HMIST: Hierarchical Multilingual Isometric Speech Translation using Multi-Task Learning Framework for Automatic Dubbing

Nidhir Bhavsar[$], **Aakash Bhatnagar**[&], **Muskaan Singh**[#] and **Petr Motlicek**[#]
[$]University of Potsdam, Potsdam, Germany
[&] Boston University, Boston, Massachusetts
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
bhavsar@uni-potsdam.de, aakash07@bu.edu,
(msingh,petr.motlicek)@idiap.ch

Abstract

In this paper, we present an approach and impact of isometric neural machine translation on the automatic dubbing process. The length of generated isometric translated sentences ranges within a $\pm 10\%$ of the source text. We propose a hierarchical and multilingual approach toward generating isometric translation via publicly available MUST-C, WMT, and IIT-Bombay(en-hi) datasets. Our experiments use namely, German(de), French(fr), Russian(ru), Italian(it), and Hindi(hi) languages. Additionally, we implement a paraphrasing module with Opusparcus(fr,de,ru), PAWS-X(fr,de) and Topaco(ru) datasets for German, French, and Russian languages to enhance the vocabulary and maintain the isometric constraints. In performance analysis, we report the average length range of source to translation, 55.15% for all languages, while ru exhibits the highest with 62.475% and a relative improvement of 23.04% from the baseline OPUS-MT model.

1 Introduction

Isometric translation is a relatively new concept in neural machine translation. As video content reaches worldwide, it becomes crucial to localize it for different regions. One of the major problems faced while translating media content is the synchrony between the translated output and the visual content. This problem mainly occurs due to a considerable variation in vocabulary between different languages. [1] state that the ideal length of the generated output should be within $\pm 10\%$ range of the source length. The recent machine translation models do not have any parameters to control the length of the output sequences.

To solve the problem mentioned earlier, we fine-tune different pre-train language models using the prompt engineering method. The initial step in all our methods is to identify the appropriate prompt. We use the approach described in [2] to generate prompts while training. Prompt engineering is an efficient way to perform transfer learning while fine-tuning a model.

In this paper, we present, an empirical analysis of our different translation and paraphrasing models. In our approach, the best performing translation model is OPUS MT and the most efficient paraphrasing model is mBART [3]. However, we train the paraphrasing model only for German, Russian and French because of the limited number of languages supported by paraphrasing datasets. We use a combination of Opusparcus and Topaco datasets for Russian, and Opusparcus and PAWS-X datasets for French and German. As per our knowledge, while writing this paper, we cannot find a standard paraphrasing corpus for Italian and Hindi languages.

2 Background

In this section, we further explain neural machine translation in section 2.1, and its corresponding controlling output length in section 2.2 with lexical and length constraint.

2.1 Neural Machine Translation

For a language pair with parallel data as 1, an MT model parameterized with θ, trains to maximize likelihood on the training sample pairs 2.

$$\mathcal{D} = \{(s_i, t_i) : i = 1, \ldots, N\} \tag{1}$$

$$L(\theta) = \theta \arg\max \sum_{i=1}^{N} \log p\left(t_i \mid s_i, \theta\right) \tag{2}$$

2.2 Controlling output length of MT

Several attempts have been made to control the output length attributes, this includes user preference for desired length summarization [4] or using of multiple extractive summarization algorithms for strict length constraints [5], use of side-information [6] or source text involvement and formality [7] [8]. There can be 2 major approaches for constraining sequence length using MT: i) lexically constrained translation, ii) length constrained translation.

2.2.1 Lexically constrained MT

This section includes lexical integration of length-constraint in NMT, either via constrained training or decoding. [9] replaced recognized entities (URL and number) with place-holders which are then detokenized during post-processing. [10] employed a transformer model, augmenting source phrases with target translations to maintain translation consistency while also allowing the machine to learn lexicon translations by duplicating source-side target terms. On the contrary, [11] leverage the effectiveness of Levenshtein Transformers by injecting terminology constraints at inferences time without any significant impact on decoding speed while also mitigating the re-training procedure.

2.2.2 Length constrained MT

[1] injects length control information via the positional encodings of the self-attention, thus enriching the input embedding in source and target with positional information. [12] extend this approach by computing the distance from every position to the end of the sentence, which is further summed with input embedding in the decoder network. [1] combines the methods for biasing the output length by i) conditioning the output to target-source length ratio and ii) enriching the decoder input with relative length embeddings computed according to the desired target string length. [13] involves translating sentences in source language containing pause marker information and integrates verbosity control of phrases between consecutive pause markers.

3 Proposed Methodology

Our methodology incorporates the approach presented by [2]. As shown in figure 1, our model architecture comprises two main components: 1) Translation module and 2) Paraphrasing Module. We experiment with multiple models for both translation and paraphrasing. Our best approaches consist of OPUS MT [14] as the translation model and

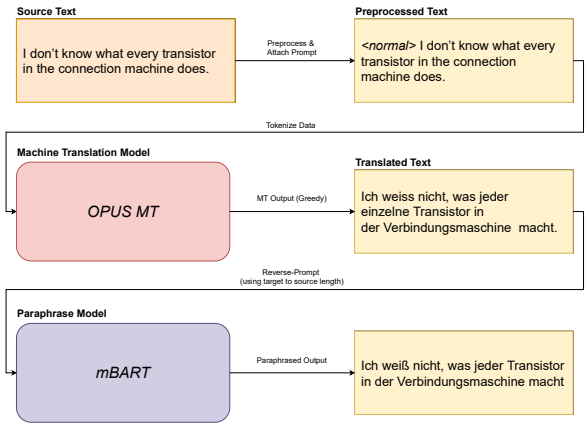


Figure 1: This is the system pipeline of our best-proposed method. After the first step of attaching length prompts, we pass the data from our fine-tuned translation model(OPUS-MT). Further, we pass these translated sentences through a paraphrasing model(mBART) for de,fr,ru.

mBART paraphrasing model. The intuition behind paraphrasing is to increase the vocabulary of the target language. As the primary purpose of isometric translation is to control output length, the paraphrasing model, when used with prompt engineering, helps us adhere to this constraint. For example, the sentence "The name of this person is John" can also be written as "This is John." This example is a precise instance of how paraphrasing can vary the length of generated translated sentences. It is only possible when the model can understand the semantics of a sentence and have the vocabulary to rewrite it. The following two subsections will further elaborate on our two modules.

3.1 Machine Translation

As mentioned earlier, we use OPUS MT for translation. OPUS MT is trained on 108 different languages and performs very efficiently on most languages. Figure 2 describes the general architecture of the OPUS MT model. The OPUS MT model is trained using the Marian NMT framework [15] and constitutes 6-self attentive layers in both encoder and decoder network with 8 attention heads in each layer. However, using baseline OPUS MT does not yield ideal results, as evident in table 4. To comply with the length constraints, we make use of prompt engineering methods and leverage the flexibility of multi-task learning. Our model learns which sentences fall within the normal, short, and long range through prompts. While testing, we add a normal prompt to all input sentences so that model

language	model	dataset	BLEU Score	BERT Score	Length Ratio	Length Range
de	OPUS-MT	MuST-C + WMT	**42.3**	**0.85**	**1.087**	49.81
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	29.1	0.83	1.04	50.55
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	29.9	0.83	1.05	**51.95**
it	OPUS-MT	MuST-C	**34**	**0.84**	**1.045**	**57.032**
fr	OPUS-MT	MuST-C	**44.8**	**0.87**	1.08	49.6
	OPUS-MT+ MT5	MuST-C	42.3	0.85	1.12	51.3
	OPUS-MT + MT5	MuST-C	38	0.86	1.11	46.4
	OPUS-MT + few short mBART	MuST-C + Opusparcus + Paws-X	40.9	0.85	**1.03**	57.33
	OPUS-MT + mBART	MuST-C + Opusparcus + Paws-X	41.2	0.85	1.04	**61.81**
ru	OPUS-MT	MuST-C + WMT	**22.7**	**0.84**	**1.005**	54.517
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	20.8	0.82	0.95	58.934
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	21.7	0.83	0.967	**62.475**
	MT5	MuST-C + WMT	5.6	0.76	0.732	19.3
hi	OPUS-MT	IITB-En-hi	**11.9**	**0.84**	**0.941**	**42.521**

Table 1: Evaluation scores of various experiments. In this table, we state the language-wise experiments along with the datasets used

Source - Target Langauge	Total Instances	Avg. Source Length	Avg. Target Length	Length Ratio	Length Range %
MuST-C					
en-fr	275K	101.78	112.31	1.141	37.65
en-de	229K	100.77	108.84	1.319	36.93
en-it	253K	103.97	108.2	1.076	47.66
en-ru	229K	104.25	102.14	1.044	43..212
IIT-B Corpus					
en-hi	1.65M	74.81	72.92	1.043	46.95
WMT					
en-de	4.5M	138.26	152.45	1.204	28.12
en-ru	2.5M	107.33	98.75	1.18	38.29
Tapaco					
ru-ru	29K	26.38	26.35	1.025	49.45
Opusparcus					
ru-ru	150K	15.81	15.84	1.059	54.948
fr-fr	940K	19.3	19.31	1.079	39.34
de-de	590K	19.63	19.64	1.074	39.33
PAWS-X					
fr-fr	940K	120.94	122.32	1.004	82.61
de-de	50K	119.02	118.24	1.003	85.27

Table 2: Dataset Statistics

generates isometric output.

3.2 Paraphrasing & Length Correction

This module significantly improves our score concerning the isometric constraints. As mentioned above, the main goal of applying the paraphrasing module is to enhance the vocabulary to write sentences with similar meanings in different ways. After exhaustive experimentation, we find that the mBART model is most suitable for paraphrasing. We chose a multilingual version of BART [3] because of its auto-encoding capabilities, which allows it to fully comprehend the language of the text it is parsing, thus making it the best fit for paraphrasing tasks. mBART is a model for text generation in different languages. As seen in table 2, it is evident that the paraphrasing module improves the length ratio and length range significantly.

This module also implies a few-short learning approach via prompt engineering as described by [16].

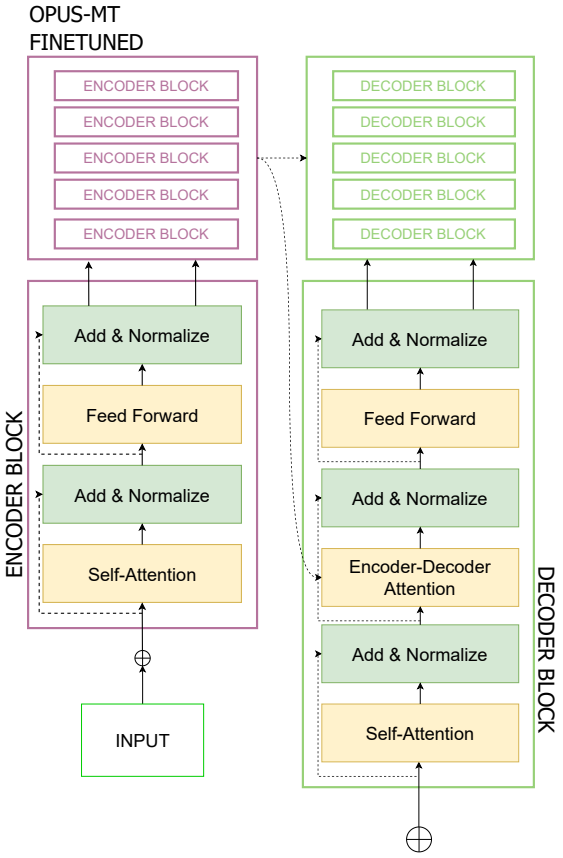


Figure 2: A detailed architecture diagram of our translation model OPUS-MT

The prompts while training are decided similarly to the translation module. However, the difference here is in selecting prompts while making inferences. Given that this model follows the translation module, we slightly employ a different methodology for choosing the prompts for the predictions. We use the paraphrasing module to shorten the

long translated outputs or lengthen the short translated outputs. Equation 3 represents the process of finding the appropriate prompt for paraphrasing prediction.

$$f(x) = \begin{cases} long, & LR < 0.95 \\ short, & LR > 1.10 \end{cases} \quad (3)$$

In equation 3 LR is computed by the length of generated translated text divided by the source text. We do not apply paraphrasing to the translated sentences under the normal range. The paraphrasing module's reverse prompts improve our results and indicate that few short learning is performing as expected. This also indicates that the mBART model successfully understands the length constraint during paraphrasing.

4 Experimental Setup

In this section, we describe dataset details in section 4.1, hyper-parameter settings in Section 4.2 and training procedures in Section 4.3

4.1 Dataset

We implement different datasets for translation and paraphrasing module. As shown in table 1, for machine translation, we use the Multilingual Speech Translation Corpus (MuST-C) [17] for translating majority of our source languages. We also use the Statistical Machine Translation Dataset (WMT) [18] for German (de) and Russian (ru). Additionally for translating Hindi (hi) we use the IIT-B English-Hindi Corpus [19]. Next, we use a combination of Opusparcus [20] and PAWS-X [21] datasets for most of our Paraphrase training tasks, However, due to unavailability of PAWS-X dataset for Russian (ru), we utilize the Tapaco dataset [22] which is a sub-extracted paraphrase corpus derived from the Tatoeba database [23].

4.2 Hyper-parameter Settings

We used 4 Tesla V100-PCIE GPU for all experiments with a memory size of 32510 MiB each. Due to resource constraints, we train each of our models for 1 epoch with a batch size of 32. We apply a learning rate of 2×10^{-5} with a weight decay of 0.01. We implement the AdaFactor optimizer [24], which internally adjusts the learning rate based on the scale parameter and relative/warmup steps.

4.3 Training

In this section, we will discuss details of all experiments performed. In [2], the hierarchical approach was implied only on the French MUST-c dataset with OPUS-MT and MT5 [25] models. This paper extends that approach to five different languages and includes mBART in the paraphrasing module. Our results stand out, and the mBART model performs better than MT5 as a paraphrasing model. This paper also uses reverse-prompt in the paraphrasing module, significantly impacting the results.

We employ prompt engineering on OPUS-MT and MT5 translation models because languages like Italian (it) and Hindi (hi) lack a standard paraphrase dataset. There are very few MT models that support en-hi translation, we utilize only OPUS-MT for this task.

We also implement a singleton MT5 model that performs translation and paraphrasing of 5 supported languages using the prompt engineering method. We utilize the MT5 model for this singleton approach, one of the most optimized multi-task learning models. We use two additional prompts in this approach: 1) Translation and 2) Paraphrasing. The translation prompt signifies that the model will translate the given input, and the paraphrasing prompt signifies that the model will generate isometric sentences from the translated sentences. Further, these two prompts were combined with length prompts. In this model, we use the MUST-C dataset for translation and PAWS-X and Topaco for paraphrasing. However, the sentences generated by this MT5 model are very absurd. After our analysis, we find that the model is mixing up different languages. One possible reason for this is the small dataset size (approx 200K) for each language.

In our experimental setup, we adopt a prompt-based few-shot learning strategy for the paraphrasing task. The model utilizes a small sample of the training dataset(approx 500) and then tries to integrate the derived model with the predictions obtained from the MT model. The comparative scores achieved by assessing using the same technique are listed in table 1. The few-shot model can constrain the output length adequately while also preserving the semantical aspects of the MT. We employ the few-shot learning strategy to train the pre-trained mBART model like the main paraphrase module. Instead of facilitating downstream fine-tuning via pre-training on different corpora, we focus on using

Source Text	Translated Text	Reverse-Pormpt	Paraphrased Sentence	CL Length	BERT Score
I don't know what every transistor in the connection machine does.	Ich weiss nicht, was jeder einzelne Transistor in der Verbindungsmaschine macht.	Short	Ich weiß nicht, was Jeder Transistor in der Verbindungsmaschine macht.	70	0.964
		Normal	Ich weiß nicht, was der einzelne Transistor in der Verbindungsmaschine macht.	77	0.951
		Long	Ich weiß nicht, was der einzelne Transistor in der Verbindungsmaschine macht.	77	0.951
Say, "Please repeat that process." Score them again.	Sag: "Bitte wiederholen Sie diesen Vorgang. " Zählen Sie sie noch einmal.	Short	Bewerte sie nochmal	20	0.749
		Normal	Sag, "Bitte wiederhole diesen Prozess." Bewerte sie.	52	0.979
		Long	Sag, "Bitte wiederhole diesen Prozess" und bewerte sie nochmal.	63	0.908

Table 3: In this table the first & second column represents the source & the generated translated text respectively. The third column shows the value of reverse-prompt that we append on the translated output generated by our model. fourth column represents the paraphrased text generated by our paraphrasing module. CL is the character length of the paraphrased sentences.

accessible data samples to perform few-shot learning. We utilize the prompt-engineering techniques to extend the pre-trained model's performance for the specific task of paraphrasing non-isometric text. We use a similar data configuration for the following source languages: de, fr, and ru.

5 Result and Analysis

For evaluating isometric translation outputs for we use BLEU, BERTScore and length compliance.

- *BLEU* [26] score is a statistical method that evaluates on the basis of n-grams in translated and reference text. Particularly for isometric translation, where the length of translated sentence may vary from the reference text, BLUE score is unable to capture the semantic meaning.

- *BERT score* [27] however, uses pre-trained contextual word embeddings to calculate cosine similarity between translated sentences and reference text. BERT score is the most appropriate option because it can evaluate sentences based on semantics and is comparatively more robust while evaluating short translation sentences.

- *Length Compliance* [28]. is isometric constraint specific evalaution metric which comprises of two measure: (1) Length Ratio and (2) Length Range. Length Ratio is defined as the ratio of source text by generated text. Length Range is defined as the percentage of sentences that falls within ideal span of length ratio 0.90-1.10.

5.1 Evaluation Measures Analysis

Adhering to isometric constraints can negatively affect the derived BLEU score as it depends on the number of characters. The best performing models for the isometric task across each source language have a lower BLEU score, as seen in table 1. This dip in the BLEU score is fairly evident in the languages that use the paraphrase module. The paraphrasing module modulates sentence length to conform to the interchangeable vocabulary. Each of the following language pairs, (en-de, en-fr, en-ru) have a high BLEU score for the baseline OPUS-MT. However, the BLEU score changes abruptly when the paraphrase module is applied, although the value of length compliance metrics improves. Consequently, we recommend the BERT score as a similarity measure since it provides a more precise similarity assessment while taking the semantical

Language	Model	BLEU Score	BERT Score	Length Ratio	Length Range(%)
de	baseline OPUS	**33.1**	**0.84**	1.14	35.74
	finetuned OPUS + mBART	29.9	0.83	**1.05**	**51.95**
it	baseline OPUS	31.3	0.82	**1.037**	54.662
	finetuned OPUS	**34**	**0.84**	1.045	**57.032**
fr	baseline OPUS	**45.4**	**0.86**	1.149	35.41
	finetuned OPUS + mBART	41.2	0.85	**1.04**	**61.81**
ru	baseline OPUS	20.4	0.83	**1.001**	50.776
	finetuned OPUS + mBART	**21.7**	**0.83**	0.967	**62.475**
hi	baseline OPUS	9.9	0.83	0.844	31.911
	finetuned OPUS	**11.9**	**0.84**	**0.941**	**42.521**

Table 4: Comparison of our language-wise best performing systems with the baseline OPUS-MT models

component of translations into account.

5.2 Comparison with Baseline OPUS

We see a significant difference in the length compliance metrics when we compare our results with the pre-trained OPUS-MT model. As depicted in table 4, our best performing models have shown improvement compared to the respective baselines OPUS models. In contrast to the baseline OPUS-MT models, our models can provide length-controlled outputs. In table 4, it is visible that the BLEU score of the pre-trained OPUS-MT model is better than our models in most of the cases; however, there is no significant difference in the BERT score. These statistics further reinforce our point of using the BERT score as an evaluation measure for isometric translation. Particularly for Russian, our predictions Exhibit a high length range of 62.475% and a Length Ratio of 0.967. Moreover, even though the OPUS en-hi corpus is used for pre-training, OPUS-MT consists of only 24M entries compared to 421.5M for de, 550.7M for fr, 241.4M for it, and 160M for ru, we were still able to improve the BLEU score and BERT score.

5.3 Qualitative Analysis

Table 3 provides two instances that demonstrate our reverse-prompt method. As mentioned in earlier sections, reverse-prompt was designed to assist the paraphrasing module in constructing length-controlled phrases. As shown in the second instance of table 3, when the short prompt is applied, our model ignores the content enclosed within the double-inverted commas. At the same time, the long prompt tries to append extra vocabulary to increase the prediction length. As shown in table 3, all three types of outputs (short text, long text, and normal text) exhibit a similar BERT score except for some extreme cases. This consistency in the BERT score represents that our paraphrasing model maintains the semantic meaning while controlling the length of the generated output. When we applied the reverse-prompt approach to Google's MT5 model, the results were not promising compared to mBART. We believe that mBART is more optimized for paraphrasing tasks and prompt-engineering mechanisms.

5.4 Automatic Dubbing Analysis

The main purpose of isometric translation is to establish synchrony between the source speech and the translated speech. As stated previously, the ideal range of LR is $0.90 - 1.10$ for source-text translation with isometric constraints so that text-to-speech(TTS) modules can produce a more synchronous output. To analyse this statistic, we use Amazon's Polly(Joanna speaker) [1] and Google's text-to-speech(default speaker) [2] models. Figure 3 shows eight graphs, each representing the time taken by AWS Polly for speaking source language, target language, and our generated isometric translated sentences. We can see that the green and blue lines tend to come together in most cases. This reinforces the claim made by authors of [28] regarding the ideal LR value.

In contrast to AWS Polly, Google's text-to-speech model does not differentiate much between our generated translated outputs and the target outputs. In figure 4 we can see that there is a lot of overlap between the red line and the green line. It is evident that AWS Polly was producing different duration of speech, while Google's Text-to-Speech is not recognizing the difference for the same sen-

[1] https://aws.amazon.com/polly/

[2] https://cloud.google.com/text-to-speech

(a) Opus-mt-en-de (b) mBART-en-de (c) Opus-mt-en-ru

(d) mBART-en-ru (e) Opus-mt-en-fr (f) mBART-en-fr

(g) Opus-mt-en-it (h) Opus-mt-en-hi

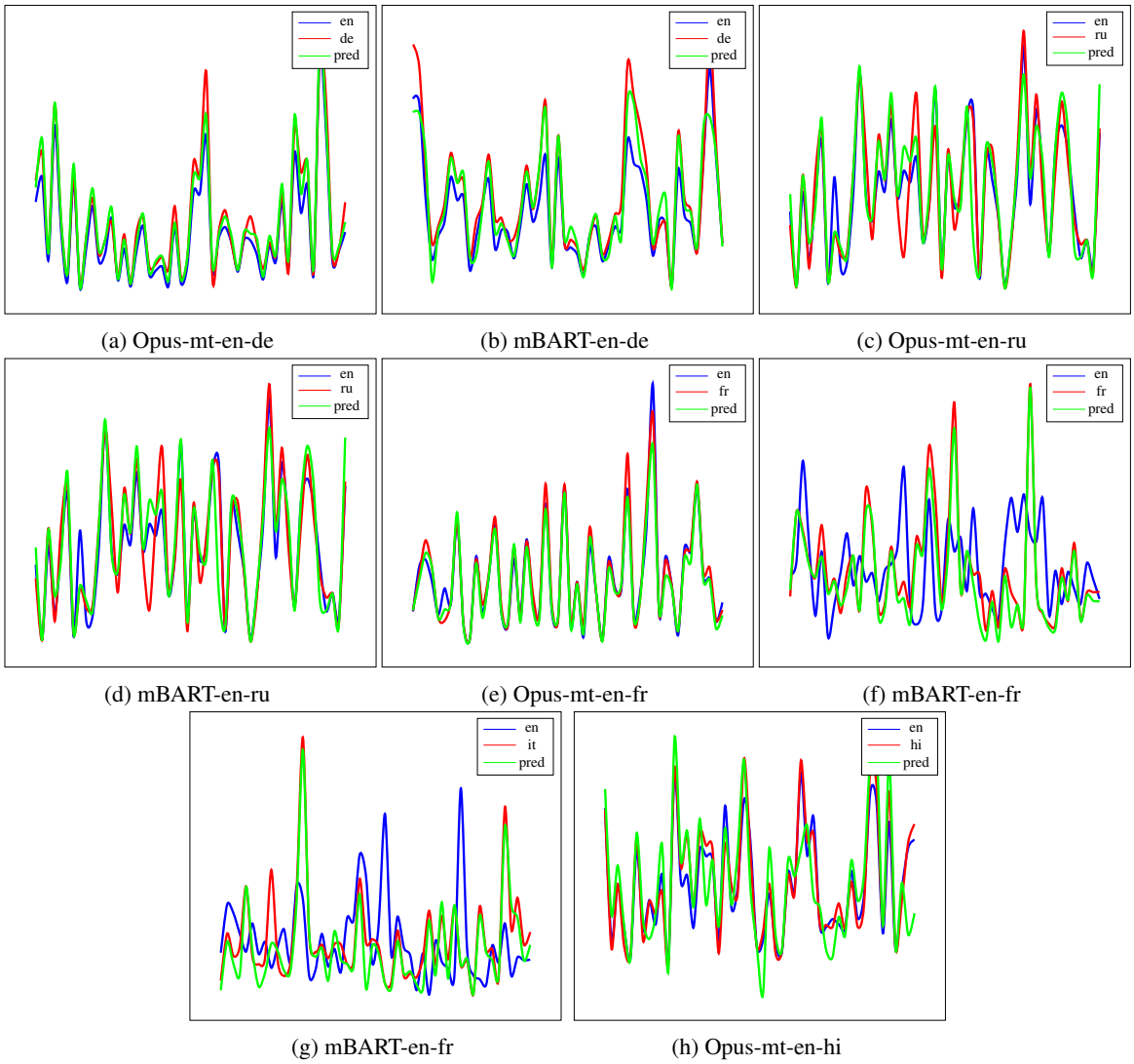
Figure 3: Comparison of time-duration across source, target and prediction for all of the aforementioned models as well as languages (de, fr, ru, it, hi) using Amazon Polly text-to-speech API. Here y-axis represents the duration of speech and *Blue*, *Red*, and *Green* lines show the time duration taken by the model for uttering the source text, target text, and generated isometric translated text.

tences. One possible reason can be that there can be a different ideal range of LR for Google's text-to-speech model to generate isometric outputs. Another reason can be that Google cloud uses wavenet-generated voices[3], which are trained using raw audio samples of actual humans speaking, which lead to a more human-like emphasis and inflection on syllables, phonemes, and words. On the contrary, the AWS Polly produces a more auto-tuned version of voices.

A point worth noting in figure 3 is that although OPUS-MT + mBART model of en-ru achieves the highest length range, the graph of fine-tuned OPUS-MT model seems more convincing and aligned with the source speech. From table 1 we can see that fine-tuned OPUS-MT of ru exhibits the Length Ratio of 1.005, which is extremely close to the ideal value 1. On the other hand, OPUS-MT + mBART achieves the LR of 0.967. After our analysis, we observe that although the LR of OPUS-MT + mBART falls within the isometric constraints, it is shortening most of the sentences.

In our analysis, we also found that the ideal LR can be changed based on the speed of the target language. For example, French is a faster-paced language than English, while German is slower than English. Therefore the value of the ideal LR for generating isometric outputs can be changed. As for a less number of characters, a faster language will match up with English while speaking.

[3] https://cloud.google.com/text-to-speech/docs/basics



(a) Opus-mt-en-de (b) mBART-en-de (c) Opus-mt-en-ru

(d) mBART-en-ru (e) Opus-mt-en-fr (f) mBART-en-fr

(g) mBART-en-fr (h) Opus-mt-en-hi

Figure 4: Comparison of time-duration across source, target and prediction for all of the mentioned models as well as languages (de,fr,ru,it,hi) using Google Cloud's text-to-speech API. Here y-axis represents the duration of speech and *Blue*, *Red*, and *Green* lines show the time duration taken by the model for uttering the source text, target text, and generated isometric translated text.

However, if the number of characters are more in a slower-paced language, it can cope with the English's speed

6 Conclusion & Future Work

In this work, we present a multilingual multitask learning system, which derives relations from the prompt-engineering technique, for fine-tuning the MT models as well as discuss the influence of reverse prompt engineering strategy, which can assist in paraphrasing text by utilizing the reverse prompts obtained using the target to the source character length ratio. We also present a comprehensive study for integrating several neural machine translation models with paraphrase models for source language translations with output length constraints. Additionally, We also investigate the application of a prompt-based few-shot learning technique for paraphrase models extended using the previously trained fine-tuned MT models. However, other enhancements may be incorporated to produce more optimal results. Firstly, there is a shortage of generalized isometric data, limiting the ability to evaluate the MT predictions for statistical metrics such as the BLEU score while also imposing significant constraints on training models for downstream isometric tasks. Next, Our research on the singleton system reveals that existing state-of-the-art multilingual models lack the ability to generalize to the use-case of multilingual MT tasks. Although, a generalization might be added

by inferring the MBart model, which can assess language distinction. However, the tokenization criteria, which prevents the simultaneous usage of many target languages, poses a barrier. This could be overcome by utilizing a reasonably distributive word tokenizer. Finally, a lot of improvements can be employed to the current output length control techniques. A combination of positional encoding via self-attention and prompt engineering technique could be employed to signify a more robust isometric MT. Additionally, We can even evaluate the systems predictability for various text-to-speech model thus creating a more diversified length compliance metrics enhanced for Automatic Dubbing.

7 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. Controlling the output length of neural machine translation. *CoRR*, abs/1910.10408, 2019.

[2] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *ACL*, 2022.

[3] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020.

[4] Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. *CoRR*, abs/1711.05217, 2017.

[5] Yashar Mehdad, Amanda Stent, Kapil Thadani, Dragomir Radev, Youssef Billawala, and Karolina Buchner. Extractive summarization under strict length constraints. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3089–3093, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[6] Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. Improved neural machine translation using side information. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 6–16, Dunedin, New Zealand, December 2018.

[7] Xing Niu and Marine Carpuat. Controlling neural machine translation formality with synthetic supervision. *CoRR*, abs/1911.08706, 2019.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540, 2016.

[10] Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. Lexically constrained neural machine translation with levenshtein transformer. *CoRR*, abs/2004.12681, 2020.

[12] Sho Takase and Naoaki Okazaki. Positional encoding to control output sequence length. *CoRR*, abs/1904.07418, 2019.

[13] Derek Tam, Surafel Melaku Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Prosody-aware neural machine translation for dubbing. *CoRR*, abs/2112.08548, 2021.

[14] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[15] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344, 2018.

[16] Chujie Zheng and Minlie Huang. Exploring prompt-based few-shot learning for grounded dialog generation, 2022.

[17] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[19] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT bombay english-hindi parallel corpus. *CoRR*, abs/1710.02855, 2017.

[20] Mathias Creutz. Open subtitles paraphrase corpus for six languages. *CoRR*, abs/1809.06142, 2018.

[21] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828, 2019.

[22] Yves Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association.

[23] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, 2018.

[24] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235, 2018.

[25] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[27] A. A. Vetrov and E. A. Gorn. A new approach to calculating bertscore for automatic assessment of translation quality. *CoRR*, abs/2203.05598, 2022.

[28] Surafel Melaku Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. Isometric MT: neural machine translation for automatic dubbing. *CoRR*, abs/2112.08682, 2021.