On Translating Technical Terminology: A Translation Workflow for Machine-Translated Acronyms

Richard Yue yue.r@northeastern.edu
Northeastern University, San Jose, CA

John E. Ortega j.ortega@northeastern.edu
Institute for Experiential AI, Northeastern University, Boston, MA

Kenneth Ward Church k.church@northeastern.edu
Institute for Experiential AI, Northeastern University, Boston, MA

Abstract

The typical workflow for a professional translator to translate a document from its source language (SL) to a target language (TL) is not always focused on what many language models in natural language processing (NLP) do – predict the next word in a series of words. While high-resource languages like English and French are reported to achieve near human parity using common metrics for measurement such as BLEU and COMET, we find that an important step is being missed: the translation of technical terms, specifically acronyms. Some state-of-the art machine translation systems like Google Translate which are publicly available can be erroneous when dealing with acronyms – as much as 50% in our findings. This article addresses acronym disambiguation for MT systems by proposing an additional step to the SL–TL (FR–EN) translation workflow where we first offer a new acronym corpus for public consumption and then experiment with a search-based thresholding algorithm that achieves nearly 10% increase when compared to Google Translate and OpusMT.

1 Introduction

With the myriad of artificial intelligence tools available for professional translators, it can be hard for translators to select solutions that address their core needs. Ideally, translation approaches based on machine learning techniques should improve translator proficiency and achieve higher overall quality. One such approach focuses on *technical terminology* (TT) where domain-specific terms in the form of acronyms in a source language (SL) must be translated into their target language (TL) counterpart.

TT is considered important to translators as it is one of the main sources of error a professional translator might encounter on a daily basis. The importance of TT is further displayed by the latest machine translation (MT) workshops (Semenov et al., 2023; Molchanov et al., 2021; Hasler et al., 2018) that stress the importance of correctly addressing terminology issues—including correctness of technical terms. While modern MT systems do not seem to focus on acronym and term disambiguation[1], workshops like the "Machine Translation using Terminologies" workshop[2] (Jon et al., 2021) clearly state that they focus on both translation accuracy and consistency. Since the dominant metric used (BLEU) for most MT approaches does not center so much on terminology expansion with acronyms and other mechanisms, we present in this article a novel method that hones in specifically on the day-to-day work in terminology that a professional translator may encounter, which has not been

[1] MT research generally use metrics such as BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020).

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

addressed by most of the recent literature.

In this article, we present two main novelties that are based on the translation of acronyms: (1) the introduction of a new corpus made publicly available for others to use and (2) a fact-checking step that is used to verify the combination of a technical term and its acronym (long form (LF) and short form (SF)). We do this for several published articles in the TL, which is English. We aim to show that acronym disambiguation can improve term error rate by reducing the risk of default MT models that generally do not have an acronym approach. Our claim is that translators can use this method as a novel verification step in the normal translation pipeline. We also believe that other automated work such as generative MT may be able to include this step as a mechanism for evaluation.

To that end, we present the following sequence. First, we introduce relevant work in Section 2. Second, we describe our motivation and high-level proposal methodology in Section 3. We then cover the details of our corpus creation in Section 4.4. Afterwards, we show the results of our SF/LF method in Section 5 and finally we conclude our work in Section 6.

2 Related Work

In the WMT 2023 Shared Task on Machine Translation with Terminologies, Semenov et al. (2023) emphasize the effectiveness of incorporating terminology dictionaries and respecting domain-specific terminology constraints. The authors also distinguish terminology incorporation from general MT methods.

Post et al. (2019) explore the use of masking to properly isolate and translate specific named entities such as terminology spans. Their findings show that masking solves some problems, but relies entirely on the masks being properly aligned.

Ghazvininejad et al. (2023) propose a method for translating rare words such as technical terminology. The method, called DiPMT, is a prompting technique that provides an LLM with multiple translation choices from a dictionary as well as hints about their meaning for a subset of input words. It outperforms baselines for low resource and out-of-domain MT. The authors also extract bilingual dictionaries from the training data to assist in this process. Doing so allows for fine-grained control over the use of domain-specific terminology.

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.

3 Background and Motivation

In order to better understand how ineffective acronym disambiguation may be for translators, we investigate the performance of LFs and their SF acronyms within the realm of commercial MT systems. We perform this necessary step in order to confirm our hypothesis that: **acronym disambiguation in the current state-of-the-art French MT systems is not being addressed properly**. In Table 1, we provide a specific agreement comparison that uses a widely-used commercial MT system – Google translate[3]. For both cases (LFs and SFs) agreement is between 54% and 63%, giving way to a high amount of room for improvement. We illustrate

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

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

Table 1: Google Translate agreement for long- and short-form acronyms.

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

Table 2: Erroneous Google Translate examples on long forms (LFs).

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

Table 3: Erroneous Google Translate examples on short forms (SFs).

this with further analysis in Tables 2 (long forms) and 3 (short forms).

As a way of mitigating the room for improvement, we propose the following novel method for MT that decomposes translation into four high-level steps by taking into account that Google Translate is more successful on LFs than SFs. For other MT systems, this may not be the case; we focus solely on Google Translate here as the oracle for our experiment.

1. Use Google Translate to translate each LF from French (FR) to English (EN).
2. Extract the LF from Google Translate's EN pair output (using a simple split command).
3. Generate several SF hypotheses using the extracted LF from Step 2.
4. Use a search technique to verify and evaluate certainty of hypotheses.

To better describe Steps 1 through 4, we provide the following in-depth description. A term such as "acide désoxyribonucléique **(adn)**" would first be translated in Step 1 from French to English as "deoxyribonucleic acid **(dna)**". We then extract the English LF (deoxyribonucleic acid) and SF (dna) for use in the next steps. Step 3 consists of the use of AB3P[4] (Sohn et al., 2008; Church and Liu, 2021), an acronym tool that provides LFs in English created by the United States government and contains acronyms from crawls of PubMed[5] and arXiv[6]. If a sufficient number of documents is not found that contain the English LF and SF together, we then generate a list of acronym hypotheses translations from the translated LF. Each hypothesis is generated using a fine-tuned version of the Scibert (Beltagy et al., 2019) model described in section 4.

Step 4 consists of the verification process, also known as "Fact Checking". Typically, the translation process for technical terms involves a significant component of researching the meaning of

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

[5] https://pubmed.ncbi.nlm.nih.gov/

[6] https://arxiv.org

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.[7] According to Bowker (2021), professional translation term 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. Domain expertise from professional translation trade unions such as the ATA[8] point to two or three sources being sufficient to substantiate use of a given term. We replicate that process using the search method below.

We implement a Boolean retrieval system that contains 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).

The translation of acronyms is further complicated by non-English languages opting to adopt a better known English acronym alongside a translation of the term. The French translation for "large language model" (grand modèle de langue) is condensed using the English acronym "LLM," even though the acronym does not correspond to the first letters of each word. Despite this limitation, our search step allows for the verification of such cases, as the pairing of term and acronym is likely to occur in the literature if they have found consensus in the field. Thus, verification would succeed and the disambiguation step would not be performed. Furthermore, fine tuning on corpora such as Pubmed was foregone due to the non-compositionality of many technical terms; boolean search ensures that the term is verified as a fixed unit.

While an exact match (e.g. 'RCP' to 'CPR') is the objective of our system, it is important to note that for evaluating the system we distinguish between *agreement* (an exact match) and *verification* (verified by a search) as noted:

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).

4 Experimental Settings

4.1 Translation Models

For **Google Translate**, experiments were performed using the Google API[9] as available to the public on October 14, 2023. For **Opus MT**, the vanilla model was used without any fine tuning. The French-to-English language variant from Hugging Face[10] was downloaded for this purpose.

4.2 Baselines

We compare the inclusion of our method against several baselines that are executed with and without our proposed step. Our experiments are performed on the acronym corpus that we created and allow for public consumption. Our first set of experiments focuses on three main baseline approaches found in Table 5 that we call: (1) *Identity*, (2) *Reverse*, and (3) *Google/Opus*[11]. The Identity baseline is the most straightforward experiment which is when the English SF output is **equal to** the French SF input (e.g. ADN in French is equal to ADN in English). The Reverse baseline is when the English SF output is the **reverse** of the French SF input (e.g. ADN in French is equal to NDA in English). The Google/Opus baseline takes the LF and SF in French and outputs an SF in English.

4.3 Hypothesis Generation

For the disambiguation of acronyms, we use a SciBERT (Beltagy et al., 2019) model that is fine-tuned on 1.8M term-acronym pairs in the target language (English) with these parameters: Adam as the optimizer, an initial learning rate of 2e-5, 1,000 warmup

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

[8] https://www.atanet.org/growing-your-career/terminology-management-what-you-should-know/

[9] https://cloud.google.com/translate

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

[11] We use the OpusMT system for an extra comparison https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

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

Table 4: Training data for SF candidate generation.

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

Table 5: Examples of our three baseline methods.

steps, and a weight decay of 0.01. We use data downloaded from arXiv[12] and then processed by AB3P for fine-tuning as shown in Table 4. The final model accepts input in the form: "LF ([MASK])" and outputs an n-best list of SF candidates.

4.4 Acronym Corpus

A new test set[13] (called the **acronym corpus** here) 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[14], a repository of French academic papers, many of which are from medicine and science. The pairings contain an LF and SF for each term in both French (source) and English (target). Examples were selected such that no offensive content or personal information was to be included.

The HAL 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. An example of the acronym translation task is to input a French LF such as **réanimation cardiopulmonaire** and its corresponding SF, in this case **RCP**. The output should be the correct translation of the SF: **CPR**.

5 Results

We compare the baselines first in Table 6. We provide both agreement and verification for consistency purposes, which show that verification is generally much lower than agreement for all systems.

When compared, our proposed technique, which includes search and verification, achieves 9.9% improvement (43.9%) for agreement and 17.8% improvement (32.7%) for verification compared to the baseline when using the OpusMT system. Google translate scores are also markedly higher, with 8.3% improvement (**62.6%**) and 13.6% (**42.8%**), respectively. It is clear that through the use of our proposed system, the acronym resolution is much higher for both agreement and verification.

Additionally, we illustrate the comparisons in more detail from a precision and recall perspective in Table 7 for all experimental systems. Our experiments show that through the use of our proposed step which uses agreement and verification, professional translators that use the Annotated Corpus will have more success using our system. Precision is presented here as the portion of agreed terms that are

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

[13] https://github.com/rtotheich/acronym_corpus/tree/main

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

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%

Table 6: Agreement and verification for the baseline experiments on the *Acronym Corpus*.

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 7: Precision and recall comparisons for all experimental systems.

verified and recall as the portion of verified terms.

6 Conclusion

Professional translators must be well versed in the source and target languages that they are translating. Translating technical terminology can be so important that it has been compared to the job of a terminologist (Cabré, 2010). Quality translations will take into account several units of measurement such as fluency, adequacy, and more. However, it has been the case in the past that, more often than not, terminology, specifically the translation of acronyms, is not included as a major improvement to a translator's pipeline. Domain-specific standards (GHENŢULESCU, 2015), nonetheless, have been set such that verification of terminology like acronyms is considered an important step in translation.

Translators and AI practitioners could benefit highly from the use of a system like the one presented in this article. We believe that our corpus and findings provide sufficient evidence and materials to reproduce a benefit to warrant future work on the topic.

7 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. The working languages of the authors being French and English, hand curating a corpus was limited to these only. Our solution also may not scale to longer texts; the method is based on working with term-acronym pairs and working on a full text would require a pre-processing 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.

8 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

Anastasopoulos, A., Besacier, L., Cross, J., Gallé, M., Koehn, P., Nikoulina, V., et al. (2021). On the evaluation of machine translation for terminology consistency. *arXiv preprint arXiv:2106.11891*.

Beltagy, I., Lo, K., and Cohan, A. (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.

Bosca, A., Nikoulina, V., and Dymetman, M. (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.

Bowker, L. (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.

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

Church, K. and Liu, B. (2021). Acronyms and opportunities for improving deep nets. *Frontiers in Artificial Intelligence*, 4:732381.

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

Ghazvininejad, M., Gonen, H., and Zettlemoyer, L. (2023). Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

GHENŢULESCU, L. R. (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.

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

Hasler, E., De Gispert, A., Iglesias, G., and Byrne, B. (2018). Neural machine translation decoding with terminology constraints. *arXiv preprint arXiv:1805.03750*.

Jon, J., Novák, M., Aires, J. P., Variš, D., and Bojar, O. (2021). Cuni systems for wmt21: Terminology translation shared task. *arXiv preprint arXiv:2109.09350*.

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

Lee, H. A. and Kim, G. C. (2002). Translation selection through source word sense disambiguation and target word selection. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Molchanov, A., Kovalenko, V., and Bykov, F. (2021). Promt systems for wmt21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Post, M., Ding, S., Martindale, M., and Wu, W. (2019). An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Semenov, K., Zouhar, V., Kocmi, T., Zhang, D., Zhou, W., and Jiang, Y. E. (2023). Findings of the wmt 2023 shared task on machine translation with terminologies. In *Proceedings of the Eight Conference on Machine Translation (WMT)*. Association for Computational Linguistics.

Skadiņš, R., Pinnis, M., Gornostay, T., and Vasiļjevs, A. (2013). Application of online terminology services in statistical machine translation. In *Proceedings of Machine Translation Summit XIV: Posters*.

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

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