Technical Terminology Verification for Neural MT

Anonymous EACL submission

Abstract

Translating technical acronyms is a problematic task for MT systems, with an error rate around 50% for Google Translate and around 65% for Opus-mt. Incorrect acronym translation is a fatal error. We present a turnkey solution for translating long form (LF)–short form (SFs) pairs and verifying their use by the scientific community. Since MT models perform better on LFs than SFs, our proposed method takes advantage of this observation to improve translations of SFs, by introducing a novel verification process. This process is motivated by standard practice in professional translation.

1 Introduction

While large language models are remarkably fluent, there are challenges with hallucinations (Church and Yue, 2023). With hallucinations defined as "[generated] text that is factually incorrect or nonsensical,"[1] we postulate that technical acronym translation errors are a form of hallucination. Similarly to with LLMs, priorities in the machine translation field must shift to address problems with technical acronym (and, more generally, technical term) hallucinations if we are to propose systems that perform at the level of a professional translator.

The potential sources of error a professional translator might encounter on a daily basis are not being properly evaluated by the metrics used by neural MT practitioners, namely BLEU, COMET and the computation of loss (Callison-Burch et al., 2006). Several workshops on terminology stress the importance of correctly addressing terminology issues—including correctness of technical terms—in the machine translation space (Molchanov et al., 2021; Hasler et al., 2018). The evaluation strategy for the "Machine Translation using Terminologies" workshop (Jon et al., 2021) states that "it will focus on both translation accuracy and consistency."[2] Additionally, current neural MT requires copious amounts of human-generated translations in order to successfully translate domain-specific terminology (Elliott et al., 2004). In this paper, we present a path forward for dealing with technical MT inaccuracies that better aligns with the stringent quality control of a human translator.

2 Towards Professional-Level Translation

Technical terminology is important to professional translators. BLEU (Papineni et al., 2002) and COMET (Amrhein and Sennrich, 2022) have other priorities. Hangya et al. (2021) found that "improving the translation of a few selected words [e.g., technical terms] could lead even to a slight drop in BLEU." Additionally, reliance on human-annotated reference translations is a major hindrance to accelerating MT improvement and generalizing it to low resource languages (Agić and Vulić, 2019).

Obviously, some errors are more serious than others. There is a considerable literature on Responsible AI (O'Neil, 2016; Bender et al., 2021; Blodgett et al., 2020). Most errors may not be that serious, but some errors can be offensive, and can lead to lawsuits[3] and product cancellations.[4]

This paper will focus on the translation of acronyms, a special case of technical terminology. We will introduce a fact-checking step to verify the combination of the long form (LF) and short form (SF) in at least two published articles in the target language. More generally, we believe there is an opportunity to use search to fact-check assertions from chatbots to reduce the risk of hallucinations.

There is an asymmetry in professional transla-

[1] https://towardsdatascience.com/llm-hallucinations-ec831dcd7786

[2] https://www.statmt.org/wmt21/terminology-task.html

[3] https://www.zdnet.com/article/microsoft-sued-for-racist-application/

[4] https://en.wikipedia.org/wiki/Tay_(chatbot)

tion (Pokorn, 1998). Most translators are stronger in one language than the other. They prefer to translate from their weaker language and into their stronger language. This asymmetry is rarely mentioned in the literature on machine translation, though there may be a motivation for the asymmetry under the proposed verification step. The proposed verification step is performed on the target language and not on the source language. Verification can take advantage of massive amounts of data in the target language, where available.

2.1 A New Test Set for Translating Acronyms

A new test set[5] has been created for evaluating machine translation systems on acronyms. The test set consists of 437 LF-SF pairs obtained from a corpus of 13,500 abstracts crawled from HAL,[6] a repository of French academic papers, many of which are from medicine and science. The examples were all hand-picked by the authors so as not to include any offensive content or personal information.

The repository provides abstracts in both French and English. These abstracts contain many technical terms. An example of an abstract is "[...] 42/194 patients (21%) did not want **cardiopulmonary resuscitation (CPR)** and 15/36 (41%) did not prefer intensive care unit (ICU) admission [...]." When the abstract introduces an acronym, the gold labels in the test set specify the long form (LF) and the short form (SF) in both French and English. The acronym translation task is illustrated in Table 1.

Input LF	Input SF	Gold SF
réanimation cardiopulmonaire	RCP	CPR

Table 1: The acronym translation task inputs LFs and SFs in French and outputs a candidate SF in English. Ideally, the candidate will agree with the gold label.

For evaluation purposes, we distinguish *agreement* from *verification*.

Agreement The candidate SF is an exact match with the gold SF.

Verification The candidate SF was found near the LF in at least two published papers in the target language (English).

[5]This test set will be posted in GitHub after the paper has been accepted. In the meantime, the test set can be found in the supplemental materials.

[6]https://theses.hal.science/?lang=en

We will use a search process to verify candidate SFs. As will be discussed later in Section 2.5, verifying acronyms can be viewed as a special case of fact-checking. Using search to fact-check assertions in ChatGPT output may also be a promising path forward for addressing hallucinations.

2.2 Google Translate

How well do commercial machine translation products work on technical terms? Table 2 shows that Google Translate[7] is more successful on long forms than short forms. Though there is considerable room for improvement in both cases (as illustrated in Tables 3-4), this paper will focus on SFs, where there is more opportunity for improvement.

Type of Term	Agreement
Long Forms (LFs)	62.1%
Short Forms (SFs)	54.3%

Table 2: Google is better on LFs than SFs

Input French	Output English	
	Google	Gold
indice moteur	engine index	motricity index
fréquence cardiaque	cardiac frequency	heart rate
roue polaire	polar wheel	claw pole

Table 3: Google errors on long forms (LFs)

Input French	Output English	
	Google	Gold
AOMI	PAAD	PAD
DE	DE	EE
ICMI	CIMI	CLI

Table 4: Google errors on short forms (SFs)

2.3 Proposal for Translating Acronyms

The proposed method decomposes translation into four steps. This decomposition takes advantage of the fact that Google Translate is more successful on long forms than short forms.

1. Use Google to translate LFs from FR to EN.

[7]https://translate.google.com/

2. Extract the EN LF from Google's output.
3. Generate candidate SFs from the EN LF.
4. Use search to verify candidates.

The first two steps are self-explanatory; the last two steps will be described in Sections 2.4-2.5.

2.4 Step 3: SF Candidate Generation

We use a fine-tuning process to generate SF candidates. We start with a pre-trained model, Scibert (Beltagy et al., 2019), and fine-tune a fill-mask task with the data formatted as illustrated in Table 5[8].

Scibert was trained on 1,800,000 term-acronym pairs with Adam as the optimizer, an initial learning rate of 2e-5, 1,000 warmup steps, and a weight decay of 0.01. The training data was obtained from arXiv papers[9] processed by AB3P[10] (Sohn et al., 2008; Church and Liu, 2021). After fine-tuning, the post-trained model can input strings of the form: "LF ([MASK])" and output n-best lists of candidates for the appropriate SF.

Input: LF ([MASK])	**Gold SF**
cardiopulmonary resuscitation ([MASK])	CPR
deoxyribonucleic acid ([MASK])	DNA
Organization of the Petroleum Exporting Countries ([MASK])	OPEC

Table 5: Training data for SF candidate generation.

Proposed method for acronym translation: First, the LF-SF pair is translated using the Google Translate API and Opus-mt model[11] in the format "acide désoxyribonucléique (ADN)". Second, the translated term pair is searched for as an exact match to see if it is used by domain experts in multiple published papers. Search is performed on output from AB3P of crawls of Pubmed and arXiv containing acronyms, their long forms, and document IDs. If insufficient evidence is provided for the use of the generated acronym translation (fewer than two document IDs where the pair was found), generate a list of candidate acronym translations from the machine translated LF using the fine tuned version of Scibert and verify each candidate translation in the list through search.

[8]The model will be posted after the paper is accepted

[9]https://info.arxiv.org/help/bulk_data/index.html

[10]https://github.com/ncbi-nlp/Ab3P

[11]https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

2.5 Step 4: Verification (Fact Checking)

The professional human technical term translation process involves a significant component of researching the meaning of a source language term, identifying multiple target language candidate terms, and finally, proceeding through the n-best list in order and seeking out the use of a chosen term in context in similar target language texts, written by experts in the field in question.[12] According to Bowker (2021), verification is done on the basis of observed frequency in a corpus; if enough experts use the selected term in context, it is considered to be valid. We replicate that process using search.

We implemented a Boolean retrieval system containing acronyms extracted from AB3P output on a crawl of arXiv and Pubmed along with the long forms they map to and source paper ID. If a sufficient number of sources have been found to employ the desired term-acronym pair (in the form *cardiopulmonary resuscitation (CPR)*), term validation is deemed to be successful and the term pair is returned to the user alongside the list of sources for verification. This re-appropriates the term verification method employed by professional translation agencies in the field (and facilitates verification by a reviewer, who may need to fact check term sources at a later stage).

3 Evaluation

3.1 Baselines

Baseline	**Input**	**Output**
Identity	ADN	ADN
Reverse	ADN	NDA
Google/Opus	acide désoxyribonucléique (ADN)	DNA

Table 6: Examples of three baseline methods

Table 7 compares outputs from the proposed method with the three baselines (in Table 6).

Identity: The candidate SF (EN) = input SF (FR).
Reverse: Same as above, but reverse input SFs.
Google/Opus: Use the given system to translate the SF in context. That is, we provide the model/API with both the LF and the SF (in French), with the SF in parentheses. Then we

[12]https://www.technitrad.com/how-to-perform-terminology-research/

extract an SF from the generated translation (in English) and use that as the candidate SF.

Due to the high technicity of many terms, our retrieval system was unable to verify a number of term pairs, including many of the gold labels. In several cases, translating the term through Google allowed us to obtain the correct SF via a more common concurrent LF than the gold label, which resulted in the proposed method beating the gold verified percentage.

Method	Agreement	Verified
Identity Baseline	21.5%	0.06%
Reverse Baseline	28.5%	14.6%
Opus Baseline	34%	14.9%
Google Baseline	54.3%	29.2%
Gold Labels	100%	42%
Proposed (Opus)	43.9%	32.7%
Proposed (Google)	**62.6%**	**42.8%**

Table 7: Proposed method outperforms all baselines

3.2 Results

Table 7 shows the proposed method is well above the three baselines when verification succeeded in finding evidence for one of the candidates in at least two published papers. In Table 8, we report precision as the portion of agreed terms which were verified and recall as the portion of verified terms.

Method	Precision	Recall
Identity Baseline	0.28	0.06
Reverse Baseline	0.51	0.15
Opus Baseline	0.43	0.15
Google Baseline	0.54	0.29
Gold Labels	0.42	0.42
Proposed (Opus)	0.75	0.33
Proposed (Google)	**0.68**	**0.43**

Table 8: Precision and Recall

4 Related Work

Anastasopoulos et al. (2021) stress the importance of taking terminology into account in neural MT and propose metrics to measure MT output consistency with regard to domain constraints. Dagan and Church (1994) propose a system to identify technical terms in a source text as well as their translations. The system uses part-of-speech tagging and word alignment techniques to assist translators during the translation process. Smadja et al. (1996) address the issue of translating collocations in a variety of domains.

Grefenstette (1999) offers an example-based method for dealing with terminology problems in translation as well as other NLP tasks. The method proposed uses search to find the most statistically likely translation of an entire noun phrase. Lee and Kim (2002) provide a knowledge-based approach to translation that includes using word-sense disambiguation to semantically derive the meaning of a word before seeking a target translation corresponding to that meaning.

Skadiņš et al. (2013) demonstrate the use of a cloud-based terminology search system that fully integrates with statistical methods to address the need for domain-specific terms and their integration into neural MT systems. Meanwhile, Bosca et al. (2014) stress the importance of term verification and consistency in the translation process and propose using external terminological databases to assist in fact checking and correcting domain-specific terminology.

5 Conclusion

A technical translator's job is more akin to that of a terminologist than to that of a bilingual copywriter (Cabré, 2010). Target text quality depends not on producing more fluent texts, but rather on translating and verifying technical terminology correctly and according to domain-specific standards (GHENŢULESCU, 2015). This is a problem not easily solved with parallel corpora, as acquiring such data is an expensive and laborious process.

Neither translators nor neural MT practitioners are benefiting from the current situation. As NLP systems take on increasingly challenging tasks, the need for guidance by domain experts becomes all the more important (Van den Broek et al., 2021). Moreover, the dismissal of professional translation concerns in favor of higher BLEU/COMET scores and lower loss by NLP and AI experts only discourages the type of collaboration we call for here. Terminology verification is a major cause for concern in technical translation, one for which we have outlined a path forward through search. Other issues of a stylistic nature are beyond the scope of this paper but may be addressed in future work.

Limitations

The results of applying our method may not transfer to languages that are very different from English in orthography (e.g., Chinese, Japanese) and/or morphology. Our solution also may not scale to longer texts; the method is based on working with acronym pairs and working on a full text would require a preprocessing step to identify term pairs as well as inference time for each acronym. Training a model for this task also requires access to GPU resources. Additional information about model parameters, hardware used, and number of training examples is available in the supplemental materials.

Ethics Statement

In line with the concept of professional translator ethics presented by Lambert (2020), it is of paramount importance to guard against translations that "represent their source texts in unfair ways." This refers to unfaithful translations that do not correctly transfer the true meaning in the source language, a prime example being incorrect or unverifiable terminology. Our system upholds this doctrine of translation ethics and adheres to ethics policies outlined by the translation community.

References

Željko Agić and Ivan Vulić. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet. *arXiv preprint arXiv:2202.05148*.

Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, et al. 2021. On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Alessio Bosca, Vassilina Nikoulina, and Marc Dymetman. 2014. A lightweight terminology verification service for external machine translation engines. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–52.

Lynne Bowker. 2021. Machine translation literacy instruction for non-translators: A comparison of five delivery formats. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 25–36, Held Online. INCOMA Ltd.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Kenneth Church and Boxiang Liu. 2021. Acronyms and opportunities for improving deep nets. *Frontiers in Artificial Intelligence*, 4:732381.

Kenneth Ward Church and Richard Yue. 2023. Emerging trends: Smooth-talking machines. *Natural Language Engineering*, 29(5):1402–1410.

Ido Dagan and Kenneth Church. 1994. Termight: Identifying and translating technical terminology. In *Fourth Conference on Applied Natural Language Processing*, pages 34–40.

Debbie Elliott, Anthony Hartley, and Eric Atwell. 2004. A fluency error categorization scheme to guide automated machine translation evaluation. In *Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004. Proceedings 6*, pages 64–73. Springer.

Lect Raluca GHENŢULESCU. 2015. The importance of terminology for translation studies. *In the Beginning Was the Word". On the Linguistic Matter of Which the World Is Built. Bucureşti: Ars Docendi*, pages 54–61.

Gregory Grefenstette. 1999. The world wide web as a resource for example-based machine translation tasks. In *Proceedings of Translating and the Computer 21*.

Viktor Hangya, Qianchu Liu, Dario Stojanovski, Alexander Fraser, and Anna Korhonen. 2021. Improving machine translation of rare and unseen word senses. In *Proceedings of the Sixth Conference on Machine Translation*, pages 614–624.

Eva Hasler, Adrià De Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. *arXiv preprint arXiv:1805.03750*.

Josef Jon, Michal Novák, João Paulo Aires, Dušan Variš, and Ondřej Bojar. 2021. Cuni systems for wmt21: Terminology translation shared task. *arXiv preprint arXiv:2109.09350*.

Joseph Lambert. 2020. Professional translator ethics. *The Routledge Handbook of Translation and Ethics Routledge*, pages 165–179.

Hyun Ah Lee and Gil Chang Kim. 2002. Translation selection through source word sense disambiguation and target word selection. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. Promt systems for wmt21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841.

Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nike Kocijancic Pokorn. 1998. Translation into a non-mother tongue in translation theory: Deconstruction. In *Translation in context: selected contributions from the EST Congress, Granada*, pages 61–72.

Raivis Skadiņš, Mārcis Pinnis, Tatiana Gornostay, and Andrejs Vasiļjevs. 2013. Application of online terminology services in statistical machine translation. In *Proceedings of Machine Translation Summit XIV: Posters*.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1):1–10.

Elmira Van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2021. When the machine meets the expert: An ethnography of developing ai for hiring. *MIS quarterly*, 45(3).