Estimating the Quality of Translated Texts using Back Translation and Resource Description Framework

Vinay Neekhra and Dipti Misra Sharma

Language Technology Research Center (LTRC)
Kohli Center on Intelligent Systems (KCIS)
International Institute of Information Technology, Hyderabad, India
vinay.neekhra@research.iiit.ac.in
dipti@iiit.ac.in

Abstract. How can we effectively estimate the quality of translated texts in the medical field, where back-translation is usually available and/or recommended for sensitive documents. This paper proposes a novel metric, GATE[1], for translation quality estimation task, leveraging the Resource Description Framework (RDF) to encode both semantic and syntactical information of the original and back-translated sentences into RDF graphs. The distance between these graphs is measured to get the semantic similarity score to assess the quality of the translation. Unlike traditional metrics like BLEU and METEOR, our approach is reference-less, capturing both semantic and syntactical information for a comprehensive assessment of translation quality. Our results correlate better with human judgment, giving a better Pearson correlation (0.357) as compared to BLEU (0.200), thereby showing ~70% improvement over BLEU. Our research shows that, in the field of translation evaluation, existing resources like back-translation and RDF could be useful.

Keywords: Resource Description Framework (RDF) · Back Translation · Translation Quality Estimation · GATE

1 Introduction

A drug trial in the medical domain incorporates a mandatory consent form called a Medical Consent Form (MCF), which informs the patient about the experiment and its potential side effects. There is a legal requirement for the MCF to be in the patient's mother tongue and for it to be easy to understand. A human translator translates the original MCF into the patient's mother tongue. As MCFs are sensitive documents, evaluating the quality of translated texts is crucial to ensure faithfulness to the original texts (see Section 1.1 for an example).

One way to evaluate the quality of the translated texts is using back-translation (see Section 3.1), wherein the translated text is translated back into the original

[1] GATE: Graphical Assessment for Translation quality Estimation

language. The original and back-translated texts are then compared to estimate the quality of the translation. Back-translation is a prominent way to assess the quality of translated texts in domains, such as medical documents, where accuracy and precision are paramount [7][4].

Experienced professionals are responsible for carrying out all three procedures (see Figure 1), namely: initial translation from the source language to the target language, followed by translation from the target language back to the source language, and ultimately, comparison between the original text and the back-translated texts. Our efforts are focused on reducing the efforts of human evaluators comparing the original and back-translated texts by automating the task of evaluating the quality of translated texts.

In this paper, we propose a novel translation evaluation metric, GATE (Graphical Assessment for Translation quality Estimation), which leverages back-translation (see Section 3.1) and the Resource Description Framework (RDF) (see Section 3.2). GATE encodes both semantic and syntactical information of the original and back-translated sentences into RDF graphs, allowing for a reference-less, semantically-aware assessment of translation quality.

For sensitive documents in the medical field, such as medical consent forms and qualitative research, back-translation is a common practice to ensure the faithfulness of translations [7][4]. GATE capitalizes on this by integrating back-translation into its evaluation framework, providing a comprehensive and reliable assessment of translation quality. To estimate the quality of translated texts, we encode the meaning of these sentences into graphs using the Resource Description Framework (RDF) and then compare these graphs to come up with a similarity score (See Figure 4). GATE shows a higher correlation (0.357) with human judgment than BLEU (0.200). (see Section 4 for the experiment details). In the next Section 1.1, we discuss the significance of translation evaluation, highlighting the context and motivation behind our research efforts.

1.1 Significance of Translation Evaluation

Consider the following sentence from a medical consent form for a vaccine trial, translated to the patient's mother tongue *(Tamil language)* where the original consent form is in English.

- *Source text:*

There are no side effects mentioned previously.

To comply with legal requirements, the consent form was translated into Tamil by hospital authorities, resulting in two translated versions. For evaluating the translation quality, the translated MCF was back-translated to English, yielding the following results:

- *Back Translation 1:*

No side effects which were mentioned previously

- *Back Translation 2:*

It has been already mentioned that it does not have any side-effects

As seen above, the first back-translated sentence is semantically similar to the source text and preserves the original intent. The second back translated text, on the other hand, conveys that *—as previously mentioned, there are no side-effects—*, whereas the original intent was that no side-effects have been observed yet, thus raising ethical and legal concerns.

Thus, it is crucial, that translated texts are evaluated for their faithfulness to the original text, especially in the medical domain. In the next subsection, we highlight the contributions of our work.

1.2 Contributions

1. This paper presents a novel approach, GATE, for translation quality estimation task by utilizing back-translation and leveraging knowledge graphs (namely, Resource Description Framework) for encoding the meaning of original and back-translated texts to come up with a translation quality estimation score.
2. GATE incorporates both syntactic and semantic information, leading to improved evaluation scores. Our approach is applicable to both machine-translated and human-translated texts. Our experiments demonstrate a better correlation with human judgment compared to BLEU, with a Pearson correlation of 0.357 compared to the most commonly used metric, BLEU's 0.200.
3. Our approach eliminates the need for reference texts by comparing the source text directly with its back-translated counterpart. This makes our approach reference-less and thus valuable for scenarios where reference texts are not available for translation evaluation (such as medical consent forms).
4. While our results do not surpass the current state-of-the-art, our metric, GATE, offers distinct advantages such as requiring no training, being computationally lightweight, being available for low-resource languages, and operating without the need for extensive training data, unlike neural network-based methods like COMET [11].

The paper is structured as follows: Section 2 reviews related work in the area of translation evaluation, discussing the limitations of existing metrics. Section 3 builds the foundation of our work, providing an overview of back-translation along with its significance, introduces Knowledge Graphs in general, and describes Resource Description Framework (RDF) and FRED RDF graphs. Section 4 details the experiment design and methodology leading to the creation of GATE. The results of our experiments are presented in Section 5, along with a discussion of the insights gained from our research efforts while also addressing the current limitations of our metric. Finally, Section 6 and Section 7 conclude the paper along with outlining the directions for future research.

2 Related work

Existing metrics for translation evaluation, such as BLEU[9], METEOR[2], NIST[5], and TER[12], have been widely utilized in the field, with BLEU being the most commonly used among them. BLEU compares the translated sentence with a reference sentence. It operates on word group matching using an n-gram model and remains popular due to its simplicity. In contrast, METEOR was developed as a successor to BLEU to account for synonyms and other variations in language. Usually, the quality of translation is evaluated at the sentence level, but word and document level QE are also possible [13].

However, these metrics have inherent limitations. Many traditional metrics are categorized as n-gram matching metrics, relying on handcrafted features to estimate translation quality by counting the number and fraction of n-grams shared between a candidate translation hypothesis and one or more human references. This restricts their ability to capture nuanced meaning, particularly in complex and domain-specific texts. They often rely on surface-level similarity measures and may necessitate reference translations, typically provided by humans as a standard of perfection.

More recent approaches have explored the use of word embeddings as an alternative to n-gram matching for capturing word semantic similarity. Metrics like BLEU2VEC[14], BERT SCORE[18], and COMET[11] create alignments between reference and hypothesis segments in an embedding space to compute a score reflecting semantic similarity. COMET, a notable metric in this domain, has demonstrated remarkable results for translation evaluation. However, to train these models, the availability of word embeddings for low-resource languages remains a significant challenge.

However, these metrics may still need to catch up in capturing the full range of nuances captured by human judgments. Challenges with existing metrics include their reliance on reference texts for comparison, requiring semantic exactness at the word level, susceptibility to differences in lexical structure (such as word order), and the tendency to measure semantic relatedness rather than semantic similarity, huge data requirement for training models thus not well-suited for low-resource languages.

3 Preliminaries

This section lays out the foundation required for our experiment design.

3.1 Back Translation:

Back translation is a process where a translated text is translated back into the original language (source language) by a different translator [10]. In Figure 1, translation and back-translation processes between English and French are illustrated, as depicted by [15].

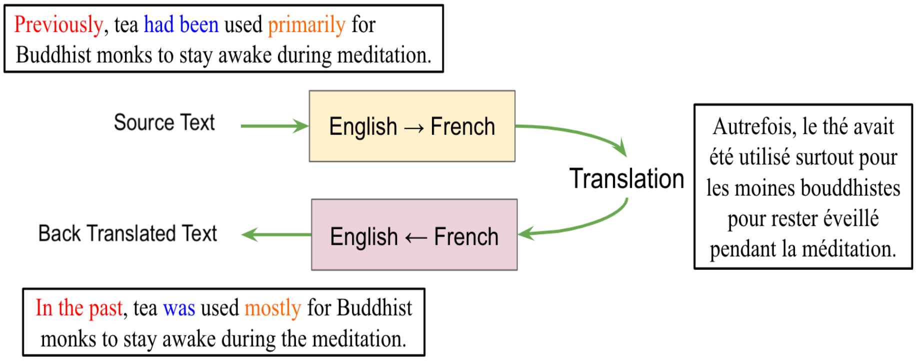

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.
Source Text
English → French
Translation
Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.
Back Translated Text
English ← French
In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

Fig. 1. Example of Back Translation (best viewed in color)

Back translation is recommended in the domains where the content subjected to translation is too sensitive and needs to be double-checked. The back-translation method is widely used in medical research and clinical trials, as it is required by Ethics Committees and regulatory authorities in several countries [7]. This allows us to compare the back-translated text with the original text to evaluate the quality of the translation.

The rationale behind using back-translation is that for sensitive documents in the medical domain, back-translation is a recommended practice to cross-verify that the translation adheres to the intended meaning. Usually, back-translation is mandatory in case of quality assessment of medical consent forms, so this is not an overhead in this particular scenario and is generally recommended for medical, legal, market research, and government agencies working in public health, safety, and legal matters. We are utilizing this for translation evaluation. We aim to address the specific needs of these domains to ensure the faithfulness of the translated texts. Our efforts are to use already available back-translation texts for the translation evaluation tasks.

3.2 Resource Description Framework

The Resource Description Framework (RDF) is a W3C standard for data representation on the Web. RDF provides a foundation for encoding information in a structured way for the Semantic Web [17]. It is particularly useful for representing knowledge about entities and the relationships between them.

Components of RDF RDF consists of triplets, which are fundamental units of information. These triplets, also known as RDF triples, form the building blocks for representing knowledge within an RDF graph. Each RDF triple is composed of three elements:

1. **Subject:** The resource (entity) being described. (e.g., "The patient")
2. **Predicate:** The property or characteristic of the subject, denoted by directed arrows. (e.g., "has diagnosisof")
3. **Object:** The value associated with the predicate for the subject. (e.g., "pneumonia")

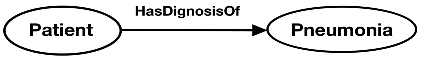


Fig. 2. RDF Triple for the sentence "The patient has diagnosis of pneumonia"

In Figure 2, the RDF triple depicts a statement about a patient having a diagnosis of pneumonia. In the context of our research, we leverage RDF to capture the semantics of the sentences, enabling a more nuanced evaluation of translation quality compared to traditional metrics.

FRED RDF Graphs Our research is based on RDF graphs provided by FRED (Framework for RDF-based Extraction and Disambiguation) [6] to capture semantic nuances in translated texts. At its core, FRED leverages the Resource Description Framework (RDF) to construct semantic graphs that capture the relationships and entities present in the text. FRED bridges the gap between unstructured text and structured knowledge representation, employing Semantic Web technologies to extract and disambiguate information from textual data. Figure 3 shows the RDF graph for the sentence "An experimental drug is one which has not been approved by FDA.".

4 Experiment Design

We conduct a comparative experiment to evaluate the efficacy of our proposed RDF-based evaluation metric, GATE, in comparison to the baseline metric BLEU and its correlation with human judgment. To obtain baseline BLEU scores, we are using iBLEU [8]. The evaluation procedure, outlined in Algorithm 1, explains the comparison of RDF graphs generated through the FRED API, which can be accessed at http://wit.istc.cnr.it/stlab-tools/fred/demo/.

4.1 Dataset

Our experiments were done on the selected medical consent forms and the sentences from Semantic Textual Similarity (STS) Benchmark Dataset [3] to evaluate the effectiveness of GATE in capturing semantic similarity compared to

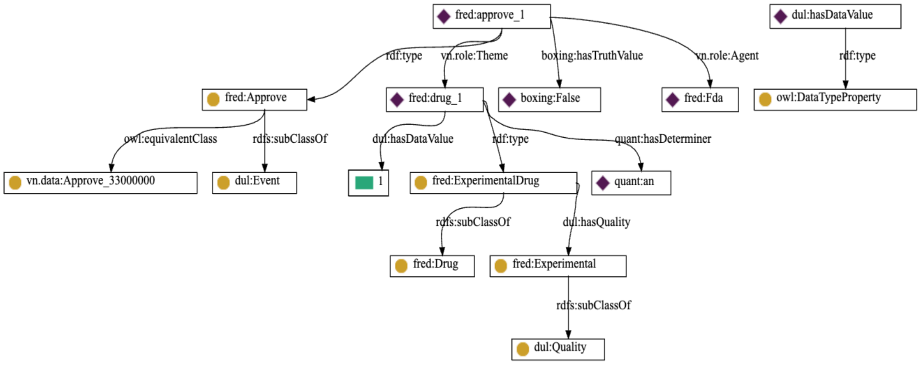


Fig. 3. FRED RDF graph for "An experimental drug is one which has not been approved by FDA." taken from a medical consent form.

BLEU. The medical consent forms dataset has around 250 original sentence, their corresponding translations, and the back-translated texts, all provided by human translators. Due to the selected availability of medical data, we augmented our analysis with the STS benchmark dataset. In total, our experiments were conducted on 500 sentence pairs, with 250 pairs sourced from medical consent forms provided by a medical institute.

4.2 Graph comparison and GATE Score

We are comparing the source sentence with the back-translated text by constructing RDF graphs for both. The distance between graphs is measured as the Jaccard similarity coefficient [16] between the entities in the graphs. This way, the distance between the source and the back-translated sentence graph is normalized between 0 and 1, where 1 denotes an exact match, and 0 denotes no similarity. Algorithm 1 outlines the steps in the evaluation process. Specifically, for source sentence $\mathbf{s}_k$, and the back-translated text $\mathbf{b}_k$, the GATE Score is calculated as follows:

$$G_k = \frac{entities(\mathbf{s}_k) \cap entities(\mathbf{b}_k)}{entities(\mathbf{s}_k) \cup entities(\mathbf{b}_k)}$$

In the next section, we present the findings of our experiments along with a discussion of the insights gained from our research efforts while also addressing the current limitations of our metric.

5 Results & Discussion

Our experiment implemented the proposed GATE metric alongside the baseline metric, BLEU. We calculated the Pearson correlation between the BLEU score

Algorithm 1 : GATE Score evaluation process

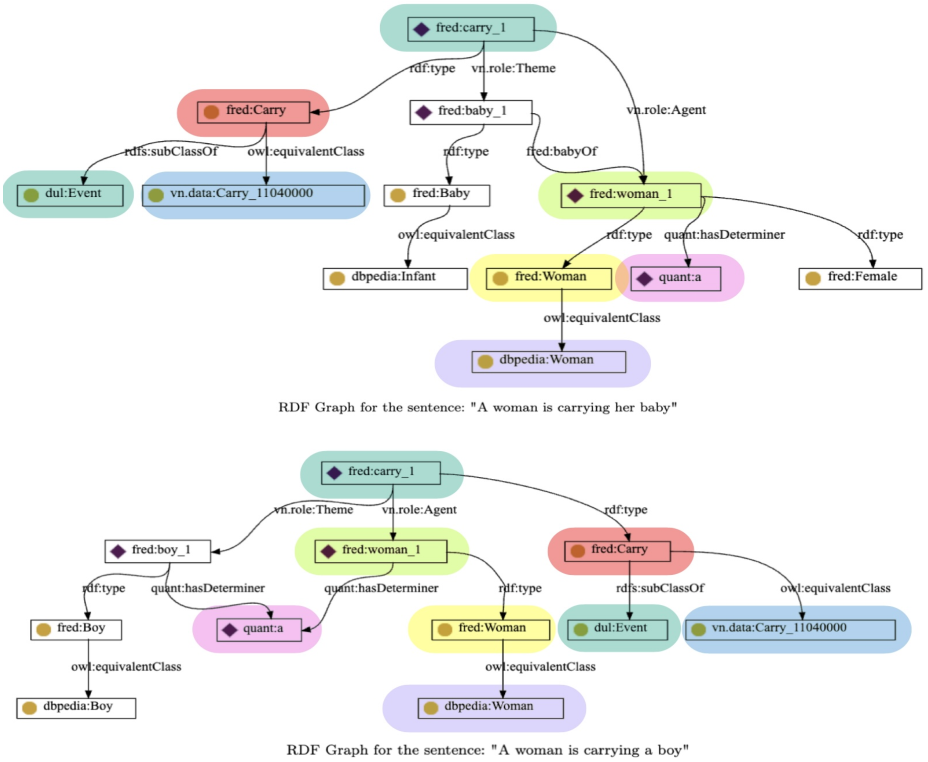
```
Require: All source sentences $\mathbf{s}_k \in \mathbf{S}$ and target sentences $\mathbf{t}_k \in \mathbf{T}$ of $n$ sentence pairs
Ensure: sentence-level scores $\mathbf{G}_k$
for each sentence pair $\{\mathbf{s}_k, \mathbf{t}_k\} \in \{\mathbf{S}, \mathbf{T}\}$ do
    $\mathbf{b}_k \leftarrow$ back-translation of $\mathbf{t}_k$ (either already available or obtained using Google Translate)

    $entities(\mathbf{s}_k) \leftarrow$ RDF graph nodes of $\mathbf{s}_k$ using FRED
    $entities(\mathbf{b}_k) \leftarrow$ RDF graph nodes of $\mathbf{b}_k$ using FRED

    common $\leftarrow \{x \mid x \in entities(\mathbf{s}_k)$ and $x \in entities(\mathbf{b}_k)\}$
    unison $\leftarrow \{x \mid x \in entities(\mathbf{s}_k)$ or $x \in entities(\mathbf{b}_k)\}$

    $\mathbf{G}_k \leftarrow \frac{common}{unison}$

end for
```

and GATE score against human judgment on the experiment dataset. Our results in Table 1, show that GATE achieves a significantly higher correlation with human judgment in translation evaluation tasks compared to the widely used metric, BLEU. Specifically, GATE exhibits a approximately 70% improvement in correlation on the experiment data, with a Pearson correlation coefficient of 0.357 compared to BLEU's 0.200. The higher correlation underscores the effectiveness of leveraging RDF graphs in capturing semantic information, thereby improvement in correlation with human judgments.

Table 1. System-wide Pearson correlation of BLEU and GATE with human judgments on MCFs Data and STS Benchmark Dataset

Metric	Pearson Correlation
BLEU	0.200
GATE	**0.357**

Table 2 shows examples with corresponding human evaluation scores, GATE scores, and BLEU scores. These examples serve to highlight GATE's capability to better reflect human perception of semantic similarity, as evidenced by its closer alignment with human judgments compared to BLEU scores. In summary, our findings indicate that integrating RDF graphs with already existing back-translated texts holds promise for reference-free translation evaluation. This metric can potentially assist human evaluators who evaluate the translation of sensitive documents using back-translated texts.



RDF Graph for the sentence: "A woman is carrying her baby"

RDF Graph for the sentence: "A woman is carrying a boy"

Fig. 4. Graph Comparison for measuring semantic similarity. Common nodes are highlighted in multiple colors. In these two graphs there are 8 common nodes, and total unique nodes are 15. (best viewed in color)

For the Figure 4, the GATE Score is calculated as:

$$G = \frac{8 \text{ (number of common entities)}}{15 \text{ (total unique nodes in both the graphs)}} = 0.53$$

Using RDF for translation evaluation could be helpful as they 'encode' real-world semantics akin to how embeddings work in neural network frameworks (such as COMET), contrasting with metrics that are based on lexical level information for translation evaluation (such as BLEU). This work has the potential to pave the way for utilizing knowledge graphs in the field of translation evaluation alongside existing resources, such as word embeddings and LLM-based frameworks. Our experiments reinforce our belief, demonstrating that using knowledge graphs to encode meaning is helpful and gets better results than the baseline metrics.

Given that RDF is currently available only in English and our metric compares graphs of original and back-translated texts for translation evaluation, our metric is presently only applicable where English is the source language. How-

Table 2. GATE vs. BLEU score against human evaluation.
Selected examples from the experiment run on STS dataset. Higher correlation with human judgment is marked in bold.

Hypothesis	Reference	Human	GATE	BLEU
A man is erasing a chalk board	The man is erasing the chalk board	*1.00*	**0.65**	0.60
Three men are playing guitars	Three men on stage are playing guitars	*0.75*	0.45	**0.60**
A woman is carrying a boy	A woman is carrying her baby	*0.47*	**0.53**	0.63
A woman peels a potato	A woman is peeling a potato.	*1.00*	**1.00**	0.52

ever, the target language can be any other language as long as back-translation is available.

While our results do not surpass state-of-the-art performance, they serve as a proof-of-concept, showcasing the effectiveness of leveraging RDF graphs for translation evaluation tasks. As FRED accommodates large sentences as well, our future work will involve working with more extensive real-world translated medical data and testing our methodology on larger sentences to demonstrate its effectiveness comprehensively. These results underscore the advantages of GATE over traditional metrics like BLEU and motivate further validation of GATE's applicability on real-world data particularly in domains like medicine, along with continuing our exploration for further improvement of the metric.

6 Conclusion

In this paper, we introduce GATE, a novel metric based on the Resource Description Framework (RDF) designed for assessing the quality of translated medical texts for which back-translation is available. To showcase the effectiveness of our metric, we conducted experiments using selected medical data and the STS benchmark dataset, comparing the results against the baseline metric, BLEU, and human judgment scores. Notably, GATE exhibits a stronger correlation with human judgment than BLEU, achieving a higher Pearson correlation coefficient (0.357 compared to BLEU's 0.200), representing approximately a ˜70% improvement over BLEU, the most commonly used metric.

By leveraging back-translation and using RDF graphs to encode both semantic and syntactical information, GATE provides a reference-less and semantically aware assessment of translation quality. In comparison with the more advanced Large Language Model (LLM)-based metrics such as COMET, our metric is computationally much lighter. It works for any target language, including low-resource languages, and does not require any data training. Our research shows that, in the field of translation evaluation, existing resources like back-translation and Resource Description Framework could be helpful in real-world scenarios such as the medical domain.

7 Future Directions

As part of future work, we would like to explore:

1. Conducting further experiments to validate the efficacy of GATE on real-world translated medical data.
2. Since Translation and Summarization can both be viewed as natural language generation from a textual context, we aim to explore knowledge graphs such as RDF in the area of evaluating summarization or similar natural language generation tasks. Investigate the utilization of knowledge graphs for tasks beyond translation evaluation, such as summarization.
3. For calculating GATE score, experimenting with different formulas incorporating variations in weights of entities, incoming edges, and outgoing edges.
4. Addressing the challenge of language dependency in GATE by incorporating multilingual knowledge graphs since FRED works only with English texts. A primary avenue for future work, will be looking into the inclusion of other knowledge graphs available in other languages, making GATE language independent.
5. Development of a software similar to iBLEU for integrating FRED API to facilitate automatic scoring of source and back-translated texts, enhanced visualization, and accessibility of the RDF metric.
6. [1] shows that back-translation could be useful for improving the translation quality for low-resource languages. Our future work is to combine neural networks with back-translation and knowledge graphs in the area of translation evaluation for low-resource languages. Our future work aims to combine these technologies along with knowledge graphs (such as Knowledge Graph Embeddings) to improve our metric, making it suitable for evaluating translated sensitive texts and investigating the potential of combining neural networks with back-translation and knowledge graphs to improve translation quality, particularly for low-resource languages.

Supplemental Material Statement: Source code for Algorithm 1 is available from Github[2]. Selected Medical Consent Forms Dataset cannot be made available as it incorporates private data. However, STS Benchmark Dataset used in the experiment is publicly available.

Acknowledgments. Our sincere gratitude to the late Prof. Ravi Kothari, on whose suggestions this research work was started. We thank anonymous reviewers for their time and valuable suggestions for improving the paper. We also express our gratitude to Supriya Ranjan, Bhavesh Neekhra, Mamatha Alugubelly, and others for their invaluable feedback and support.

[2] https://github.com/vinayneekhra/GATE

References

1. Abdulmumin, I., Galadanci, B.S., Isa, A.: Enhanced back-translation for low resource neural machine translation using self-training. In: Misra, S., Muhammad-Bello, B. (eds.) Information and Communication Technology and Applications. pp. 355–371. Springer International Publishing, Cham (2021)
2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics (Jun 2005), https://aclanthology.org/W05-0909
3. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/s17-2001, http://dx.doi.org/10.18653/v1/S17-2001
4. Chen, H.Y., Boore, J.R.: Translation and back-translation in qualitative nursing research: methodological review. Journal of Clinical Nursing **19**(1-2), 234–239 (2010)
5. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. pp. 138–145 (2002)
6. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., MongiovÃ¬, M.: Semantic Web Machine Reading with FRED. Semantic Web **8**(6), 873–893 (2017)
7. Grunwald, D., Goldfarb, N.M.: Back translation for quality control of informed consent forms. Journal of Clinical Research Best Practices **2**(2), 1–6 (2006)
8. Madnani, N.: ibleu: Interactively debugging and scoring statistical machine translation systems. In: 2011 IEEE Fifth International Conference on Semantic Computing. pp. 213–214 (2011). https://doi.org/10.1109/ICSC.2011.36
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. pp. 311–318. Association for Computational Linguistics (Jul 2002). https://doi.org/10.3115/1073083.1073135
10. Q, A.: What is back translation? (Dec 2021), https://gtelocalize.com/what-is-back-translation/
11. Rei, R., C. de Souza, J.G., Alves, D., Zerva, C., Farinha, A.C., Glushkova, T., Lavie, A., Coheur, L., Martins, A.F.T.: COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 578–585. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.52
12. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 223–231 (2006)
13. Specia, L., Scarton, C., Paetzold, G.H., Hirst, G.: Quality estimation for machine translation, vol. 11. Springer (2018)
14. Tättar, A., Fishel, M.: bleu2vec: the painfully familiar metric on continuous vector space steroids. In: Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Kreutzer, J. (eds.) Proceedings of the Second Conference on Machine Translation. pp. 619–622.

Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4771, https://aclanthology.org/W17-4771

15. Trinh, T.H., Le, T., Hoang, P., Luong, M.: A tutorial on data augmentation by backtranslation. https://github.com/vietai/dab (2019)
16. Wikipedia contributors: Jaccard similarity. https://wikipedia.org/wiki/Jaccard_index/ (Apr 2023)
17. World Wide Web Consortium: Resource description framework (rdf) syntax specification (revised) (1998), https://www.w3.org/TR/PR-rdf-syntax/
18. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)