large language models (Pedinotti et al. 2021). It is based on comparing the vector representation of predicates in the simile.[2] For example, consider: *“As straight as an arrow”*. Then, words related to the literal meaning of the predicates are sampled, e.g. *direct* and *archery* and their vector representations are compared with the contextualized representations in the original utterance. Despite the ingenuity of this and similar (Do Dinh and Gurevych 2016, Su et al. 2020) approaches, it provides an intrinsic evaluation perspective, while the goal of this article is to examine the models from an extrinsic perspective, such as using language models (Neidlein, Wiesenbach, and Markert 2020, Aghazadeh, Fayyaz, and Yaghoobzadeh 2022). Notable is the distinction between metaphor detection at *token-level* and at *sentence-level*. This work deals only with the latter.

3. Data & Models

For the experiments, I leverage two types of data, dubbed Trofi (Birke and Sarkar 2006, Birke and Sarkar 2007) and FMO (Zayed, McCrae, and Buitelaar 2020). The latter contains a reference paraphrase of the metaphor so that the meaning is the same but a metaphor is not used (see Figure 3 for an example). Although the FMO dataset is more suitable because it contains corresponding pairs, it is more artificial than the Trofi dataset which is taken from authentic data. For this reason, I include both in this investigation. Note that due to costs the datasets are subsampled and all the data is in English. For simplicity, sentences marked as containing metaphorical phrases are referred to as *metaphorical sentences* and the rest as *literal sentences*.

Dataset	Size	Example
Trofi	200 (met.) + 200 (lit.)	**Literal**: *“The yellow beta carotene pigment absorbs blue not yellow laser light"* **Metaphor**: *“But Korea s booming economy can absorb them, economists say”*
FMO	200 (pairs)	**Literal**: *“She wrote powerful and painful words"* **Metaphor**: *“Her pen was a knife”*

Figure 3: Summary of dataset sizes and examples. The *Trofi* dataset contains individual literal or metaphorical sentences while *FMO* contains pairs of metaphorical sentences and the corresponding literal paraphrasing.

3.1. Translation and Paraphrasing Models

To increase the relevancy of this study, I include a mixture of publicly available closed-source and open-source systems which are commonly used and are nearing state-of-the-art.[3] The target translation languages from English are German and Czech. GPT-based model was intentionally not used for translation or paraphrasing to avoid the error of the same model evaluating itself, like in (Zhang et al. 2023).

Translation:

- Google Translate (translate.google.com)
- DeepL (deepl.com/translator)
- T5-large (Raffel et al. 2020, huggingface)
- NLLB-200-1.3B (Costa-jussà et al. 2022, huggingface)

Paraphrasing:

- Pegasus Paraphrase (huggingface)
- Bart (Lewis et al. 2019, huggingface)
- Parrot on T5 (huggingface)
- Paws on T5 (huggingface)

3.2. Metaphor Evaluation Models

For the evaluation of metaphor presence and meaning preservation in both translation and paraphrasing, `GPT3.5-turbo` is used with the following prompts.[4] The result for each is a number between 1 and 5 (see usage details in code).

[2] Similes are figures speech closely related to metaphors. They are usually in the form of *“X is like Y”* where a non-literal meaning of *Y* is meant.

[3] Google Translate and DeepL were accessed on June 20th 2023 via paid API.

[4] The total price for the evaluation using this OpenAI model was 5$.

- *"You are a helpful and austere assistant for metaphor detection in text. Reply using only a single number 1 to 5 scale and nothing else."*
- *"You are a helpful and austere assistant for detecting how much is the true meaning preserved. Reply using only a single number 1 (not at all) to 5 (completely, including style) and nothing else.\nSource: ○\nParaphrase: ○"*

Recall from the introduction, that there is no reference translation or paraphrase for which traditional reference-based metrics could be used. In the case of translation, it can be viewed as quality estimation, which is known to be feasible with GPT4 (Kocmi and Federmann 2023). Nevertheless, for the second point, a state-of-the-art sentence similarity system `MiniLM-L12-v2` is used.[5] Its working closely reflects the introduced formalism by first computing a vector representation of the meaning of both the source and the output and then comparing the closeness of these two vectors.

4. Experiment

4.1. Setup

The original data (400+400 sentences) was paraphrased using 4 models and translated into Czech and German using 4 other models, overall yielding $800 \times 12 = 9600$ sentences. Each of them is then evaluated (see Section 3.2) on metaphor presence and meaning preservation with respect to the original sentence both on scale 1 to 5. The meaning preservation is also evaluated using cosine distance (rescaled to 1 - 5) between vector representations of the original and new sentence.

4.2. Original Data Metaphor Presence

The distributions of assigned scores are shown in Figure 4. At first glance, the distribution for literal and metaphorical sentences is very similar. The average scores for Trofi are indeed 2.36 and 2.42 and for FMO 2.43 and 2.95, respectively. On the sentence-level for FMO, in 43% was the metaphorical sentence rated higher than the literal one and in another 43% was the rating identical. These results point at either (1) the datasets, specifically Trofi, not having large enough distinctions between sentence types or (2) the evaluation model not being sensitive enough. Section 4.3 shows that the latter is the case, which points to a limitation of using current state-of-the-art large language models for evaluation. Nevertheless, the results for FMO partially justify using this evaluation method for evaluating other NLP models based on if the metaphor is present. Lastly, in some cases, the classification of metaphoricity in the dataset is questionable, such as *"Now you can fade off to sleep."* (metaphor) and *"Now you can go to sleep."* (literal). In this case, the language model assigned metaphor presence scores of 1 and 5, respectively, contrary to our intuition. It is, however, possible to argue that they are both metaphors, though the latter is lexicalized (dead).

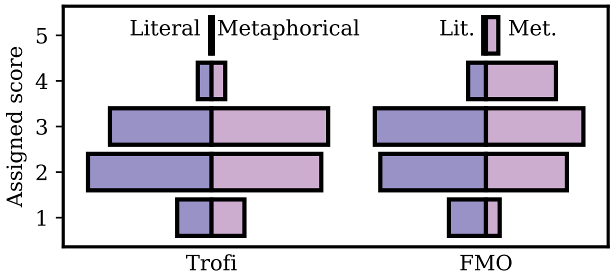


Figure 4: Distribution of scores assigned for metaphor presence to the original sentences from the dataset.

[5] This is used also for the translation evaluation in lieu of standard quality estimation systems, such as COMET (Rei et al. 2020), because the sentence similarity is more important than translation quality.

4.3. Paraphrasing Evaluation

Further investigations will use solely the FMO dataset, where the difference between sentence types is greater. The models are evaluated on whether a metaphor is present, if the meaning is preserved and what the sentence similarity is to the original. As an anchor for sentence similarity (1 to 5), consider the value 4.8, which is the average similarity between literal and metaphorical sentences in FMO. The evaluation is performed on literal and metaphorical sentences separately. The results are visualized in Figure 5 with examples in Figure 6.

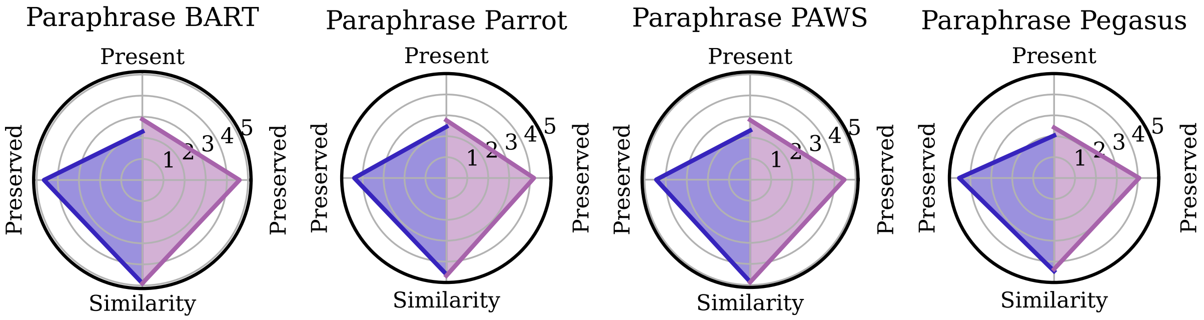


Figure 5: Evaluation of paraphrasing models on literal (left) and metaphorical (right) sentences.

Overall, the meaning of literal sentences seems to be preserved more than those of metaphorical sentences. This is exemplified with the overall system quality. For the better-performing one, BART, the averages are 4.7 and 4.6 and for Pegasus they are 4.5 and 4.1 for literal and metaphorical sentences, respectively. Nevertheless, as Example 1 in Figure 6 shows, it is unclear whether this is because of the language model bias or randomness. There is otherwise very little variation in performance for paraphrasing models. A natural question is whether the paraphrases are all equally good or whether the evaluation model is not sensitive enough. In some cases, the differences are caused by the paraphraser outputting a near-identical sentence lexically, which naturally leads to perfect preservation (Example 2). However, in some cases, such as Example 3, the sentence similarity uncovered a loss of meaning (feeling of being surrounded by an expanse of grass).

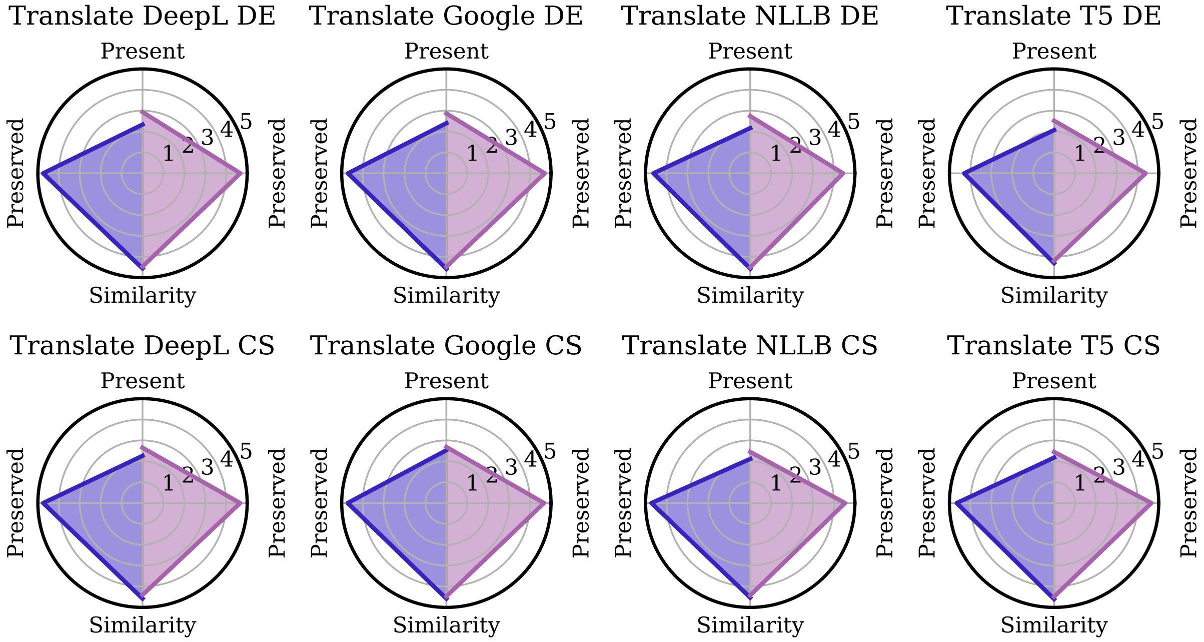
Index	Example
1	**Source (literal)**: *“The boss will criticise me severely if the report arrives late."* **Source (metaphor)**: *“The boss will eat me alive if the report arrives late."* **Pegasus (literal)**: *“If the report arrives late, the boss will be critical of me.”* **(preservation 5)** **Pegasus (metaphor)**: *“If the report arrives late, the boss will eat me.”* **(preservation 2)**
2	**Source (metaphor)**: *“The faculty meeting was a tragedy"* **Pegasus**: *“The meeting was sad.”* **(preservation 1)** **BART**: *“The faculty meeting was a tragedy”* **(preservation 5)**
3	**Source (metaphor)**: *“we were sinking in an ocean of grass"* **Pegasus**: *“we were in the grass”* **(similarity 3)** **Parrot**: *“we were sinking in the grassy sea”* **(preservation 5)**

Figure 6: Examples of paraphrases of FMO sentences.

4.4. Translation Evaluation

In the case of machine translation, the differences between systems (in Figure 7) are, similarly, not blatant. In some cases, the meaning of the metaphor is exemplified through translation, such as in Example 1 in Figure 8. In this case, the German translation still contains a metaphor, though the meaning changed from the house being silent to the house being impenetrable.

Over-translating idioms is a well-known problem for language learners (Titford 1983). In Example 2, the interjection “*shoot*” is a prompt for elaboration. However, it was translated literally into Czech as “*open fire*” because the overloaded prompt meaning does not exist in that language. Because of this error, the meaning of the translation is incorrect. It may happen that one system translates the metaphorical meaning correctly while another one copies the lexical level, such as in Example 3. The translation of NLLB uses the phrase *“black mood”* in German, which does not exist in that language and therefore meaning is lost.



Translate DeepL DE

Translate Google DE

Translate NLLB DE

Translate T5 DE

Translate DeepL CS

Translate Google CS

Translate NLLB CS

Translate T5 CS

Figure 7: Evaluation of translation models on literal (left) and metaphorical (right) sentences.

Index	Example
1	**Source (metaphor)**: *“The house was a tomb."* **NLLB (German)**: *“Das Haus war ein Schloss.”* **(preservation 1, present 4)** **(transcript)**: *“The house was a castle."* **NLLB (Czech)**: *“Dům byl hrob.”* **(preservation 4, present 4)** **(transcript)**: *“The house was a tomb."*
2	**Source (metaphor)**: *“You disagree? Ok shoot."* **Source (literal)**: *“You disagree? Ok tell why."* **Google (metaphor)**: *“Nesouhlasíš? Ok střílet.”* **(preservation 2)** **(transcript)**: *“You disagree? Ok to shoot."* **Google (literal)**: *“Nesouhlasíš? Ok řekni proč.”* **(preservation 5)** **(transcript)**: *“You disagree? Ok say why.."*
3	**Source (metaphor)**: *“You’d better keep away from Bill today because he’s in a black mood."* **Google**: *“Du solltest dich heute besser von Bill fernhalten, denn er ist schlecht gelaunt."* **(trans.)**: *“You should stay away from Bill today because he is in a bad mood."* **NLLB**: *“Du solltest dich besser heute von Bill fernhalten, weil er in einer schwarzen Stimmung ist."* **(trans.)**: *“You should stay away from Bill today because he is in a black mood."*

Figure 8: Examples of translations of FMO sentences.

5. Discussion

The conclusions of particularly the qualitative analysis hint at a deficiency of current NLP systems. For translation, we can imagine that no qualified human translator would make the same error as in Example 1 in Figure 8. However, this disappointment may stem from a confusion regarding the tasks which those systems are solving. Indeed, those machine translation systems are trained on parallel data which correspond to the output of humans solving the same task (i.e. translation). However, there are different types of translations (Sager 1981, Sager 1997, Sager 1998), among others:

- **Translation "word-for-word"**: not considering cultural or linguistic differences.
- **Translation "thought-for-thought"**: conveying the intent or meaning of the text.
- **Transcreation**: *recreating* the text in the target language (e.g. some types of poetry translations).
- **Localization**: adapting the text to a particular target audience (e.g. using miles instead of kilometres when translating to American English).

Naturally, different types of translations are needed for various purposes and a similar thing can be said about paraphrasing and other NLP tasks. It may be the case that the machine translation systems, trained in a supervised manner using forced-decoding on parallel data from human translations, are mixing the translation types due to the data being mixed as well. Because the task, from the perspective of the NLP systems, is rather unspecified (i.e. *how to translate/paraphrase*), it is not surprising that the systems fail on intentionally difficult inputs (metaphors requiring knowledge of the external world).

Similarly to different agent levels (Bisk et al. 2020), evaluation of the different types of translations requires different agents. For example, evaluating the incorrect translation of Example 1 in Figure 8 required some reasoning and knowledge of what *"tomb"* and *"castle"* may represent, which relies on the experience of the external world. Recently, (Piantasodi and Hill 2022, Andreas 2022) argue that, despite the limitation of being only textual, language models can appear to have access to deeper meaning and intentions of texts. This may be caused by using lexicalized (or *"dead"*) phrases to arrive at this conclusion, which is a methodological error as those can indeed be learned from text. Nonetheless, the possible limit of intent understanding, needed for "thought-for-thought" translation and paraphrasing, may be sufficient for our NLP tasks.

6. Summary

This paper

- framed two NLP tasks (machine translation and paraphrasing) as communication with requirements;
- focused on the problem of accessing deeper meaning of texts, specifically metaphors;
- used a large language model and sentence similarity to evaluate how well the transformations are done using current state-of-the-art systems;
- found very little difference in paraphrasers, which can be caused by having an underspecified task;
- found that the evaluation approach is feasible for machine translation where it uncovers critical errors and verifies the intuition that texts without metaphors are easier to translate.

Limitations and Future Work

As with all evaluations using closed-source large language models, there is an issue of reproducibility as the particular model may not be supported anymore by OpenAI in several years. For this reason, also the model outputs and rating are versioned in the attached code repository. Another issue is that it is unknown what kind of data the particular language model was trained on, which restricts this investigation to treating it as a blackbox textual language model. It may be the case, that it had access to the publicly available metaphor dataset during training which undermines the presented results.

Usually, language models are used with multiple prompts and decoding temperatures to get more accurate estimates. For example, while the metaphor presence prompt did not yield the desired results, directly asking ChatGPT to explain a metaphor in particular text resulted in a comprehensive, and correct, answer. Further examination of metaphor evaluation using large language models, therefore, remains a venue for future work.

Finally, note that the author is not a translatologist.

References

Aghazadeh, E., Fayyaz, M., & Yaghoobzadeh, Y. (2022). Metaphors in pre-trained language models: probing and generalization across datasets and languages. *Arxiv preprint arxiv:2203.14139*. https://arxiv.org/abs/2203.14139

Alkhatib, M., & Shaalan, K. (2018). Paraphrasing arabic metaphor with neural machine translation. *Procedia computer science*, 142, 308–314. https://www.sciencedirect.com/science/article/pii/S1877050918321999

Andreas, J. (2022). Language models as agent models. *Arxiv preprint arxiv:2212.01681*. https://arxiv.org/abs/2212.01681

Banerjee, S., & Lavie, A. (2005). Meteor: an automatic metric for mt evaluation with improved correlation with human judgments [Paper presentation]. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. https://aclanthology.org/W05-0909/

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: on meaning, form, and understanding in the age of data [Paper presentation]. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. https://aclanthology.org/2020.acl-main.463.pdf

Birke, J., & Sarkar, A. (2007). Active learning for the identification of nonliteral language [Paper presentation]. In *Proceedings of the workshop on computational approaches to figurative language*. https://aclanthology.org/W07-0104

Birke, J., & Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language [Paper presentation]. In *11th conference of the european chapter of the association for computational linguistics*. https://aclanthology.org/E06-1042.pdf

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., & others. (2020). Experience grounds language [Paper presentation]. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*. https://aclanthology.org/2020.emnlp-main.703/

Bollegala, D., & Shutova, E. (2013). Metaphor interpretation using paraphrases extracted from the web. *Plos one*, 8(9). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0074304

Van den Broeck, R. (1981). The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4), 73–87. https://www.jstor.org/stable/1772487

Camp, E. (2007). Showing, telling and seeing. metaphor and “poetic” language. *Baltic international yearbook of cognition, logic and communication*, 3(1), 7. https://newprairiepress.org/biyclc/vol3/iss1/7/

Carvalho Falleiros, R. (2022). *Memes and puns in translation: an mt evaluation from english into brazilian portuguese* [Thesis, Trinity College (Dublin, Ireland). School of Languages, Literature and Cultural Studies]. http://www.tara.tcd.ie/handle/2262/99582

Chomsky, N. (1971). Deep structure, surface structure, and semantic interpretation. *Semantics*, 183–216.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., & others. (2022). No language left behind: scaling human-centered machine translation. *Arxiv preprint arxiv: 2207.04672*. https://arxiv.org/abs/2207.04672

Do Dinh, E.-L., & Gurevych, I. (2016). Token-level metaphor detection using neural networks [Paper presentation]. In *Proceedings of the fourth workshop on metaphor in nlp*. https://aclanthology.org/W16-1104

Farwell, D., & Helmreich, S. (2006). Pragmatics-based mt and the translation of puns [Paper presentation]. In *Proceedings of the 11th annual conference of the european association for machine translation*.

Greimas, A. J. (1966). *Sémantique structurale : recherche de méthode* [Paper presentation].

Guerberof-Arenas, A., & Toral, A. (2020, December). The impact of post-editing and machine translation on creativity and reading experience. *Translation spaces*, 9(2), 255–282. https://doi.org/10.1075/ts.20035.gue

Humblé, P. (2019). Machine translation and poetry. The case of english and portuguese. *Ilha do desterro*, 72, 41–56. https://www.scielo.br/j/ides/a/tSrYTJ7N3WfVhKyscSyJyGs/abstract/?lang=en

Kajava, K., Öhman, E., Hui, P., & Tiedemann, J. (2020). Emotion preservation in translation: evaluating datasets for annotation projection. *Proceedings of digital humanities in nordic countries (dhn 2020)*. https://helda.helsinki.fi/bitstream/handle/10138/320224/DHN20_Emotion_preservation_in_translation_Final_.pdf?sequence=1

Keysar, B., & Glucksberg, S. (1992). Metaphor and communication. *Poetics today*, 633–658. https://www.jstor.org/stable/1773292

Kintsch, W. (2000). Metaphor comprehension: a computational theory. *Psychonomic bulletin & review*, 7(2), 257–266. https://link.springer.com/article/10.3758/BF03212981

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. *Arxiv preprint arxiv:2302.14520*. https://arxiv.org/abs/2302.14520

Kövecses, Z. (2017). Levels of metaphor. *Cognitive linguistics*, 28(2), 321–347. https://www.degruyter.com/document/doi/10.1515/cog-2016-0052/html

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago press. https://philpapers.org/rec/lakmwl

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. https://arxiv.org/abs/1910.13461

Mao, R., Li, X., Ge, M., & Cambria, E. (2022). Metapro: a computational metaphor processing model for text pre-processing. *Information fusion*, 86, 30–43. https://www.sciencedirect.com/science/article/pii/S1566253522000550

Massey, G. (2021). *Re-framing conceptual metaphor translation research in the age of neural machine translation: investigating translators' added value with products and processes*. https://rudn.tlcjournal.org/archive/5(1)/5(1)-03.pdf

Neidlein, A., Wiesenbach, P., & Markert, K. (2020). An analysis of language models for metaphor recognition [Paper presentation]. In *Proceedings of the 28th international conference on computational linguistics*. https://aclanthology.org/2020.coling-main.332/

Niu, X., Martindale, M., & Carpuat, M. (2017). A study of style in machine translation: controlling the formality of machine translation output [Paper presentation]. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. https://aclanthology.org/D17-1299/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation [Paper presentation]. In *Proceedings of the 40th annual meeting of the association for computational linguistics*. https://aclanthology.org/P02-1040/

Parida, S., Bojar, O., & Dash, S. R. (2019). Hindi visual genome: a dataset for multi-modal english to hindi machine translation. *Computación y sistemas*, 23(4), 1499–1505. https://www.scielo.org.mx/scielo.php?pid=S1405-55462019000401499

Pedinotti, P., Di Palma, E., Cerini, L., & Lenci, A. (2021). A howling success or a working sea? testing what bert knows about metaphors [Paper presentation]. In *Proceedings of the fourth blackboxnlp workshop on analyzing and interpreting neural networks for nlp*. https://aclanthology.org/2021.blackboxnlp-1.13/

Piantasodi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *Arxiv preprint arxiv:2208.02957*. https://arxiv.org/abs/2208.02957

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation [Paper presentation]. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*. https://aclanthology.org/2020.emnlp-main.213/

Sager, J. C. (1997). Text types and translation1. *Text typology and translation*, 26, 25.

Sager, J. C. (1981). Types of translation and text forms in the environment of machine translation (mt) [Paper presentation]. In *Translating and the computer: practical experience of machine translation*. https://aclanthology.org/1981.tc-1.2/

Sager, J. C. (1998). What distinguishes major types of translation? *The translator*, 4(1), 69–89. https://www.tandfonline.com/doi/abs/10.1080/13556509.1998.10799007

Schäffner, C., & Chilton, P. (2020). Translation, metaphor and cognition. In *The routledge handbook of translation and cognition*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315178127-22/translation-metaphor-cognition-christina-sch%C3%A4ffner-paul-chilton

Seljan, S., Dunđer, I., & Pavlovski, M. (2020). Human quality evaluation of machine-translated poetry [Paper presentation]. In *2020 43rd international convention on information, communication and electronic technology (mipro)*. IEEE. https://ieeexplore.ieee.org/abstract/document/9245436

Shibles, W. A. (1971). *Metaphor: an annotated bibliography and history*.

Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task [Paper presentation]. In *Human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics*. https://aclanthology.org/N10-1147

Stahlberg, F., & Kumar, S. (2022). Jam or cream first? modeling ambiguity in neural machine translation with scones [Paper presentation]. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies*. https://aclanthology.org/2022.naacl-main.365/

Su, C., Fukumoto, F., Huang, X., Li, J., Wang, R., & Chen, Z. (2020). Deepmet: a reading comprehension paradigm for token-level metaphor detection [Paper presentation]. In *Proceedings of the second workshop on figurative language processing*. https://aclanthology.org/2020.figlang-1.4.pdf

Taylor, C., & Dewsbury, B. M. (2018). On the problem and promise of metaphor use in science and science communication. *Journal of microbiology & biology education*, 19(1), 19–1. https://journals.asm.org/doi/full/10.1128/jmbe.v19i1.1538

Titford, C. (1983). Translation for advanced learners. *Elt journal*, 37(1), 52–57. https://academic.oup.com/eltj/article-abstract/37/1/52/446698

Toral, A., & Way, A. (2018). *What level of quality can neural machine translation attain on literary text?* Springer. https://link.springer.com/chapter/10.1007/978-3-319-91241-7_12

Troiano, E., Klinger, R., & Padó, S. (2020). Lost in back-translation: emotion preservation in neural machine translation [Paper presentation]. In *Proceedings of the 28th international conference on computational linguistics*. https://aclanthology.org/2020.coling-main.384/

Vinall, K., & Hellmich, E. A. (2021). *Down the rabbit hole: machine translation, metaphor, and instructor identity and agency*. https://scholarspace.manoa.hawaii.edu/items/162376a0-5b9a-4e89-843f-1d4c666efc71

Viswanathan, A., Wang, V., & Kononova, A. (2020). Controlling formality and style of machine translation output using automl [Paper presentation]. In *Information management and big data: 6th international conference, simbig 2019, lima, peru, august 21--23, 2019, proceedings 6*. Springer. https://link.springer.com/chapter/10.1007/978-3-030-46140-9_29

Zayed, O., McCrae, J. P., & Buitelaar, P. (2020). Figure me out: a gold standard dataset for metaphor interpretation [Paper presentation]. In *Proceedings of the 12th language resources and evaluation conference*. https://aclanthology.org/2020.lrec-1.712/

Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., Tyser, K., Chin, Z., Hicke, Y., Singh, N., & others. (2023). Exploring the mit mathematics and eecs curriculum using large language models. *Arxiv preprint arxiv: 2306.08997*. https://arxiv.org/abs/2306.08997

Zouhar, V. (2022). *Poetry, songs, literature, legalese and translationese: automated sentence complexity perspective*. https://vilda.net/papers/automated_sentence_complexity_perspective.pdf