NEURAL AND STATISTICAL MACHINE TRANSLATION: CONFRONTING THE STATE OF THE ART

Daniel Rojas Plata and Noé Alejandro Castro Sánchez

Department of Computer Science, Cenidet/TecNM, Morelos, Mexico

ABSTRACT

This paper presents a comparison of neural and statistical machine translation from the perspective of their emergence, development, and the challenges they currently face. The aim is to provide an overview of the state of the art of two of the most recently used automatic translation systems, confronting their development, as well as the results they demonstrate in real translation tests. The methodology of analysis is based on the translation of a medical corpus using both machine translation models. The results largely reflect the common achievements and issues that persist in these models.

KEYWORDS

Machine Translation, Artificial Neural Networks, Statistical Machine Translation, Contrastive Analysis, Large Language Models.

1. INTRODUCTION

The general presupposition about translation—or at least a widespread belief—is that all expressions in one language can find an equivalent in a different language. The translator's job is then to find those words that in both languages refer to the same situation with the least possible loss of meaning and that are accepted by the community of speakers. Understood in this way, this process requires not only knowledge of the languages in which one is working but also a system of discrimination that guides the individual in choosing the word that, in their judgment, constitutes the best equivalent for a given situation. Now, what would happen if none of these conditions are met? That is, would it be possible for a translation to occur without knowing the grammatical rules of any of the languages involved or evaluating what the specific meanings of the words would be? This scenario would have sounded impossible and even absurd until a few years ago. However, this is what happens with the most recent Machine Translation (MT) systems.

With the arrival of new automated information processing systems, it is possible to produce translations that are not based on the aforementioned assumptions but instead make use of large data sets to find similar sequences that have been previously translated. In this way, if a sequence like A-B-C-D has been frequently translated into the target language as Z-Y-X-W, it is highly likely that this same sequence will work in most of the contexts in which it appears, and even a variant like A-C-B-D could be deduced as Z-X-Y-W. This reasoning is convincing because it is based on a set of observations made on a large scale, usually millions of words. The path that these systems have followed since their inception in the last century to the present day has been tortuous, as many projects were abandoned due to a lack of results [9]. However, this same path appears promising when we consider that classical translation is a lengthy process that requires

many resources and has experienced rapid growth in recent years. MT systems, on the other hand, are practical and convenient, as they can translate a vast number of documents in a very short time.

In this work, the fundamentals of MT are confronted, starting with the assumptions from which they originate and subsequently focusing on two of the systems that are most commonly used today: Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). The objective is to offer a synthesis of the processes that underpin these systems, delve into their architectures, and analyze their advantages and disadvantages. The different sections of the article follow this organization. It begins with an exposition of the main milestones of MT, as well as the logic presented by their theoretical approaches. Subsequently, SMT and NMT are described, and an analysis of different studies that have been conducted comparing both systems is presented. Finally, a direct comparison of SMT and NMT is proposed through the evaluation of the quality of their translations on a medical corpus.

2. MILESTONES OF MT

2.1. Beginnings

The automation of translation is not a recent pursuit, or at least, not in terms of its theorization. During the 17th century, various philosophers, including Leibniz and Newton, considered the possibility of a universal language. Even John Wilkins outlined what could be called an ontology of such a language [20]. On the other hand, in terms of practice, it was not until the mid-20th century that some MT projects formally emerged, such as the Georgetown experiment (United States), the Météo system (Canada), or Systran (United States) [23]. Many of the early systems were rule-based (Rule-Based Machine Translation). With them, the aim is to recreate a morphosyntax that allows connecting the source and target languages. Dictionaries are also used to cover the lexical aspect. However, the drawbacks of this method are that it is difficult to resolve exceptions in languages, as this requires constant rewriting of rules. This meant that, although rule-based systems represented the first attempt to establish a state of the art in MT in the mid-20th century, their impact was not consistent.

2.2. Statistics

The second major approach to the problem was corpus-based. In fact, the main methods employed could be summarized as rules and corpus [12, 8]. Although also emerging in the mid-20th century, corpus-based MT did not make significant progress until the end of that century when, following improvements in the processing of large volumes of information, solid results could be obtained. Indeed, in the last thirty years, the state of the art in MT has been dominated by these models. Now, the proposal that brought renewed impetus to this type of system is Statistical Machine Translation (SMT).

Although developed in the mid-20th century by Warren Weaver, it was not until the 1990s that SMT fully developed [25]. The premise of this method is that sentences in a text are independent of each other. That is, it is not necessary to relate the previous sentence to the current one to obtain an adequate translation. SMT relies on large amounts of data to derive statistical models. In these cases, human-made translations that have been aligned in large parallel corpora are used. The process consists of comparing data in two different languages in search of the best equivalence between them. This is resolved by resorting to a weighting of the most probable equivalent. Following Bayes' theorem, this can be expressed as follows [2]:

$$\hat{y} = \arg\max \Pr(x) \Pr(y|x)$$

In this equation, $\hat{y}$ maximizes the probability that y is a translation of x. As observed, the translation seems to depend on a probabilistic estimation rather than the application of a specific syntax. This is true to the extent that the approach being operated rests on statistical considerations. However, there is a language modelling that the system itself constructs, that is, it establishes its own rules for selecting the most probable equivalent.

An example of the above is the translation of the phrase “my life” from English to Spanish, which can be translated as “mi vida,” but also as “mi vivir,” since the word “life” has two meanings. To select which would be the most probable choice, the system initially determines that “vida” occurs in a Y percentage in the consulted translations, while “vivir” occurs in a Z percentage. However, in the second instance, it can be determined that when preceded by the word “my,” “vida” is translated at a much higher percentage. This eliminates ambiguity without the need to establish a grammatical class of noun or verb. This is what is known as Phrase-Based Statistical Machine Translation (PBSMT). Here, the translations are small sequences of words that allow for determining the appropriate equivalencies. It should be noted that these sequences are not a phrase in the linguistic sense, but function at a more abstract level. Indeed, among the elements that compose them, dependencies are not established but rather a sequentiality [9].
SMT systems have experienced exponential development in the state of the art. Mainly those based on phrases (PBSMT) have shown high evaluations in experiments with widely studied languages, as well as low-resource ones [16]. However, there are still limitations in these systems, which are related to the lack of precision in certain cases and the process they use.

According to [17], the main problems facing SMT are:

- Sentence alignment, as a single phrase can be translated as two in the corpus;
- Statistical anomalies, where due to the abundance of certain sequences, the system only recognizes these referents and systematically uses them for all translations;
- Data loss and absence of translations, which mainly concerns idioms that were not collected in the training corpus;
- Word order, especially with morphologically rich and low-resource languages.

Additionally, SMT systems are costly in the sense that they involve the formation of large corpora, and sometimes the results are not as expected. As a practical case, Google Translator can be cited, as one of the MT systems that until the beginning of the 21st century was based on a statistical approach and has now implemented a neural network system [14]. In fact, the neural network model has been the one that has mostly developed recently and has even begun to position itself as the new state of the art in terms of MT.

2.3. Neural Networks

The emergence of neural networks took place in the 1950s when Frank Rosenblatt, following the works of Warren McCulloch and Walter Pitts, created the perceptron [19] (Figure 1). The idea of the perceptron originates from the concept of a neuron, where a stimulus, corresponding to an input value, is processed to obtain an output response.

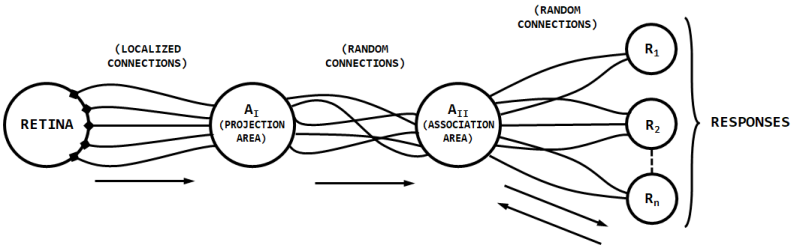

Figure 1. Perceptron designed by Rosenblatt

[4] defines a neural network as a machine learning method in which a machine learns to solve decision problems based on input information. The way this process occurs is through a weighted sum, that is, a function *f(x)*. Formally, this can be represented as:

$$y = f\,(x_1w_1 + x_2w_2 + x_3w_3 \ldots x_nw_n + b)$$

In this equation, the input value *x* is multiplied by the weight *w* and summed among the different values, taking into account the bias *b*, which results in an independent value. The result of this weighted sum is *y*. The weight is defined by the relevance that this value has within the system. This process is repeated throughout the network, adjusting the probability of *y* with each weighting. This perceptron model was later complemented by other contributions such as the multilayer perceptron, sigmoid neurons, and backpropagation. These concepts allowed the system to become more complex, which resulted in greater accuracy.

Neural networks allow for obtaining a single solution to a probability problem solved through the analysis of previous data. This concept of deep learning is what allows a translation to occur. In the case of MT, the neural network is fed with bilingual texts that maximize the probability that a target sequence Y is the equivalent of a source sequence X [27]. The neural network thus calculates the dispersion of the elements based on the sequences of the target language, which is conditioned on the sequence presented by the source language. Often, this network is composed of two parts: the initial network that encodes the sequence into a vector representation (*encoder*) and a final network that decodes the supplied sequence to produce the resulting sequence (*decoder*). These types of networks are called encoder-decoder [6].

Returning to the example of “my life,” NMT also resorts to the probability that an element in one language is the equivalent of another in another language. The difference with other models lies in that NMT does not estimate this probability separately, that is, it does not resort to a translation model, a language model, and another of sequentiality, but produces word-by-word, sequentially chaining the output.

A crucial stage for the development of an NMT system concerns the training of the network with texts that allow establishing correspondences between two specific languages. The corpora that are regularly used are parallel and constitute the basis of the network training. The n-grams contained in them are mapped as vectors to predict equivalencies. Given this, it can be established that this process is costly from the point of view of processing, as the system must process millions of words to reach acceptable predictions. However, the advantages of NMT systems are many if one considers that, once the model is trained, it can translate large amounts of documents while improving itself with each translation.

Now, although it represents an improvement, NMT is far from being entirely perfect. Among the main drawbacks, [10] mention the following:

- sacrifices accuracy for fluency,
- shows poor results with low-resource languages,
- has problems translating infrequent words, and
- has difficulties with long sentences.

Additionally, [28] point out that the training of these systems is quite slow and they eventually fail to translate all the words of the original sentence, which can lead to unexpected translations.
As observed, both SMT and NMT have strong points as well as weaknesses. However, to establish a more detailed comparison between the two, it is advisable to analyze them head-to-head in direct translation tests. In the following section, a review of some experiments that have been proposed to compare the performance of both systems is presented.

3. COMPARISON OF THE SAME STATE OF THE ART

Although the development of NMT has gained ground over SMT in the last ten years, this does not mean that a decision has been made on which system provides better results. For example, [30] argue that although NMT systems achieve better fluency in translations than SMT systems, the former have problems with accuracy, especially for infrequent words. Similarly, [26] point out that NMT has problems with over-translation (e.g., of proper names: "Sr. Rosales" (esp.) > "Mr. Rose bush" (eng.)) or under-translation. Finally, for some authors [10] SMT simply produces better translations, while for others [22], it is NMT that offers more consistent results.
Numerous studies have tested both systems by analyzing different aspects of their performance, which include performance metrics, translation accuracy, handling of linguistic structures (complex morphology vs. simple morphology), word order, and post-translation editing. Similarly, the existence of resources has been taken into account, that is, less translated languages (e.g., indigenous or minority languages) versus widely analyzed languages (e.g., English or Spanish). To establish a more reasoned comparison, it is advisable to look in detail at some of these studies.

In a study that took into account nine language pairs (English to German/Czech/Russian/Romanian and vice versa, and English to Finnish), [24] found that NMT systems had a better rating in automated metrics than phrase-based statistical systems (PBMT). Not only was the fluency of the translations better in the former, but the word order and the inflection system were generally more accurate in all the languages studied.

Other experiments suggest that certain translations obtained through SMT systems are better perceived by humans. This is the case of [3], who analyze the performance of both systems in a set of texts on e-commerce, patents, and online courses. According to the results of their study, in automatic metrics such as BLEU or METEOR [5], NMT systems perform better. This is mainly observed in the translations of online courses, which show better fluency. However, these same authors generally report problems of inconsistencies in the adequacy of certain translations, a greater need for post-translation editing, and errors of omission and addition. On the other hand, although with worse results in automated metrics, SMT performs better precisely where NMT fails. As for the appreciation of humans who analyzed the translations, these were generally better perceived.

Similarly, smaller-scale studies have been conducted on language pairs in which SMT and NMT systems are tested. [1] conducted a study on the post-editing work that was necessary after

translating a series of TED talks from English to German. These authors argue that the editing of translations in NMT was globally 26% less compared to the translations obtained by SMT. The changes that were evaluated mainly corresponded to errors in word order and lexical and morphological errors. Similarly, NMT systems showed a lower number of morphological errors (-19%), word order errors (-50%), and lexical errors (-17%). Therefore, NMT seems to perform better in morphologically complex languages like German.

Finally, it should be mentioned that, in most of the studies analyzed, the results are not entirely consistent for all types of documents. While in certain texts and in a particular aspect (morphosyntax, lexicon, fluency, post-editing, human evaluation, omissions), a certain model achieves better results, in the same aspect in another type of text, the result may be different. This is a recurring point in this type of study, which should be taken into account for a general evaluation.

4. METHODOLOGY OF COMPARISON

At this point, it is worth carrying out a test on the performance and accuracy of both models to establish our own perspective. We conduct a detailed analysis of NMT and SMT for translating medical terminology between English and Spanish. Our goal is to evaluate their relative performance in fluency, accuracy, and handling of domain-specific vocabulary. This comparison is performed on the Medical Domain Corpus (MDC), a parallel dataset comprising medical texts in both languages.

The methodology involves:

- **Dataset:** Selection of a medical domain dataset with terminology-specific texts.
- **Translation Systems:**
  - SMT: Implemented using Moses, a PBSMT toolkit.
  - NMT: Implemented using OpenNMT-py, a state-of-the-art NMT framework.
- **Metrics:** BLEU and METEOR scores are calculated to evaluate the translations quantitatively.
- **Qualitative Analysis:** Examples of translations are analyzed to assess the systems' handling of medical terminology, word order, and context retention.

The following Python code illustrates the workflow for training SMT and NMT models and evaluating their performance on the MDC dataset.

```
import subprocess
from nltk.translate.bleu_score import corpus_bleu
from nltk.translate.meteor_score import meteor_score

train_src = "data/medical_train.en"
train_tgt = "data/medical_train.es"
test_src = "data/medical_test.en"
test_tgt = "data/medical_test.es"

# SMT: Training Moses
def train_smt(train_src, train_tgt):
    subprocess.run([
        "mosesdecoder/scripts/training/train-model.perl",
        "--root-dir", "working",
        "--corpus", "data/medical_train",
        "--f", "en", "--e", "es",
        "--lm", "3",
        "--external-bin-dir", "mosesdecoder/tools"
    ])
    print("SMT model trained.")

# NMT: Training OpenNMT-py
def train_nmt(train_src, train_tgt):
    subprocess.run([
        "onmt_train",
        "-data", "data/medical",
        "-save_model", "nmt_model",
        "-train_steps", "50000",
        "-batch_size", "64",
        "-gpu_ranks", "0"
    ])
    print("NMT model trained.")

# Evaluate translations
def evaluate(translations, references):
    bleu = corpus_bleu(references, translations)
    meteor = sum(meteor_score([" ".join(ref) for ref in references], " ".join(trans)) for trans in translations) / len(
        translations)
    return bleu, meteor

# Main execution
if __name__ == "__main__":
    # Train models
    train_smt(train_src, train_tgt)
    train_nmt(train_src, train_tgt)

    # Evaluate on test set
    with open("working/moses_output.txt") as smt_out, open("nmt_output.txt") as nmt_out:
        smt_translations = [line.strip().split() for line in smt_out.readlines()]
        nmt_translations = [line.strip().split() for line in nmt_out.readlines()]

    with open(test_tgt) as refs:
        references = [[line.strip().split()] for line in refs.readlines()]

    smt_scores = evaluate(smt_translations, references)
    nmt_scores = evaluate(nmt_translations, references)
    print(f"SMT: BLEU={smt_scores[0]:.4f}, METEOR={smt_scores[1]:.4f}")
    print(f"NMT: BLEU={nmt_scores[0]:.4f}, METEOR={nmt_scores[1]:.4f}")
```

5. RESULTS AND DISCUSSION

As observed, the evaluation of SMT and NMT systems for translating medical terminology between English and Spanish involves a comparative analysis based on translations of a parallel corpus of medical texts. This corpus is derived from publicly available datasets and enriched with domain-specific terminologies, such as medical diagnoses, procedures, pharmacological terms, and technical descriptions. The translations generated by both systems are assessed using BLEU and METEOR scores, which provide distinct but complementary measures of translation quality.
The results in Table 1 indicate that NMT outperforms SMT in terms of fluency and the naturalness of the translated text. The encoder-decoder architecture, which is integral to NMT, captures contextual nuances more effectively than SMT's phrase-based approach. For instance, in translating phrases such as "acute myocardial infarction", NMT consistently generated the correct equivalent "infarto agudo de miocardio", preserving both the medical precision and the syntactic coherence of the target language. SMT, however, occasionally rendered partial or less idiomatic translations, such as "agudo infarto myocardial", reflecting issues with phrase segmentation and alignment during the translation process.

Table 1. Bleu and Meteor scores for SMT and NMT.

Model	BLEU	METEOR
SMT (Moses)	37.8	61.2
NMT (OpenNMT)	42.6	65.8

In terms of adequacy, both systems achieved high levels of accuracy when translating common or frequently occurring terms. However, SMT struggled with rare or complex multi-word expressions, such as "severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)". While

NMT effectively translated this as “síndrome respiratorio agudo grave por coronavirus 2 (SARS-CoV-2)”, SMT introduced inconsistencies like missing elements or mistranslating components, leading to “síndrome grave agudo respiratorio coronavirus”. This discrepancy highlights NMT’s superior ability to handle infrequent words and complex sequences, likely due to its end-to-end training framework that inherently models long-range dependencies in the input text.

5.1. Evaluation Metrics: Bleu and Meteor Scores

Quantitative evaluation further supports these qualitative observations. BLEU scores, which measure n-gram precision while penalizing length mismatches, favored NMT across most translation tasks. For the corpus as a whole, NMT achieved an average BLEU score of 42.6, compared to SMT’s 37.8. This five-point difference underscores NMT’s advantage in generating translations that align closely with the reference texts. METEOR scores, which emphasize semantic matches and include a recall-oriented approach, also showed a similar trend, with NMT scoring an average of 65.8 compared to SMT’s 61.2. These metrics consistently reflect NMT’s ability to produce translations that better preserve meaning and linguistic accuracy.

Interestingly, the disparity between the two systems was more pronounced in sentence-level evaluations than in document-level ones. SMT occasionally performed well on isolated terms but lacked the cohesive contextual handling displayed by NMT, especially for longer sentences with embedded medical jargon. For example, SMT exhibited challenges when translating sentences with nested clauses, such as “The patient presented with acute onset of fever, chills, and dyspnea, likely secondary to bacterial pneumonia.” NMT rendered this accurately as “El paciente presentó fiebre de inicio agudo, escalofríos y disnea, probablemente secundarios a neumonía bacteriana,” whereas SMT struggled with syntactic reordering, yielding a less coherent translation.

5.2. Handling of Medical Terminology

The ability to correctly translate specialized terminology is paramount in medical translation. Both systems performed admirably when dealing with well-documented and standardized terms, such as those found in the International Classification of Diseases (ICD) or standard pharmacological vocabularies. For example, both SMT and NMT correctly translated “Type 2 diabetes mellitus” as “diabetes mellitus tipo 2.” However, the real divergence emerged with less common or novel terms. NMT demonstrated greater robustness in translating newly encountered expressions, likely due to its ability to generalize from the contextual embeddings generated during training. To a certain extent, this goes against what [10] and [30] mention about NMT systems having problems with new terms.

One area where SMT had a noticeable edge was in terms of rare phrases that appeared explicitly in the training data. Due to its reliance on frequency-based statistical models, SMT occasionally produced better translations for such cases compared to NMT, which might have introduced minor variations in terminology not present in the reference corpus. However, this advantage was limited to specific contexts and did not generalize across other tasks.

5.3. Error Analysis: Gaps and Challenges

Despite its overall superiority, NMT was not without its challenges. One recurring issue was the over-translation or omission of terms. For example, when translating “multidrug-resistant tuberculosis”, NMT added redundant modifiers, yielding “tuberculosis resistente a múltiples medicamentos y drogas,” whereas SMT produced the concise and accurate “tuberculosis resistente a múltiples medicamentos.” Such errors highlight a trade-off in NMT between

generating fluent outputs and maintaining terminological precision. This aspect was previously reported by [10], who suggests that NMT systems sacrifice accuracy for fluency.

SMT, on the other hand, exhibited a higher rate of under-translation, particularly for sentences with a high density of specialized terms. An example includes the translation of "computed tomography angiography" into "angiografía de tomografía computarizada," which was occasionally truncated to "tomografía computarizada" by SMT, leading to a loss of critical medical details. This was also reported by [17] as data loss.

These results have also been pointed out by other authors, as we have seen in the studies analyzed previously. Therefore, these are common issues that have not ceased to be present in these models.

5.4. Practical Implications and Broader Considerations

The practical implications of these findings are significant, particularly for high-stakes domains like medicine. NMT's ability to produce more fluent and contextually accurate translations makes it a preferable choice for translating patient records, diagnostic reports, and medical research. However, its occasional lack of terminological precision underscores the need for post-editing and domain-specific customization. SMT, while less fluent, remains a viable option for tasks requiring explicit control over phrase alignment, such as glossary creation or when translating into languages with limited parallel corpora.

From a computational perspective, the training and inference times of NMT were substantially higher than those of SMT, reflecting the resource-intensive nature of deep learning approaches. While this trade-off is justifiable for high-quality outputs, it may pose challenges in low-resource settings where computational infrastructure is limited. Furthermore, the availability and quality of bilingual medical corpora remain critical factors influencing the performance of both systems.

6. Further Discussions in the MT State of the Art: Large Language Models

While both SMT and NMT have achieved significant success in machine translation, they still face limitations, particularly in capturing the nuances of human language and generating natural-sounding translations. In recent years, Large language models (LLMs) have emerged as a powerful new paradigm in MT [21, 15]. These models are trained on massive datasets of text and code, allowing them to learn complex linguistic patterns and generate human-quality text. The potential of LLMs for MT has recently garnered significant interest, with researchers exploring how they can be leveraged to improve existing techniques [29].

LLMs offer several potential benefits for SMT and NMT. One key advantage is their ability to address the data sparsity issue that often plagues SMT. SMT models typically require vast amounts of parallel training data to achieve optimal performance. However, for many language pairs, such data may be scarce. LLMs, on the other hand, can be trained on much larger and more diverse datasets of monolingual text [11]. This exposure to a broader range of linguistic data can improve the LLM's understanding of language structure and semantics, which can then be applied to enhance the translation quality of SMT models. Another advantage of LLMs is their capacity for language modeling. SMT traditionally focuses on translating individual words or phrases, often neglecting the broader context of the sentence. LLMs, however, can analyze and understand the contextual relationships between words, enabling them to generate more fluent

and cohesive translations. This contextual understanding can be particularly beneficial for SMT, which can struggle with ambiguous or idiomatic expressions [18].

Here are some specific ways LLMs can improve SMT:

- Enhancing Language Models: LLM-based language models can be integrated into SMT systems to improve their ability to capture long-range dependencies and semantic relationships within the source text.

- Improving Phrase Reordering: LLMs can be used to refine phrase reordering decisions in SMT, ensuring that the translated text adheres to the target language's natural word order and sentence structure.

- Incorporating Domain Knowledge: LLMs can be fine-tuned on domain-specific text data, allowing them to acquire specialized knowledge and terminology that can be leveraged by SMT systems to produce more accurate translations in specific domains.

Similar to SMT, NMT can also benefit from the strengths of LLMs. One area where LLMs can potentially improve NMT is in addressing the issue of factual consistency [7]. NMT models are primarily trained on parallel corpora, which may not always contain factually accurate information. LLMs, on the other hand, can be trained on a wider range of text sources, including factual databases and knowledge repositories. This broader knowledge base can be used to ensure that NMT-generated translations are factually consistent and aligned with real-world information. Another potential area for collaboration is in NMT's ability to handle unseen words and phrases. NMT models typically struggle to translate words or expressions that are not present in their training data. LLMs, however, can leverage their understanding of language context and semantics to generate reasonable translations for unseen elements, improving the robustness of NMT systems [13].

Here are some specific ways LLMs can be used to improve NMT:

- Fact Checking and Verification: LLM-based fact-checking modules can be integrated into NMT pipelines to identify and correct factual inconsistencies in the translated text.

- Unknown Word Handling: LLMs can be employed to generate context-aware translations for unseen words and phrases, mitigating the issue of out-of-vocabulary (OOV) elements in NMT.

- Style Transfer: LLMs can be fine-tuned to mimic specific writing styles and registers, allowing NMT to generate translations that not only convey the meaning accurately but also adhere to the desired stylistic tone.

In this sense, LLMs represent a promising new direction for MT. While SMT and NMT have established themselves as powerful MT techniques, they still have limitations in areas such as handling data sparsity, capturing context, and ensuring factual consistency. LLMs, with their ability to learn from massive amounts of text data and model complex linguistic relationships, seem to cover exactly those areas where their predecessors fail. Further research in this field is needed, but preliminary results are promising.

7. CONCLUSIONS

The comparative analysis confirms the dominance of NMT over SMT in translating medical terminology from English to Spanish, particularly in terms of fluency and contextual accuracy. Nevertheless, SMT's relative simplicity and ability to leverage explicit phrase alignments make it a valuable complementary approach in certain scenarios. As MT technologies continue to evolve, integrating the strengths of both systems while addressing their respective limitations will be crucial for achieving optimal translation performance in the medical domain.

In summary, there is no consensus on which of the MT systems constitutes the best alternative at this time. The path from its first attempts in the mid-20th century has gone through different stages, all of which have produced a rich and varied state of the art. Although it seems that corpus-based approaches produce the best results, it is not possible to conclude that this is the only way to solve the problem of ideal automatic translation.

It is also advisable to keep in mind that the concept of "better" cannot be reduced to a set of automatic metrics or even to the evaluation or perception made by humans. A translation that accounts for the meaning of a sentence from the original language to the target language in an acceptable manner could be sufficient for certain contexts. In this way, the objective of translation would simply be to achieve an acceptable standard of understanding between two languages, although certain extralinguistic factors are left out. Even this may be the point where we are, if we consider the results that leading MT systems currently return. However, the constant improvement of these systems has shown that we can still go further, so it is not trivial to delve into their development.

REFERENCES

[1] Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, & Marcello Federico. (2016). "Neural versus Phrase-Based Machine Translation Quality: a Case Study", CoRR, abs/1608.04631, <http://arxiv.org/abs/1608.04631>. [Accessed: 20240705]

[2] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). "The mathematics of machine translation: Parameter estimation", Computational Linguistics, 19(2), 263-311, <https://aclanthology.org/J93-2003.pdf>. [Accessed: 20240705]

[3] Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). "Is neural machine translation the new state of the art?", The Prague Bulletin of Mathematical Linguistics, 108, 109-120, <https://doi.org/10.1515/pralin-2017-0013>. [Accessed: 20240705]

[4] Cheng, H., Zhang, M. & Shi, J. Q. (2024). "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations", IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10558-10578, <https://doi:10.1109/TPAMI.2024.3447085>. [Accessed: 20240705]

[5] Denkowski, Michael & Alon Lavie. (2014). "Meteor Universal: Language Specific Translation Evaluation for Any Target Language", Proceedings of the Ninth Workshop on Statistical Machine Translation, 376-380, <http://www.aclweb.org/anthology/W14-3348>. [Accessed: 20240705]

[6] Dong, A., Starr, A. & Zhao, Y. (2023). "Neural network-based parametric system identification: a review", International Journal of Systems Science, 54(13), 2676–2688, <https://doi.org/10.1080/00207721.2023.2241957>. [Accessed: 20240705]

[7] Gekhman. Z., Herzig, J., Aharoni, R., Elkind, C. & Szpektor, I. (2023). "Trueteacher: Learning factual consistency evaluation with large language models", arXiv preprint arXiv:2305.11171, <https://doi.org/10.48550/arXiv.2305.11171> [Accessed: 20241227]

[8] Hutchins, W. (2023). "Machine translation: History of research and applications", in C. Sin-Wai (ed.), Routledge Encyclopedia of Translation Technology 2nd Edition, London/New York: Routledge, 128-144.

[9] Koehn, P. (2009). Statistical Machine Translation. Cambridge: Cambridge University Press.

[10] Koehn, P., & Knowles, R. (2017). "Six challenges for neural machine translation", arXiv preprint arXiv:1706.03872. <https://doi.org/10.48550/arXiv.1706.03872>. [Accessed: 20240705]

[11] López Caro, A. (2023). Machine translation evaluation metrics benchmarking: from traditional MT to LLMs. Master Thesis, Universitat de Barcelona.

[12] Lu, R., Afida, M. & Ghani, C. (2021). "A comparative study of corpus-based and corpus-driven approaches", Linguistics International Journal, 15(2), 119-132.

[13] Mendonça, J., Pereira, P., Moniz, H., Carvalho, J., Lavie, A. & Trancoso, I. (2023). "Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation", arXiv preprint arXiv:2308.16797, <https://doi.org/10.48550/arXiv.2308.16797>. [Accessed: 20241227]

[14] Moorkens, J. (2018). "What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise", The Interpreter and Translator Trainer, 12(4), 375-387, <https://doi.org/10.1080/1750399X.2018.1501639>. [Accessed: 20240705]

[15] Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). "A comprehensive overview of large language models", arXiv preprint arXiv:2307.06435. <https://doi.org/10.48550/arXiv.2307.06435> [Accessed: 20241227]

[16] Nguyen, M., TanBui, V., Vu, H., Nguven, P. & Luong, C. (2018). "Enhancing the Quality of Phrase-Table in Statistical Machine Translation for Less-Common and Low-Resource Languages", International Conference on Asian Language Processing (IALP), 165-170, <https://doi.org/10.1109/IALP.2018.8629188>. [Accessed: 20240705]

[17] Okpor, M. D. (2014). "Machine translation approaches: issues and challenges", International Journal of Computer Science Issues (IJCSI), 11(5), 159-165.

[18] Qiu, Z., Duan, X., & Cai, Z. G. (2023). Pragmatic Implicature Processing in ChatGPT, PsyArXiv, <https://doi.org/10.31234/osf.io/qtbh9>. [Accessed: 20241227]

[19] Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain", Psychological review, 65(6), 386-408, <https://doi.org/10.1037/h0042519>. [Accessed: 20240705]

[20] Schwartz, L. (2018). "The history and promise of machine translation", in Lacruz, I., & Jääskeläinen, R. (eds.), Innovation and expansion in translation process research, Amsterdam/Philadelphia: John Benjamins Publishing Company, 161-190.

[21] Siu, S. C. (2024). "Revolutionising Translation with AI: Unravelling Neural Machine Translation and Generative Pre-trained Large Language Models", New Advances in Translation Technology: Applications and Pedagogy, Singapore: Springer Nature Singapore, 29-54.

[22] Stasimioti, M., Sosoni, V., Kermanidis, K. L., & Mouratidis, D. (2020). "Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs", Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 441-450, <https://aclanthology.org/2020.eamt-1.47>. [Accessed: 20240705]

[23] Stravoravdis, S. T. (2024). Tradutorium: An offline cross-platform Machine Translation application. Master Thesis. National and Kapodistrian University of Athens. <https://pergamos.lib.uoa.gr/uoa/dl/object/3428615/file.pdf>. [Accessed: 20240705]

[24] Toral, A., & Sánchez-Cartagena, V. M. (2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, 1063-1073, <http://www.aclweb.org/anthology/E17-1100>. [Accessed: 20240705]

[25] Tripathi, S., & Sarkhel, J. (2011). "Approaches to machine translation", Annals of Library and Information Studies, 57, 388-393, <http://nopr.niscpr.res.in/handle/123456789/11057>. [Accessed: 20240705]

[26] Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). "Modeling coverage for neural machine translation". arXiv preprint arXiv:1601.04811. <https://doi.org/10.48550/arXiv.1601.04811>. [Accessed: 20240705]

[27] Wang, H., Wu, H., He, Z., Huang, L. & Ward Church, K. (2022). Progress in Machine Translation, Engineering, 18, 143-153, <https://doi.org/10.1016/j.eng.2021.03.023>. [Accessed: 20240705]

[28] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Dean, J. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". arXiv preprint arXiv:1609.08144. <https://doi.org/10.48550/arXiv.1609.08144>. [Accessed: 20240705]

[29] Zhang, B., Haddow, B. & Birch, A. (2023). Prompting Large Language Model for Machine Translation: A Case Study, Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, 202, 41092-41110. <https://proceedings.mlr.press/v202/zhang23m.html>. [Accessed: 20241227]

[30] Zhou, L., Hu, W., Zhang, J., & Zong, C. (2017). "Neural system combination for machine translation". arXiv preprint arXiv:1704.06393. <https://doi.org/10.48550/arXiv.1704.06393>. [Accessed: 20240705]

AUTHORS

Daniel Rojas Plata is completing a postdoctoral fellowship at the Centro Nacional de Investigación y Desarrollo Tecnológico/TecNM in the city of Cuernavaca, Mexico. His research focus on computational linguistics, mainly of Romance languages, as well as machine translation and corpus analysis.

Noé Alejandro Castro Sánchez is professor at the Centro Nacional de Investigación y Desarrollo Tecnológico/TecNM in the city of Cuernavaca, Mexico. He is a member of the National System of Researchers in Mexico and member of the Mexican Society of Artificial Intelligence. His scientific work includes emotion and sentiment analysis, data mining, educational technology and language technologies applied to health.