Grenze International Journal of Engineering and Technology, June Issue

Uncovering Systemic Vulnerabilities in NMT: A Linguistically Motivated Stress Test for English-German Translation

Priya Goel[1], Prasanna Dwivedi[2], Pankaj Goel[3], Mahek Singhal[4], Shruti Shukla[5] and Sakshi[6]
[1-6]GL Bajaj Institute of Technology and Management, Greater Noida, India
Email: priya.goel.beruf@gmail.com, prasanna.dwivedi@glbitm.ac.in, pankaj.goel@glbitm.ac.in
singhalmahek12@gmail.com, cse23123@glbitm.ac.in, p.sakshi2810@gmail.com

***Abstract*— Neural Machine Translation (NMT) has achieved human-like fluency in machine translations for generic domains in recent years. However, there are notable errors in the accuracy and progress when we discuss the need to process linguistic, deeply structured inputs in more than just English. This paper provides a comparative analysis of two popular NMT architectures, DeepL and Google Translate, for the English-German Language Pair in order to account for the observed errors. This study uses a manually curated dataset that includes extreme scenarios appropriate for target-based stress testing, rather than using a premade dataset from another source. Idiomatic expressions, structural stress, lexical density, and literary tone are the four main linguistic categories in which its performance is evaluated. The COMET metric is utilized to perform machine-based analysis upon the derived data. This analysis highlights that these metrics don't fail completely but display their own set of strengths and weaknesses. It shows that Google Translate handles intricate syntactic dependencies in a better way alongside the involved recursive structures found in many languages, whereas DeepL is better at understanding cultural nuances and practical meaning. Regarding the drawbacks, both systems struggle to accurately depict authorial voice in conjunction with highly context-dependent lexical compounds. The goal of this study is to systematically tabulate and categorise the errors in order to highlight the limitations of generalised evaluation, even though it only discusses a single language pair. As our contribution, we contend that the development of strong, reliable, and long-lasting multilingual AI systems across languages in the multilingual domain requires focused, linguistically motivated stress testing.**

***Index Terms*— COMET metric, German linguistics, machine translation evaluation, neural machine translation.**

I. INTRODUCTION

The globalisation of digital communication has made high-fidelity machine translation systems more and more crucial. This enhances both the transparency and the semantic accuracy that must be preserved across linguistic boundaries. Although somewhat successful, previous Rule-Based and Statistical Machine Translation (SMT) models heavily relied on phrase-to-phrase substitution. The introduction of Neural Machine Translation (NMT), which provided a more comprehensive understanding of the capabilities of NMT systems, further transformed this idea.

Grenze ID: 01.GIJET.12.2.33_20
© Grenze Scientific Society, 2026

NMT systems, particularly those based on Transformer architectures [4], have been shown to achieve near human fluency in the majority of general domains by modelling sentence-level context using Artificial Neural Networks (ANN).

However, even with surface-level fluency, it has been observed that these models often succeed in concealing systemic flaws in linguistic fidelity. Many evaluation metrics and machine analysis were previously reliant on BLEU (Bilingual Evaluation Understudy) prior to the development of contemporary mechanisms. However, recent work has shown that learned metrics outperform traditional n-gram based metrics like BLEU in correlating with human judgement in translation quality evaluation [7], [3]. This metric is heavily reliant on n-gram overlap. This is helpful when lexical matching is required, but the metric is ineffective when it comes to identifying and differentiating between errors resulting from complex situations, such as unfamiliar idioms, long-distance dependencies, and cultural emotions [7]. When we discuss topologically divergent languages, such as the German used in this study, this disadvantage is brought to light. German shares several similarities with English; however, it reflects several morpho-syntactic challenges that English sufficient NMTs often ignore due to the verb-second (V2) positioning, high rate of inflection and compound noun structures. This in turn discusses the shortcomings of single-language models that are clearly apparent in translation. In situations where work cannot be advanced further without translation, this becomes an essential requirement.

Targeted linguistic stress testing must replace generalized optimization as the evaluation methodology in order to achieve true human parity. For situations where generalization shows results on a lower scale, optimizing linguistic barriers as a tool for betterment might just lead to a much-needed breakthrough. This paper aims to address this gap by evaluating two leading state-of-the-art systems, DeepL and Google Translate, using a manually curated Linguistically Motivated Challenge Set (L-MCS). Unlike existing standard datasets, the L-MCS is designed to stress-test models across four identified failure categories by consuming the most extreme and complex of scenarios coupled with common phrases:

- Idiomatic and Non-Compositional Expressions: Assessing the retrieval of culturally bound meanings where literal translation often fails to capture the sarcasm and essence of expressions.
- Structural Stress: Targeting complex syntax, recursive clauses and the German verb-final placement.
- Lexical Density: Challenging the models with compound nouns alongside high context emotional concepts questioning the ability to transmit a similar emotional stance.
- Literary and Affective Tone: Evaluating the needed preservation of authorial voice and modal particles (e.g., doch, wohl) in expressions as witnessed in books and general scenarios.

Ultimately, our analysis employs the COMET metric, a neural framework with a strong correlation with human judgement, in addition to the additional qualitative error analysis. The primary contribution of this work is to demonstrate that NMT errors are not random stochastic failures but rather predictable systemic deficiencies caused by underlying linguistic rules. This looks at a carefully chosen analysis of the linguistic obstacles that multilingualism must overcome rather than just drawing comparisons.

II. Literature Review

A. The Paradigm Shift

The use of Neural Machine Translation (NMT) has resulted in a paradigm shift in the usage of machine translation. In order to encode non-fixed length sentences into non-fixed length vectors, early NMT systems typically employed the Encoder-Decoder architecture and Recurrent Neural Networks (RNNs). Nonetheless, condensing complex source sentences into fixed representations was a significant bottleneck and it was difficult to produce more or less accurate translations. In a solution to this issue, Bahdanau et al. [11] developed the Attention Mechanism, which enables models to dynamically align the words in the source and target language during decoding. This contributed to the noticeable transition between the static translations to the dynamic translations. The result of the development was the Transformer architecture that entirely did not use recursion in favour of self-attention mechanisms and was presented by Vaswani et al. [8]. These architectural innovations allow the contemporary systems explored in this paper DeepL and Google Translate to be massively parallelised and context-aware. This increased awareness in NMT-trained models makes further and more explicit linguistic analysis possible.

B. The Challenge Set Methodology

The NMT systems continue to experience issues in important situations even with the development of architecture. According to Koehn and Knowles [9], six challenges were found to be critical in NMT is superior to Statistical Machine Translation (SMT) because of the problems with long sentences and word alignment.

They discuss the way morphologically rich languages limit the existing processes of translation and observe that NMT systems tend to achieve fluency at the expense of semantic adequacy and generate an output that is a factually hallucinated but grammatically plausible output. With the increased importance of the literality of words, the semantics and essence are being forgotten. To identify these nuanced failures, we use the so-called "Challenge Set Approach" [10] that was constructed by Isabelle et al. Their findings indicate that the regular, randomly selected test sets are not likely to identify certain linguistic flaws. To detect such minor failure, we use the Isabelle et al. formulated method known as the Challenge Set Approach [10]. According to their results, typical test sets selected randomly fail to uncover certain linguistic flaws. In the case of the subtleties that are specific to a specific scenario, specific samples are required to be more precise. The task is to build manually a corpus of challenging sentences, which is supplemented with the scenario below to investigate the ability of the systems to break certain structural divergences:

- Structure Stress: Based on research showing that morpho-syntactic divergences, such as complex subject-verb agreement and non-local movement verbs, pose a significant challenge for NMT systems [10].
- Idioms: Addressing a well-known blind spot where models are known to regularly fall back on literal translations, misidentify idioms, and fail to identify a suitable language replacement, all of which prevent them from capturing the non-compositional meaning of multi-word expressions.

C. Domain-Specific Limitations and Homogenization

This paper discusses the homogenisation of NMT output outside the field of structural mechanics. NMT models, according to recent studies, always produce less lexical diversity than human translators and are highly prone to lexical simplification. This is the reason why we have our "Literary" and "Lexical Density" datasets. In their article on the way Stochastic Parrots, the Large Language Models (LLMs) and NMT systems, mimic predominantly hegemonic training data, prefer high-frequency modes to original or creative stylistic decisions, as seen in the analysis of Bender et al. [4]. This study attempts to ascertain whether DeepL and Google Translate can maintain stylistic nuance or if they revert to a standardised, simplified "machine voice" through testing on literary texts and high-density lexical inputs. This is an effort to discuss adopting more appropriate and stylistic. This is in attempt to further talk about adopting more apt and stylistic choices instead of giving sole preference to often revisited, frequent multilingual choices.

D. Measurement Validity

A crucial part of our approach is COMET over BLEU. There is a strong argument against using common n-gram metrics like BLEU for this particular assessment. After conducting a systematic review of metric validity, Reiter [7] came to the conclusion that BLEU scores frequently do not correlate with human judgement or real-world system utility. In the context of our research specifically regarding idioms and literary translation, a correct creative translation often shares little surface-level overlap with a reference, causing BLEU to erroneously penalize valid outputs. Adapting to the idiomatic interpretation comes naturally and is flagged for a model that prefers literal translations, which degrades translation quality.

We use COMET (Cross-lingual Optimised Metric for Evaluation of Translation) to address this. COMET, which was first presented by Rei et al. [5], evaluates semantic similarity instead of just lexical overlap using a neural framework with cross-lingual embeddings. Recent developments such as COMETKiwi extend this idea by enabling reference-free machine translation evaluation using pretrained multilingual models [2]. Cross-linguistic mapping becomes essential when we move beyond a single language in order to extract the context and subtleties of an expression even when the transmission language is inconsistent. The results of the WMT20 Metrics Shared Task, where Mathur et al. [6] showed that learnt metrics like COMET considerably outperform BLEU in correlating with human judgement, especially for high-quality systems. We make sure that our analysis captures the semantic accuracy and nuance that existing metrics miss. This is done by combining the structural rigor of the Challenge Set Methodology [10] with the sophisticated neural evaluation through the use of COMET.

III. METHODOLOGY

Instead of using a typical corpus evaluation, a "Challenge Set" method was used to assess DeepL and Google Translate's (GT) performance on an English-German language pair. This approach is based on research by Isabelle et al. [10], who contend that randomly chosen test sets frequently fail to identify the precise linguistic defects in Neural Machine Translation (NMT) systems. By focusing on "hard" linguistic examples, we stress-test

the models' ability to bridge structural and semantic divergences between the two languages in order to improve the learning accuracy for specific test tests and scenarios.

A. Corpus Construction

In this research, a Linguistically Motivated Challenge Corpus (L-MCS) was developed. This corpus consists of four dissimilar datasets, each of which is devoted to a specific linguistic phenomenon that is known to create problems to NMT systems as illustrated in Table I.

- Idiomatic Expressions: Focuses on non-compositional language whereby the meaning cannot be understood through the words that make it up. This determines the ability of the models to provide pragmatic meaning with a pre-eminence over lexical literalism.
- Structural Stress: Structural stress is concerned with morpho-syntactic divergences. It contains complex structures in the sentence that necessitate considerable rearrangement in the target language, including nesting structure, crossing movement verbs and long-distance dependencies (i.e., subject-verb agreement) which are according to the classes in [10].
- Lexical Density: NMT systems are often known to fail on out-of-domain data. We have built a very lexically dense dataset, containing technical compounds and culturally specific terms, in order to challenge simplification and hallucination in context to maintaining the essence in its entirety.
- Literary Texts: Assesses the maintenance of authorial voice, tone and ambiguity. This data measures the quality of the output, which is human-like (translation of German modal particles, doch, wohl) and figurative language typically occurring in stories, narrations, typical situations and publications.

B. Systems Evaluated

Two most popular NMT systems were compared:

- DeepL Translator: The system is one of the widely considered neural systems that are fluent in European languages. Selection was done carefully to go hand in hand with the German-English language couple.
- Google Translate (GNMT): It is a massive multiple language system, which is built with the Transformer architecture and applicable to various languages in the world.

C. Evaluation Metric

In the case of the quantitative analysis, we have not applied traditional metrics of n-grams like BLEU. According to Reiter, the scores of BLEU are not always associated with practical utility and cannot be applied to test hypotheses about scientific quality systems [7].

Moreover, Mathur et al. [6] evidenced that usual measures tend to conflict with human judgement and cannot detect outliers in well-performing models.

In its place, we used COMET (Cross-lingual Optimised Metric for Evaluation of Translation) [5]. COMET is a neural model that predicts the quality of translations with the use of cross-lingual pretrained language models. In contrast to BLEU, which uses absolute surface level matching, it makes semantic judgments as to consider the source, hypothesis, and reference in a shared embedding space. It can be used to evaluate the Literary and the Idioms data sets more even-handedly: the correct translation might not have any n-grams intersection with the reference, but be semantically identical.

D. Experimental Procedure

Each of the sources in the challenge corpus was manually translated into each of the two systems DeepL and Google Translate. The scores of the outputs were then assigned using the COMET metric against a reference translation completed by humans. To identify the underlying reasons of the language failures in the translation processes, we eventually carried out a qualitative error analysis on specific low-scoring parts and categorized them as syntactic, semantic, or pragmatic errors. We seek language skills which can be expanded in the translation sector.

TABLE I. COMPOSITION OF THE LINGUISTICALLY MOTIVATED CHALLENGE CORPUS

Dataset Category	Sample Size (N)	Target Linguistic Phenomenon
Idiomatic Expressions	50	Non- compositional meaning, cultural nuances.
Structural Stress	30	Complex syntax, verb-final placement, recursion.
Lexical Density	40	Compounding rules, high- context terminology
Literary & Affective	40	Authorial tone, modal particles, ambiguity.
Total	160	

IV. Results and Discussion

This part constitutes a quantitative and qualitative comparison of DeepL and Google Translate (GT) on four linguistically motivated challenge sets. The assessment was done based on the COMET metric, which was complemented by a manual error assessment as a way of classifying systemic failures.

A. Quantitative Performance Overview

Table II lists the descriptive statistics of the two systems. The performance difference was established as evident: DeepL excels in more pragmatically dense categories (Idioms, Literary), which means that this model can be more easily interpreted outside the literal meaning, whereas Google Translate excels at handling rigid syntactic dependencies (Structure), which talks about common data that is already represented in the model in the form of large datasets.

Interestingly, both systems' mean COMET scores were negative in the Structure and Lexical categories (e.g., DeepL $\mu = -0.629$). Negative scores in the COMET embedding space indicate a substantial departure from the semantic vector of the reference translation, indicating that these linguistic phenomena are still problematic for existing NMT architectures.

B. Idiomatic Expressions

The pragmatic gap is represented by it. The Idioms category showed the biggest performance difference (Fig. 1). DeepL outperformed Google Translate (0.553) with a mean score of 0.602.

Failure Analysis: Google Translate often used compositional translation by decoding idioms word by word. The phrase "Ich verstehe nur Bahnof" was found to be rendered literally as understanding the "train station" rather than capturing its idiomatic essence, as can be seen in Table III (Example A).

Success Elements: DeepL successfully mapped the German idiom to its functional English equivalent by using a paraphrasing technique. "I don't understand" or "It's all Greek to me." This suggests that rather than focussing on instantaneous token-to-token alignment, DeepL's attention mechanism is better suited to capture the non-local context.

C. Structural Stress

In this instance, recursive complexity was taken into consideration. Unlike idioms, Google Translate ($\mu = -0.583$) outperformed DeepL ($\mu = -0.629$) in the Structural Stress category.

Long-Distance Dependencies: The finite verb is frequently positioned at the end of the subordinate clause in German syntax. DeepL sometimes "forgot" the initial subject-verb agreement by the time it reached the end of sentences with multiple nested relative clauses (Example B in Table III), which upsets the sentence hierarchy and makes the text harder to read overall.

Clause Hierarchy: While DeepL frequently flattened the structure (A → B and C), changing the casual relationship, Google Translate has proven to be more faithful in maintaining the nesting order of clauses (A → B → C). Both stay far from the translation's reference meaning, discussing. Both remain far from the reference translation meaning, talking about how NMT system lack the capacity to read structural linguistic nuances with accuracy.

D. Lexical Density

This tackles the issue of compound splits. On the Lexical Density dataset, both systems did poorly. High instability is indicated by the high standard deviations ($\sigma \approx 0.8$).

De-compounding Errors: The NMT system must carry out implicit segmentation for German compound nouns, such as Rindfleischetikettierungs. The semantic density of these terms was difficult for both systems to deconstruct into natural English noun phrases.

Terminological Precision: Google Translate can identify certain bureaucratic jargon more accurately than DeepL, probably because it has a larger training corpus of EU legal documents.

E. Literary Texts

The loss of modality is discussed in these texts. These outcomes were qualitatively different but statistically identical (DeepL 0.593, GT 0.589).

Modal particles: The variance in this category is significant, as shown in Fig. 2. The omission of German modal particles (doch, ja, wohl) was a significant source of error. It usually contains the speaker's attitude rather than the propositional content. A "flattened" emotional tone is the result of treating both systems as noise, eliminating them completely (Example D in Table II).

TABLE II. STATISTICAL SUMMARY OF COMET SCORES

Linguistic Category	System	Mean (μ)	Std Dev (σ)	Min	Max
Idiomatic Expression	Google	0.553	0.186	0.289	0.901
	DeepL	0.602	0.154	0.372	0.901
Structural Stress	Google	-0.583	0.804	-1.523	1.222
	DeepL	-0.629	0.783	-1.606	1.222
Lexical Density	Google	-0.583	0.804	-1.523	1.222
	DeepL	-0.629	0.783	-1.606	1.222
Literary Tone	Google	0.589	0.481	-1.079	1.129
	DeepL	0.593	0.450	-1.005	1.129

TABLE III. QUALITATIVE FAILURE ANALYSIS: REPRESENTATIVE EXAMPLES

ID	Category	Source Sentence (German)	Machine Output (Failure)	Linguistic Failure Analysis
A	Idiom	Ich verstehe nur Bahnof.	GT: I only understand the train station.	Literalism: The model failed to detect the non-compositional nature of the phrase, translating constituent tokens instead of the pragmatic meaning ("I don't understand anything.")
B	Structure	Nachdem er, der müde war, gegessen hatte…	DeepL: After he ate, who was tired…	Dependency Break: The relative clause (der müde war) was detached from its antecedent (er), disrupting the syntactic hierarchy and changing the sentence flow.
C	Lexical	Rindfleischetikettierungs…	Both: Beef Labelling Monitoring Task…	Compound Splitting: While technically accurate, the output is a "word salad". Models failed to restructure the complex noun into a fluent English prepositional phrase.
D	Literary	Das ist jawohl ein Witz	GT: This is arguably a joke.	Tone Mismatch: The particle ja wohl implies indignation or disbelief. GT's use of "arguably" introduces a formal, logical tone that contradicts the emotional intent.

V. LIMITATIONS OF STUDY

Although the research is a success in that it provides a high-resolution diagnostic analysis, it has certain methodological constraints.

A. Sample Size

The corpus of linguistically Motivated Challenge(L-MCS) to be used in this paper consists of 160 manually edited sentences. Although this is adequate to determine the failure modes of a particular system in a diagnostic capacity [10], the statistical power is weaker than a large-scale automated benchmark (e.g., WMT test sets with N > 10,000). These findings therefore may be seen to represent qualitative conduct as opposed to a definite global performance index.

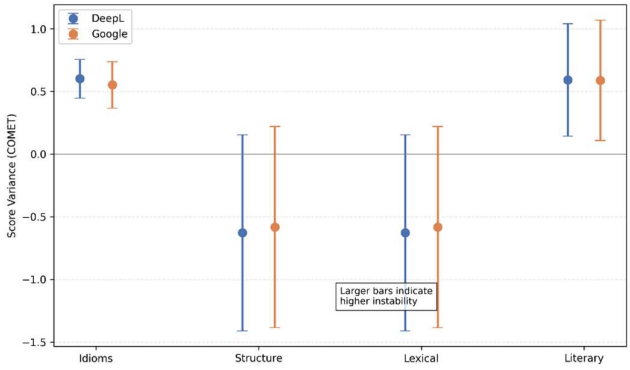

Figure 1. Assessment Of Translation Stability Through Mean Scores And Standard Deviation

B. Manual Curation Bias

The choice of hard examples is based on the human linguistic knowledge, this can introduce the selection bias. The research concentrates solely on the English-German language couple. The results of structural stress (e.g. Verb-Second word order) might not be extended to other language families at all.

C. Model Opacity

Since DeepL and Google Translate are dominant proprietary commercial systems, the specific training information and structure update are comparatively veiled. The behaviours that are observed cannot be immediately tracked down to particular weights of training and data.

VI. Conclusion

This study was an empirical study that involved a comparative analysis of the robustness of Neural Machine Translation, which put Google Translate and DeepL to certain linguistic stress tests. By using the methodology of "Challenge Set," we were in a position to uncover many inconsistencies that would not have been visible in the normal aggregate examinations.

According to our results, there is a strong difference between DeepL and other languages: DeepL is more likely to pick how pragmatics and cultural subtext are. In order to translate literature and idioms it was thus the most suitable. Nevertheless, with the high quantity of information in its data set, Google Translate was identified to require a higher level of structural fidelity. Technical terminology and deeply nested syntactic constructions can both benefit from this. Upon closer inspection, it was discovered that both systems significantly failed to maintain "Authorial Voice" and break down complex compounds. This further demonstrates that contemporary NMT models continue to use lexical simplification in spite of high fluency. These results highlight the need for linguistically focused diagnostics in place of BLEU scores in order to attain true human parity. Recent studies have also explored the use of large language models themselves as evaluators of translation quality, suggesting a new direction for semantic evaluation frameworks [1].

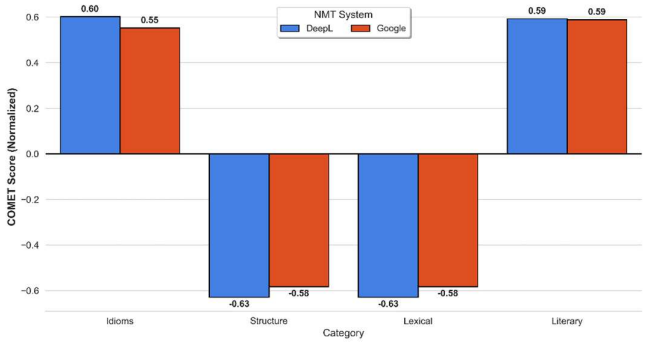

Figure 2. Comparative Performance Analysis of Deepl and GT using Comet Scores

Future Scope

This research is just a comparative study based on highlighting and tabulating the present shortcomings, ways in which those can be overcome is something that can be further explored across three categories:

A. Automation of Adversarial Sets

Further a framework can be developed that can automatically generate "Challenge Sets" eliminating the need of having to curate them manually. This can potentially be done using existing, grammatical templates. This would ensure that large-scale stress testing can be done without having to manually curate the datasets henceforth widening the scope of our study to LLMs and multiple domains.

B. Multilingual Validation

Expanding the L-MCS methodology, the same can be done for numerous other languages. Since our study has initially taken a German-English pair, we can expand our focus towards other European languages like Russian, French, Spanish) This helps determine if the trade-off between
"pragmatics" (DeepL) and "structure" (Google) is a universal architectural feature or something specific to German language which is alien to others.

C. Reference Free Evaluation

Investigating the use of Reference-Free COMET models (COMET-QE) and COMETKiwi for real-time translation quality estimation without requiring human references [2]. so that the current limitation of requiring

an accurate human-verified reference can be overcome. This allows us to evaluate the translation quality in real-time scenarios where human references are unavailable and accuracy is needed.

REFERENCES

[1] J. Robinson et al., "Evaluating Neural Machine Translation with Large Language Models," Transactions of the ACL, 2024.

[2] R. Rei et al., "COMETKiwi: Reference-Free Machine Translation Evaluation," Proc. WMT 2023.

[3] M. Freitag, R. Rei, N. Mathur et al., "Results of the WMT22 Metrics Shared Task: Stop Using BLEU?" Proc. Conference on Machine Translation (WMT), 2022.

[4] E.M. Bender, T. Gebru, A. McMillian-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency, 2021, pp. 610-623

[5] R. Rei, C. Stewart, A.C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in Proc, 2020 Conf. Empirical Methods Natural Lang, Process., 2020, pp. 2685-2702.

[6] N. Mathur, T. Baldwin, and T. Cohn, "Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrices," in Proc. 58th Annu. Meeting Assoc. Comput. Linguist., 2020, pp. 4984-4997.

[7] E. Reiter, "A structured review of the validity of BLEU," Computational Linguistics, vol. 44, no. 3, pp. 393-401, 2018.

[8] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., 2017, pp. 5998-6008.

[9] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in Proc. 1st Workshop on Neural Machine Translation, 2017, pp. 28-39.

[10] P. Isabelle, C. Cherry, and G. Foster, "A challenge set approach to evaluating machine translation," in Proc. 2017 Conf. Empirical Methods Natural Lang. Process, 2017, pp. 2486-2496.

[11] D.Bahdanau, K. Cho and Y. Benegio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473, 2014.