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

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Vinay Neekhra
200902041
vinay.neekhra@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Estimating the Quality of Translated Texts using Back Translation and Resource Description Framework**" by **Vinay Neekhra**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Dipti Misra Sharma

In the memory of
Dr. Ravi Kothari Sir

Acknowledgments

The time spent in IIIT has been life transformational. I am eternally grateful to meet great people from all walks of IIIT family, who have been a part of my journey at IIIT-Hyderabad.

I would like to express my gratitude to my guide, Dr. Dipti Misra Sharma. She has contributed not only towards my academic growth but tremendously towards my personal growth as well. She was there to support me during the challenging times and gave me the time and freedom required. She once said in an interview that she doesn't keep bonsai plants as they have restricted growth. It helped me cultivating the belief about the role of empathy in students' development, and the importance of giving them space to grow. My sincere gratitude to the late Prof. Ravi Kothari, a great human being and an amazing teacher, whose suggestions played tremendous role in this research work.

Prof. Shatrunjay Rawat Sir has been a guiding light during all my years at IIIT. This work wouldn't have been possible without his support. My Pranams to Mataji and Sangal sir for their time, wisdom, and affection.

I want to thank my friends Supriya Ranjan, my brother Bhavesh Neekhra, Prateek Saxena, Mamatha Didi, Sanchit, labmates Pruthwik sir, Arafat sir, Priyanka Mishra, Jyoti Jha and others for their support in my research journey. Grateful to have my friends, Goutam Hari Tulsiyan, Shubham Tripathi, Devender, Akshansh, Raghvendra, Avni Verma, and others who made the journey enjoyable.

I would also like to take this opportunity to Yuktahaar and NBH Mess staff support, especially Debanjan ji and Dinesh Bhaiya (Juice shop), for providing wholesome and nutritious food all these years.

Lastly, I am grateful to have immense support, love and affection of Maa-Papa, Bhavesh Bhaiya & Vivek Bhaiya, and other family members throughout these years.

Abstract

Natural language translation is an AI-complete problem meaning that if a machine can translat as well as any human being, the machine could be said to be as intelligent as any human being. Once the translation is done, an assessment of the translated text is required for evaluating the translation quality. The goal is to have automatic metrics which can measure semantic equivalence of translated text with the original text having high correlation with human judgment score to save the time and effort in the evaulation process. Existing metrics predominantly focus on syntax, which often fails to capture the intended meaning of the sentences. Our work aims to combine both syntactic and semantic information to better capture the meaning of sentences. In our research we are treating translation as a black box, and focusing exclusively on the quality of the translation once it is completed.

After the translation is completed, how can we effectively estimate the quality of translated texts, where back-translation is usually available and/or recommended for sensitive documents. This work proposes a novel metric, GATE[11], 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. We also propose novel approach of bi-directional entailment among others for measuring the faithfulness of translated texts. Using these approaches, we are able to achieve considerable accuracy on our corpus.

To the best of our knowledge, this is the first effort to use entailment classification and RDF schema representation on back-translated texts to automatically assess the quality of professionally translated texts.

[11]GATE: Graphical Assessment for Translation quality Estimation

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

This chapter provides an introduction to the thesis, addressing key aspects such as translation, the significance and importance of translation evaluation, the challenging nature of translation, related work in the field, and the shortcomings of existing approaches and our motivation behind the work. We conclude the chapter outlining our contributions, including the novel approaches developed for this task.

The focus of this thesis is on translation evaluation, specifically examining how to measure the faithfulness of professionally translated texts, such as legal and medical documents, for which back-translated texts are available. This work introduces novel methodologies, including the use of bi-directional entailment and RDF schema-based comparison, to automatically assess the quality of translated texts.

Given India's linguistic diversity, stakeholders such as hospitals, doctors, and document creators often are not well-versed in the patient's mother tongue. This linguistic barrier poses a significant challenge, since it is crucial from legal and ethical perspective that original intent of the treatment clearly conveyed to the patient. Failure to do so can raise ethical and legal concerns.

1.1 Natural Languages and Translation

At present times, there are at around 7,000 languages in the world. Whereas 23 predominant languages cover half the world population. [6](See Figure 1.1 for language diversity in India)

1.1.1 Challenges in Communication:

Communication is said to be successfull when a thought intended to be conveyed by the speaker is understood exactly as it is by the listener. In other words, communication is successful when what speaker is thinking is understood exactly by the audience. Nothing more, nothing less. Let's take an example for this. Suppose two people are talking to each other. If person "A" thinks of a Unicorn and wants to communicate this idea to the person "B", however due to mental modeling and vocabulary, "A" speaks out loud 'horse' and then person B listens to

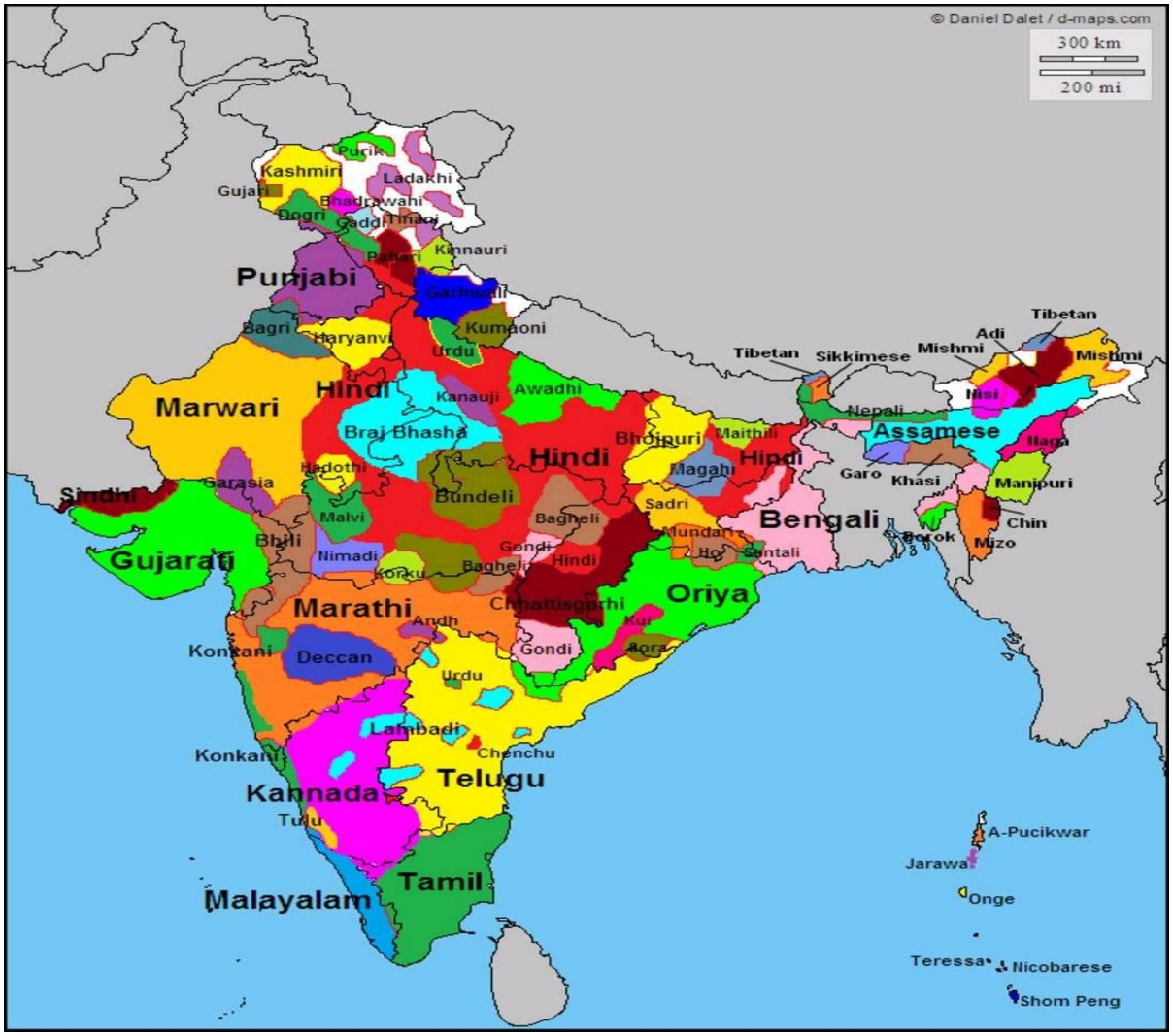


Figure 1.1 Language Diversity in India © Daniel Dalet

it as 'Ox' and forms the mental image of a cow, then communication is not successful. In this example, there are 3 ways where communication can get affected, from thought to voice, voice to hearing, and hearing to thinking.

Translation has been an important part of human civilization since the very begining.

1.2 Translation Evaluation - Significance & Challenges

The human evaluation has been the de-facto standard for translated texts since the need for translation arose. Mostly translation evaluation is done by human experts. However, human evaluation is often subjective, and often there is disagreement among human evaluators. Humans are often biased towards understanding a concept. Even the same person can give a

different evaluation of the same texts at different times. Human evaluation is also often very expensive and unreliable. By unreliable, we mean that given the same input, the same output (judgment about the quality of the translation) is not guaranteed.

1.2.1 Significance of Translation Evaluation

Consider this tweet from USA President during his visit to India

Figure 1.2 USA president tweets in Hindi for his visit in India 2020

To a native Hindi speaker, the translation has little meaning and looks more like a word-to-word substitution. As this communication impacts one of the world's largest and the oldest democracies, it is of crucial importance to evaluate the quality of the translation before it goes for public consumption.

In the 1960s, people assumed that they would solve the problem of translation by using language dictionaries and mapping them from one language to another. However, the challenges of translation remains to this day, as translation is among the most difficult tasks for machines (One of the few AI complete problems as discussed earlier).

1.2.2 Challenging nature of translation

One of the challenges is that cultural context is important in translation. The literal translation doesn't work. In the 60s, it was assumed that creating a dictionary mapping would solve the translation problem. But it was far from it. Very oftern, the same sentiment could be expressed in completely different words. e.g., consider two sentences 'What is your age' and 'How old are you.' There is a little syntactical similarity in these two sentences and they don't

share any common words, however, the intended meaning of these two sentences is almost the same.

This research work is on automatically estimating the quality of translated texts, specifically Medical Consent Forms (MCF). The need for this arises as MCFs are legally required to be in the patient's native language. As the original documents created by hospitals are in English, MCFs are translated from English to the patient's mother tongue.

Most similarity measures capture only syntactical similarity, not the inteded the meaning. e.g., This is a cat. This is not a cat. are very different sentence having opposite meaning, however in metrics like BLEU, Google USE (Universal Semantic Encoder), give an almost perfect score for the similarity of these two sentences due to high word-match count. We want to come up with an approach that somehow encodes the intended meaning of the texts and then evaluate the translation quality.

1.3 Contributions

In the efforts of creating an effective translation evaluation metric, we explored following approaches to get the fidelity score for the translated texts.

1. Rule Based methods
    - Comparing the depedency trees
    - comparing the constituent trees
    - word embeddings based word-alignment
2. Deep learning based methods:
    - Encoding the semantics of the sentences and then comparing them to get the 'semantic distance' between the two sentences.
3. Using bi-directional entailment to estimate the quality of the translated text.
4. GATE: Graphical Assessment for Translation Evaluation:
    - This is our main contribution for the Translation Evaluation problem. We are encoding the meaning of the sentences in RDF graphs and then comparing the distance between these two graphs to get the similarity score. Our experiment results are encouraging and outperform the baseline of BLEU metric.

To the best of our knowledge, this is a novel attempt of utilising back translation with resource description framework and entailment classification for the translation quality estimation task.

Medical documents are also legal documents; hence their translated versions are very sensitive to translation errors having legal and ethical implications. Mostly to check the fidelity of the translated texts Back Translation method is used by professional translators. This process takes lots of time, effort, and money. In this thesis, we propose a method to automate the evaluation process of the Back Translated texts. To the best of our knowledge, no previous work has been done in this direction.

1. This thesis 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 translation evaluation metric is reference-less making it more practical in real-world scenarios. GATE doesn't require reference texts for comparison for the translation quality estimation. This is useful in 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 [13].

1.4 Thesis Organization

Chapter 1 outlines the thesis and gives a brief introduction about the domain, research problems, and the results.

In **Chapter 2**, we discuss the related work done in the field of translation evaluation (BLEU, NIST, etc.). their approach is described in detail in this chapter, discussing the limitations of existing metrics.

Chapter 3 details the experiments design and methodology leading to the creation of GATE. In this chapter, we discuss the two main proposed approaches, a bi-directional entailment classification approach and briefly touches the RDF based translation quality comparison.

Chapter 4 describe the foundation of our work, GATE metric for translation evalution task. We briefly discuss back-translation along with its significance, introduces Knowledge Graphs

in general, and describes Resource Description Framework (RDF) and FRED RDF graphs. Subsequently, we discuss the effects of the target language on the back translation text and the effects on translation evaluation.

Chapter 5 Discusses the results of our experiments, and comparing it with baseline results along with a discussion of the insights gained from our research efforts while also addressing the current limitations of our metric.

Chapter 6 concludes the thesis and discusses possible future work, along with outlining the directions for future research.

Chapter 2

Literature Review

2.1 Translation Evaluation: Brief History

Translation evaluation is one of the most important steps for building and assessing the translation systems. This section presents evaluation methods that have been used by the translation community.

2.1.1 Human Evaluation

Since the translation needs arose, human evaluators played a significant role ensuring that the translation is accurate. After the translation is done, bilingual evaluators proficient in both the source and target languages are presented with the input and output of translation systems and are asked to rate the output on a predefined scale. Typically, two parameters are considered while evaluating any translation: fluency and adequacy. Fluency assesses whether the output sentence is grammatically correct, with judgments given on a scale of 1-4, where '1' signifies intangible output and '4' signifies perfect translation. Adequacy evaluates whether the meaning conveyed by the source sentence has been retained in the target sentence, with a similar scale of 1-4, where '1' denotes no meaning and '4' denotes complete meaning retained in the translated text.

To obtain reliable results, this evaluation is performed by multiple evaluators to mitigate the biasness. The Kappa Coefficient measures the correlation (inter-annotator agreement) between the evaluators, calculated as

$$K = \frac{p(A) - p(E)}{1 - p(E)} \tag{2.1}$$

where $p(A)$ is the proportion of time evaluators agree, and $p(E)$ is the proportion of time evaluators agree by chance.

For comparing outputs among multiple translations, the aforementioned methods are generally inconsistent. Instead of judging sentences on an absolute scale, it is advisable to rank the translations relative to each other.

2.1.2 BLEU

BLEU (Bilingual Evaluation Understudy) is one of the most widely used methods of automatic evaluation for machine-translated text, where n-gram precision is computed with respect to a reference translation. To account for shorter translations, a brevity penalty (BP) is added when words are missing. The primary limitation of this method is that it solely relies on matching n-grams between the translation and the reference output, failing to capture word sequences that have a similar meaning. Conversely, it is possible to achieve a high BLEU score even if the meaning is entirely different due to minor adjustments in n-gram placement. In BLEU evaluation, it is advisable to use multiple reference translations against a system translation to account for all acceptable translations of ambiguous parts. The brevity penalty is defined as:

$$\text{Brevity Penalty} = \min\left(1, \frac{\text{output length}}{\text{reference length}}\right) \tag{2.2}$$

The BLEU score is then calculated as:

$$\text{BLEU} = \text{Brevity Penalty} \times \exp\left(\sum_{i=1}^{4} \log(\text{precision}_i)\right) \tag{2.3}$$

BLEU was one of the first metrics to achieve a high correlation with human judgments of quality and remains one of the most popular automated and inexpensive metrics due to its simplicity and explainability. The BLEU score ranges from 0 to 1, where 0 indicates low translation quality and 1 indicates the best translation quality with respect to the reference translation.

2.1.3 NIST

The NIST framework is an enhancement of the BLEU metric with several notable modifications. While BLEU calculates n-gram precision by assigning equal weight to each n-gram, NIST introduces a measure of informativeness for each n-gram. In this approach, when a correct n-gram is found, the weight assigned to it depends on its rarity; rarer n-grams receive higher weights.

For instance, if the bigram ”on the” is correctly matched, it will receive a lower weight compared to the correct matching of the bigram ”interesting calculations,” as the latter is less likely to occur and more valuable from perspective of meaning being conveyed.

Additionally, NIST differs from BLEU in its calculation of the brevity penalty. In the NIST framework, small variations in translation length have a less significant impact on the overall score, making it more tolerant of minor discrepancies in translation length.

2.1.4 METEOR

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric has demonstrated better correlation with human judgment scores in translation evaluation. Unlike BLEU and NIST, METEOR computes the harmonic mean of unigram precision and recall, with a higher weight given to recall than to precision. This approach not only checks the overlap between translations but also incorporates stemming and synonymy matching using WordNet. For example, words such as "good" and "well," which are treated differently in BLEU and NIST, are considered equivalent in METEOR, thereby being closer to human evaluation.

Precision (P) and recall (R) are calculated as follows:

$$P = \frac{n}{n_c} \quad \text{and} \quad R = \frac{n}{n_r}$$

where n is the count of unigrams present in both the candidate and reference translations, n_c is the count of unigrams in the candidate translation, and n_r is the count of unigrams in the reference translation.

The F-measure is then computed with recall given more weight than precision:

$$F = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

To account for the contiguous occurrence of larger segments, METEOR introduces a penalty score p, calculated as follows:

$$p = 0.5 \cdot \left(\frac{\text{number of chunks}}{\text{mapped unigrams}} \right)$$

The final METEOR score for a sentence is computed as:

$$\text{METEOR} = F \cdot (1 - p)$$

2.1.5 Word Error Rate

The Word Error Rate (WER) is a translation evalution metric based on the Levenshtein distance. The Levenshtein distance is calculated at the word level. Originally utilized for assessing the performance of speech recognition systems, WER is also employed in the evaluation of machine translation systems. The metric calculates the number of words that differ between a machine-translated text and a reference translation.

2.1.6 Google Universal Sentence Encoder

Universal Sentence Encoder (USE) transforms the text into high-dimensional vectors suitable for various natural language processing tasks such as text classification, semantic similarity, and clustering.

USE leverages a diverse range of data sources for accommodating different natural language understanding applications. This model accepts variable-length English text as input and generates a 512-dimensional vector as output. We are utilising this model to convert the text into vector space for evaluating semantic similarity using the Semantic Textual Similarity (STS) benchmark dataset. Google USE uses a deep averaging network (DAN) for encoding the text meaning in vectors.

2.1.7 Semantic Textual Similarity (STS)

Semantic Textual Similarity (STS)[1] measures the degree of semantic equivalence between a pair of sentences and is applicable to tasks in Machine Translation and Summarization among others (Agirre et al., 2012). These approaches can be categorized into three broad types (Han et al., 2013):

- **Vector Space Approaches**: Texts are represented as bag-of-words vectors, and a vector similarity measure (e.g., cosine similarity) is used to compute the similarity score between two texts.
- **Alignment Approaches**: Words and phrases in the two texts are aligned, and the quality or coverage of the alignments is used as the similarity measure.
- **Machine Learning Approaches**: Multiple similarity measures and features are combined using supervised machine learning. This approach relies on the availability of training data.

2.2 Conclusion

In this chapter, we discussed several translation evaluation techniques with their pros and cons. We also explained in detail the working of a statistical model as the major research work focuses on the improvements of these models.

In the next chapter, we provide a description of few of our approaches to estimate the quality of translated texts.

[1]From https://www.aclweb.org/anthology/S15-2046

Chapter *3*

Our Approaches

3.1 Design requirements

A good translation evaluation metric should adhere to the following:

3.1.1 Functional Requirements

1. Consistent: same input -> same output
2. Reliable: main evaluation attributes: grammatical (adequacy and fluency) usability, accesibility

3.1.2 Non-functional Requirements

1. **General:** applicable irrespective of text domain

Other important criteria to consider:

1. **Fidelity** refers to the extent to which a given translation accurately represents the underlying message or meaning of the source text without distortion. Could also be seen as faithfulness to the original text. Transliteration is closely aligned with this approach, although it often fails to properly convey the message due to its rigid faithfulness to the original document. It is essential to distinguish between fidelity and fluency; a translation can be fluent, presenting grammatically correct sentences, yet not be faithful if it does not accurately convey the source text's intended meaning. For example, given any source text if the output is a given presidential speech, the translation could be said to be fluent but it lacks faithfulness to the original text.
2. **Transparency** pertains to the degree to which a translation caters to native speakers and the target audience, such that idiomatic, syntactic, and grammatical conventions are followed while cultural, political, and social context is kept in mind while translating. Some creative freedom is required by the translators to adapt to their audience.[2]

In our research, we have focused on Back Translation approaches for translation evaluation as in many areas like legal and medical, back translation is usually recommended and it makes the translation evaluation easier for humans.

3.2 Back Translation

For legal and medical texts such as medical consent forms (MCFs), back translation is crucial in quality assessment. Back translation is discussed in detail in Chapter 4.

3.2.1 Shortcomings of Back Translation

Back translation often fails to capture nuances and may ignore ambiguities, as illustrated by the example:

- Source Sentence in English: She saw him.
- Translation in Hindi: Usne use dekha.
- Back Translation: Multiple (4) options (e.g., He saw her, She saw her, etc.)

Back translation in this case would fail to adhere to its original text, and while translating back in the source language may create the output not similar to the original text.

3.3 Experiments with Different Approaches

In this section, we layout various approaches we tried for the translation evaluation task. A more detailed explanation of the bi-directional entailment and GATE Score metric is provided in the next chapter.

3.3.1 Rule-Based Fidelity Scoring

1. Dependency Tree-Based Approach
2. Constituent Tree-Based Approach
3. Word Alignment with Word Embeddings Distance

Rule-based translation evaluation faces numerous challenges, such as handling lexical ambiguity and structural variability, which often lead to suboptimal results.

3.3.2 Deep Learning-Based Models

We have investigated several deep learning-based models for translation evaluation. Existing systems often fail to adequately measure semantic similarity due to issues like 'Lexical Overlap' as discussed by R. Thomas McCoy in "Right for Wrong Reasons".

3.4 Using Bi-Directional Entailment for Translation Evaluation:

In this approach, we propose using bi-directional entailment to measure translation faithfulness.

3.4.1 Entailment Definition

Entailment is defined such that a sentence A entails sentence B if the truth of B is inherently present in the truthfulness of sentence A. For example:

- A: Some men are playing basketball.
- B: People are playing a sport.

In this case, A entails B. However if sentence B is true, A may or may not be true.

Bi-directional entailment involves ensuring entailment from sentence A to sentence B and vice versa. Bi-directional entailment ensures that both the sentences discuss have the same subject matter. For example:

- Sentence 1: This is a cat.
- Sentence 2: This is not a cat.

Although semantically close, these sentences should not be considered equivalent as the entailment is not present.

Our approach uses the concept that if sentence A entails sentence B and B entails A, then both sentences are indeed discussing the same thing.

Using the Stanford Natural Language Inference (SNLI) corpus, we classify the source text and back translated text pair into one of three categories: 'Entailment', 'Neutral', or 'Contradiction'. We assess the bi-directional entailment of the source and back-translated sentences, and based on this classification, a score is assigned to the sentence pair (e.g., E-E, E-N, N-C). Our exploration of bi-directional entailment using 'Entailment classification' with Neural Semantic Encoding (Munkhdalai 2017) did not yield satisfactory results. And led us to explore Knowledge Graphs for encoding the meaning of the sentences.

premise	hypothesis	relation	output
there have been good results with the use of a superoxide spray towards the hea	the use of superoxide spray in the direction of treatment of long standing wounds has had good results	entailment	neutral
there will be no side effects of using the spray	using spray will not have any side effects	entailment	entailment
it may decrease the duration of dressings to half as compared to the standard di	it can reduce the length of the dressing compared to standard dressing	entailment	neutral
there are various tools in the attempt to enhance wound healing	there are various tools in an effort to increase wound healing	entailment	neutral
superoxide spray will be used in this trial	in this test superoxide spray will be used	entailment	contradiction
there are no side effects mentioned previously	there are no side effects mentioned earlier	entailment	contradiction
it is considered safe and easy to use	it is considered safe and easy to use	entailment	entailment
if you are unhappy you can withdraw from the study with no implications	if you are unhappy you can withdraw from the study without any implication	entailment	neutral
if there are any enquires please contact me	if you have any inquiries please contact me	entailment	neutral

Figure 3.1 Entailment classification examples

3.5 Proposed approach: GATE: Graphical Assessment for Translation Evaluation using RDF Graphs comparison

Figure 3.2 RDF XML Graph for 'There is a cat on the mat'. This figure showcases the complexity of RDF graphs even for simple sentences.

In this approach we are encoding the meaning of the sentences into RDF graphs for a more informal structure to make the comparison of two sentences easier.

For graph comparison, we utilize the NX library to compute Graph Edit Distance (GED). There are various algorithms available for comparing GED. As illustrated in Figure 3.2, RDF graphs tend to be very complex, even for simple sentences. This complexity often leads to a poor correlation between the semantic comparison of the sentences, as observed in our results 3.3. The chart presents the score comparison for different sentence pairs, comparing the BLEU, RDF_similarity, and Ground Truth scores. As we can see from the chart, BLEU is much closer to the human judgment than RDF Similarity scores.

Instead of comparing complex RDF XML graphs, we experimented with comparing simplified graphs to reduce the unnecessary complexity for graph comparison. For instance, Figure 3.4 shows the simplified graph for the sentence 'There is a cat on the mat'. Similarly, Figure 3.5 depicts the RDF Simplified Schema for the sentence 'Cat is my favourite animal'. A more

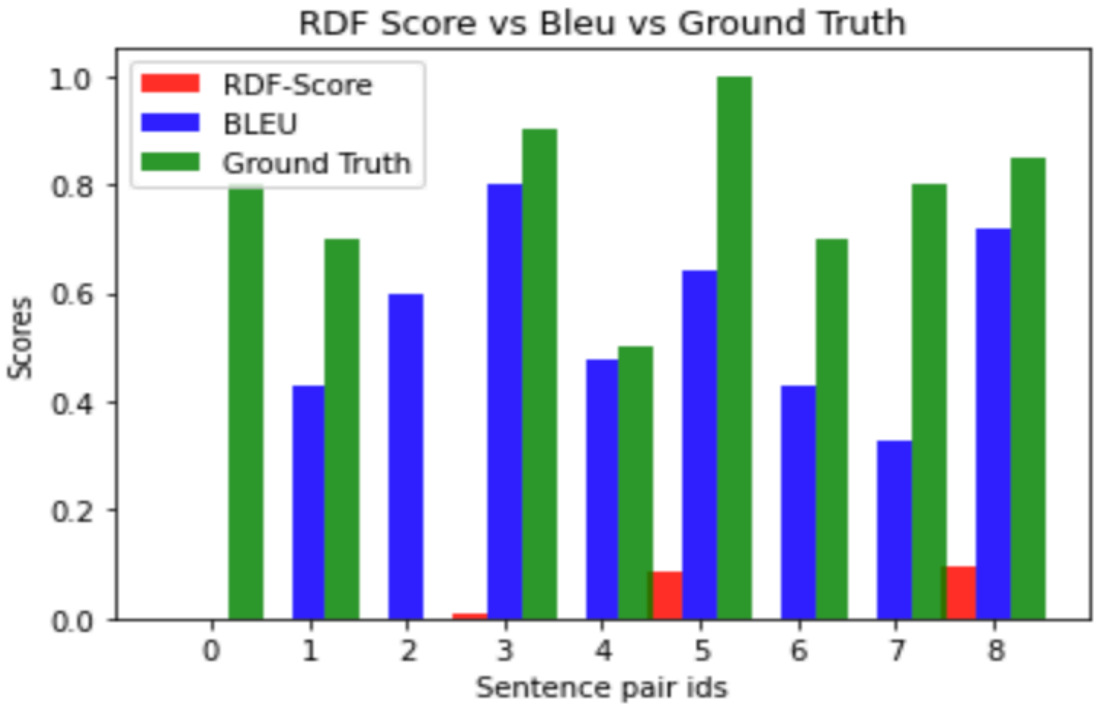


Figure 3.3 Sentence pair IDs and their corresponding BLEU, RDF_similarity, and Ground Truth scores.

detailed explanation of the RDF GATE metric is provided in the next chapter. This section serves to provide a glimpse of one of the many approaches we have tried.

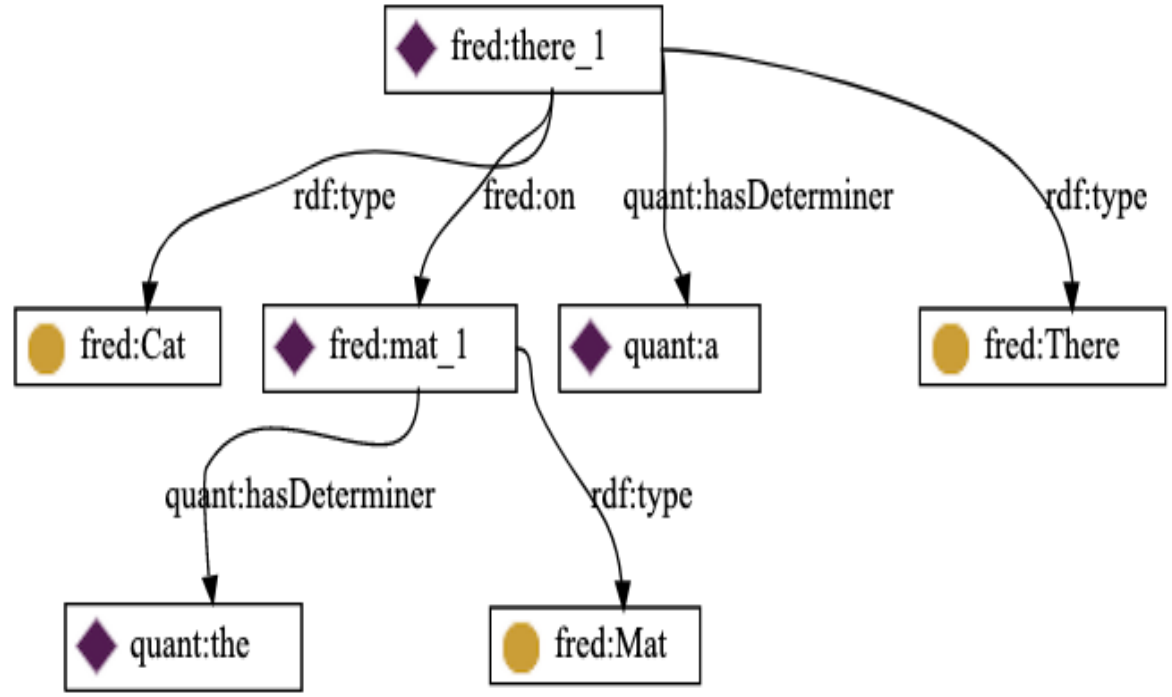


Figure 3.4 RDF Simplified Graph for 'There is a cat on the mat'.

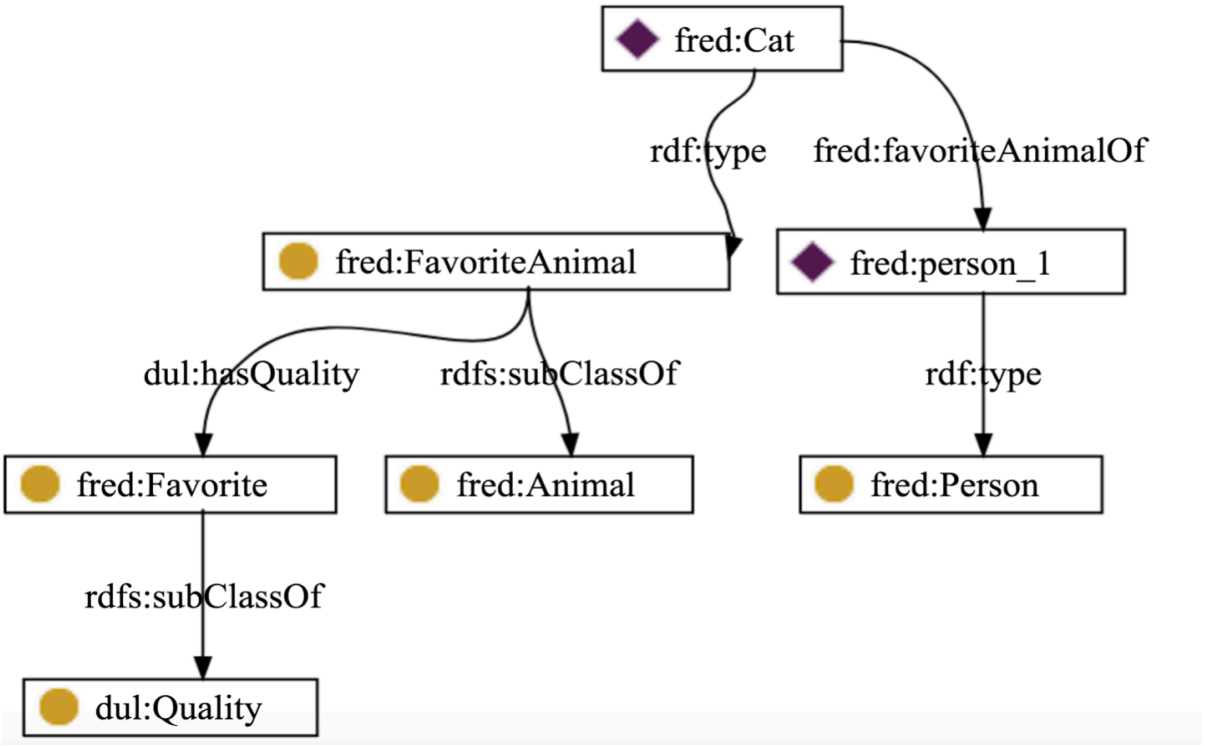


Figure 3.5 RDF Simplified Schema for the sentence: 'Cat is my favourite animal'.

Chapter 4

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

This chapter describes a part of the work done in the paper titled "Estimating the Quality of Translated Medical Texts using Back Translation Resource Description Framework", which has been published in the 7th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics SeWeBMeDA-2024.

4.1 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 medical treatment/trial 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 4.1.1 for an example).

One way to evaluate the quality of the translated texts is using back-translation (see Section 4.3.1), wherein the translated text is translated back into the original 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 [9][5].

Experienced professionals are responsible for carrying out all three procedures (see Figure 4.1), 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.

While human evaluation has traditionally served as a benchmark for assessing translation quality, it is often expensive, time-consuming, and subjective. As an alternative, automatic

evaluation metrics such as BLEU[11], METEOR[3], etc., have been developed to provide a more efficient and objective means of evaluation, with BLEU being the most commonly used metric (see Section 4.2 for related work). This field of research, called translation quality estimation (QE), is an area of research concerned with evaluating the quality of translated texts when gold standard translations (called reference texts) are unavailable.

In this chapter, we propose a novel translation evaluation metric, GATE (Graphical Assessment for Translation quality Estimation), which leverages back-translation (see Section 4.3.1) and the Resource Description Framework (RDF) (see Section 4.3.2). 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 [9][5]. Our metric, 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 4.4). GATE shows a higher correlation (0.357) with human judgment than BLEU (0.200). (see Section 4.4 for the experiment details). In the next Section 4.1.1, we discuss the significance of translation evaluation, highlighting the context and motivation behind our research efforts.

4.1.1 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 the 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 already been 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 of crucial importance, that the translated texts are evaluated for their faithfulness to the original text, especially in the medical domain. In the next section, we highlight the contributions of our work.

4.1.2 Contributions

1. This chapter 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 [13].

The chapter is structured as follows: Section 4.2 reviews related work in the area of translation evaluation, discussing the limitations of existing metrics. Section 4.3 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 4.4 details the experiment design and methodology leading to the creation of GATE. The results of our experiments are presented in Section 5.1, along with a discussion of the insights gained from our research efforts while also addressing the current limitations of our metric. Finally, Section **??** and Section 6.1 conclude the chapter along with outlining the directions for future research.

4.2 Related work

Existing metrics for translation evaluation, such as BLEU[11], METEOR[3], NIST[7], and TER[14], 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, these metrices evaluates the quality of translation at the sentence level, but word and document level QE are also possible [15].

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[16], BERT SCORE[20], and COMET[13] 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.

4.3 Preliminaries

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

4.3.1 Back Translation:

Back translation is a process where a translated text is translated back into the original language (source language) by a different translator [12]. In Figure 4.1, translation and back-translation processes between English and French are illustrated, as depicted by [17].

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

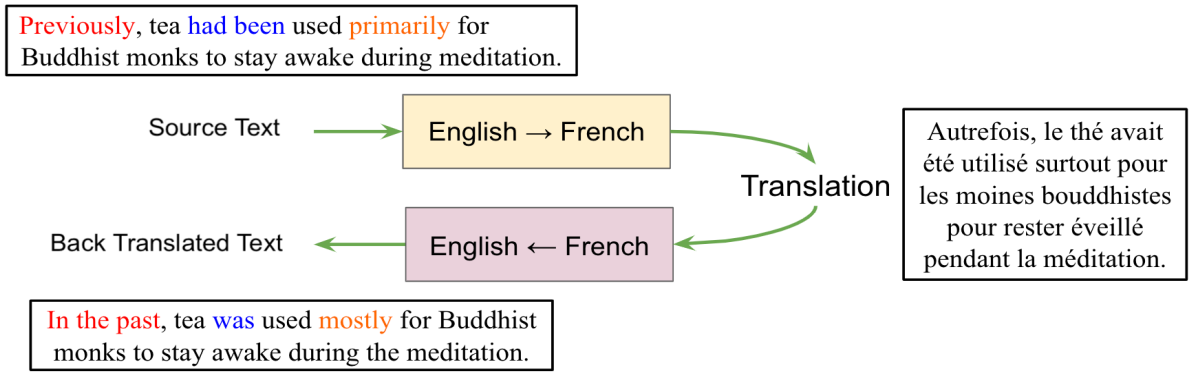


Figure 4.1 Example of Back Translation (best viewed in color)

authorities in several countries [9]. 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.

4.3.2 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 [19]. It is particularly useful for representing knowledge about entities and the relationships between them.

4.3.2.1 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”)

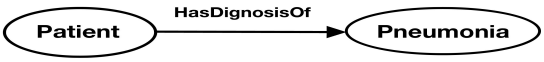


Figure 4.2 RDF Triple for the sentence “The patient has diagnosis of pneumonia”

In Figure 4.2, 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.

4.3.2.2 FRED RDF Graphs

Our research is based on RDF graphs provided by FRED (Framework for RDF-based Extraction and Disambiguation) [8] 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 4.3 shows the RDF graph for the sentence “An experimental drug is one which has not been approved by FDA.”.

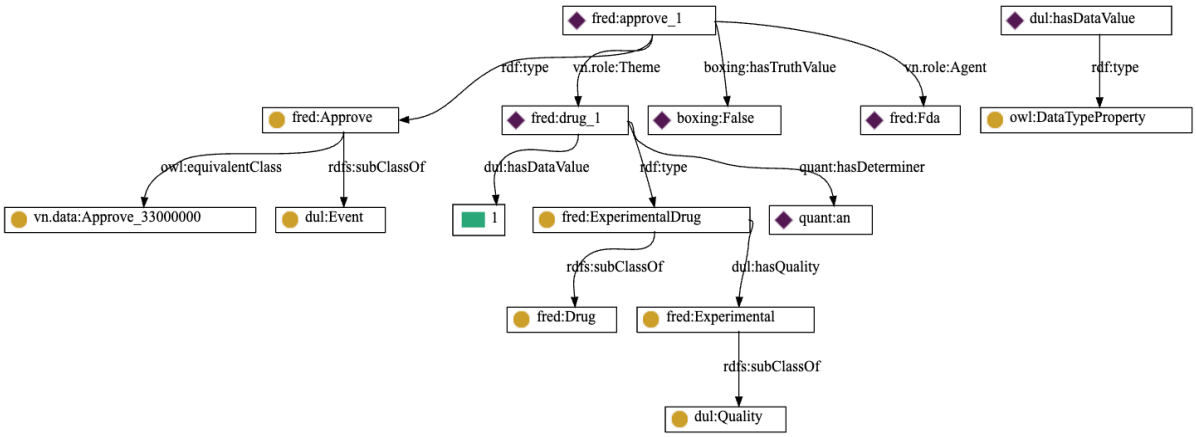


Figure 4.3 FRED RDF graph for “An experimental drug is one which has not been approved by FDA.” taken from a medical consent form.

4.4 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 [10]. 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/.

4.4.1 Dataset

Our experiments were done on the selected medical consent forms and the sentences from Semantic Textual Similarity (STS) Benchmark Dataset [4] to evaluate the effectiveness of GATE in capturing semantic similarity compared to 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.

4.4.2 Graph comparison and GATE Score

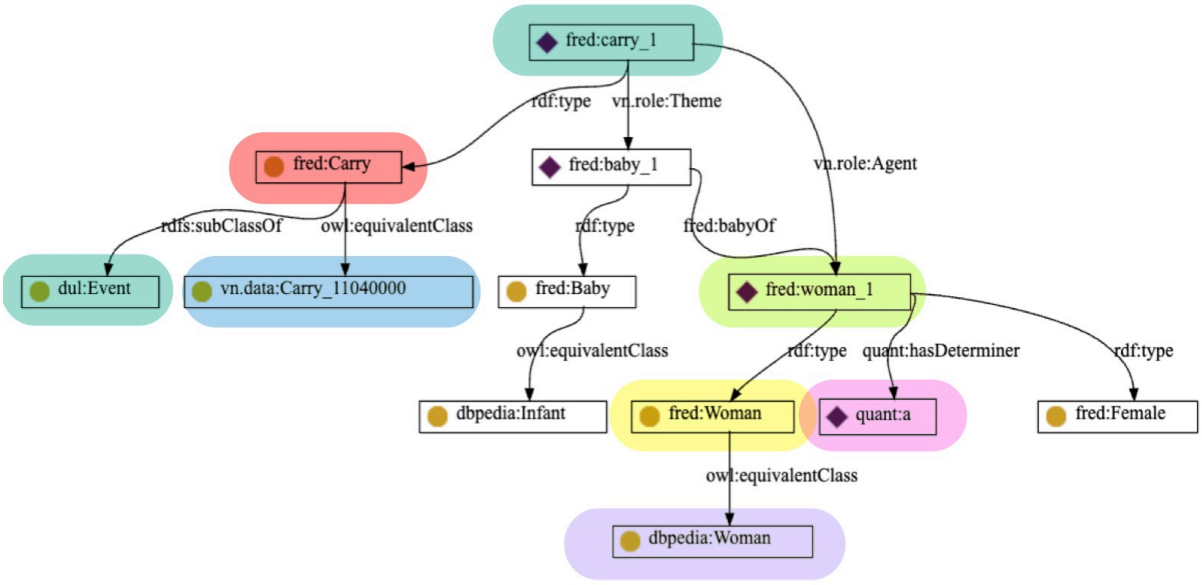
We are comparing the source sentence with the back-translated text by constructing RDF graphs for both. The distance between graphs is measured as the Jaccard similarity coefficient [18] between the entities in the graphs. This way, the distance between the source and the back-translated sentence graph 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:

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

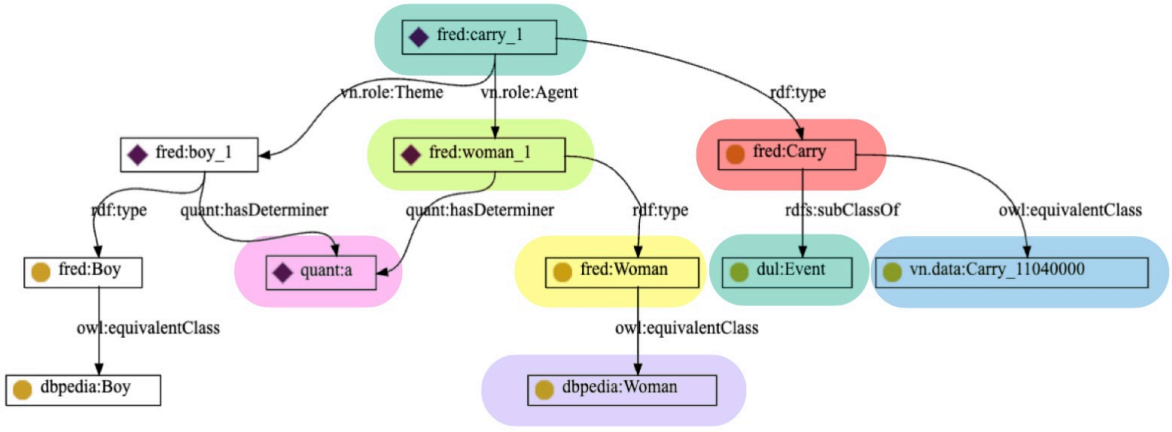
For the Figure 4.4, the GATE Score is calculated as:

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

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.



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



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

Figure 4.4 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)

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 $\mathbf{G}_k$

1: **for** each sentence pair $\{\mathbf{s}_k, \mathbf{t}_k\} \in \{\mathbf{S}, \mathbf{T}\}$ **do**
2: $\quad \mathbf{b}_k \leftarrow$ back-translation of $\mathbf{t}_k$ (either already available or obtained using Google Translate)
3:
4: $\quad entities(\mathbf{s}_k) \leftarrow$ RDF graph nodes of $\mathbf{s}_k$ using FRED
5: $\quad entities(\mathbf{b}_k) \leftarrow$ RDF graph nodes of $\mathbf{b}_k$ using FRED
6:
7: $\quad$ **common** $\leftarrow \{x \mid x \in entities(\mathbf{s}_k)$ **and** $x \in entities(\mathbf{b}_k)\}$
8: $\quad$ **unison** $\leftarrow \{x \mid x \in entities(\mathbf{s}_k)$ **or** $x \in entities(\mathbf{b}_k)\}$
9:
10: $\quad \mathbf{G}_k \leftarrow \dfrac{common}{unison}$
11:
12: **end for**

Chapter 5

Results and Analysis

5.1 Results & Discussion

Our experiment implemented the proposed GATE metric alongside the baseline metric, BLEU. We calculated the Pearson correlation between the BLEU score and GATE score against human judgment on the experiment dataset. Our results in Table 5.1, 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 ~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 5.1 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 5.2 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

Table 5.2 GATE vs. BLEU score against human evaluation. Selected examples from the experiment run on STS dataset. Higher correlation with human judgment are 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

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.

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.

- The Pearson correlation for Jaccard similarity score on RDF nodes is higher than BLEU (in the initial experiment) (0.357 vs 0.200)
- We used the semantic textual similarity(STS) benchmark data (a small subset of it) for this
- Further experiments with different touples (of incoming edge+node, outgoing edges+node, incoming+outgoing+node) showed lower scores (0.132, 0.113, 0.051)
- One more experiment just using the node for Jaccard similarity with the different group of sentences showed Pearson correlation of 0.156 compared to 0.213 of BLEU score

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. However, 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.

In this chapter, we primarily focus upon the results of RDF graph edit

Pearson correlation

Having converted the sentences into RDF graphs and then getting Graph Edit Distance between them, we come to compare how close the sentences are in terms of their semantics.

Chapter *6*

Conclusion and future work

In this thesis, 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.

6.1 Future Directions

The future scope of the present work could be:

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.

7. Future work could focus identifing if corrections are needed in the original document itself before the transalation?

8. In current experiments only node values are taken for the graph comparison. A future scope of this work could be tuples as the unit of comparison for similarity could be looked upon. Each tuple could consists of either of or combination of: a list of incoming edges, node values, a list of outgoing edges.

Related Publication

[P1] Vinay Neekhra, Dipti Misra Sharma **"Estimating the Quality of Translated Medical Texts using Back Translation & Resource Description Framework"**, in proceedings of *the 7th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics SeWeBMeDA*, 2024.

Bibliography

[1] I. Abdulmumin, B. S. Galadanci, and A. Isa. Enhanced back-translation for low resource neural machine translation using self-training. In S. Misra and B. Muhammad-Bello, editors, *Information and Communication Technology and Applications*, pages 355–371, Cham, 2021. Springer International Publishing.

[2] O. H. T. Authors. Fidelity transparency - how do they work in translation?, Apr. 2020. URL `https://www.getblend.com/blog/fidelity-vs-transparency/`.

[3] S. Banerjee and A. Lavie. 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*, pages 65–72. Association for Computational Linguistics, June 2005. URL `https://aclanthology.org/W05-0909`.

[4] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. 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. doi: 10.18653/v1/s17-2001. URL `http://dx.doi.org/10.18653/v1/S17-2001`.

[5] H.-Y. Chen and J. R. Boore. Translation and back-translation in qualitative nursing research: methodological review. *Journal of Clinical Nursing*, 19(1-2):234–239, 2010.

[6] D. Dalet. Linguistic map of india, 2024. Map image © Daniel Dalet.

[7] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[8] A. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, and M. MongiovÃ¬. Semantic Web Machine Reading with FRED. *Semantic Web*, 8(6):873–893, 2017.

[9] D. Grunwald and N. Goldfarb. Back translation for quality control of informed consent forms. 2, 03 2006.

[10] N. Madnani. ibleu: Interactively debugging and scoring statistical machine translation systems. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 213–214, 2011. doi: 10.1109/ICSC.2011.36.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318. Association for Computational Linguistics, July 2002. doi: 10.3115/1073083.1073135.

[12] A. Q. What is back translation?, Dec. 2021. URL `https://gtelocalize.com/what-is-back-translation/`.

[13] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.52`.

[14] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 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*, pages 223–231, 2006.

[15] L. Specia, C. Scarton, G. H. Paetzold, and G. Hirst. *Quality estimation for machine translation*, volume 11. Springer, 2018.

[16] A. Tättar and M. Fishel. bleu2vec: the painfully familiar metric on continuous vector space steroids. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4771. URL `https://aclanthology.org/W17-4771`.

[17] T. H. Trinh, T. Le, P. Hoang, and M. Luong. A tutorial on data augmentation by back-translation. *https://github.com/vietai/dab*, 2019.

[18] Wikipedia contributors. Jaccard similarity. `https://en.wikipedia.org/wiki/Jaccard_index`, 2023.

[19] World Wide Web Consortium. Resource description framework (rdf) syntax specification (revised), 1998. URL `https://www.w3.org/TR/PR-rdf-syntax/`.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.