Hierarchical Multi-task learning framework for Isometric-Speech Language Translation

Aakash Bhatnagar and Nidhir Bhavsar
Navrachana University
Vadodara, India
(18124526,1803488)@nuv.ac.in

Muskaan Singh and Petr Motlicek
IDIAP Research Institute,
Martigny, Switzerland
(msingh,petr.motlicek)@idiap.ch

Abstract

This paper presents our submission for the shared task on isometric neural machine translation in International Conference on Spoken Language Translation (IWSLT). There are numerous state-of-art models for translation problems. However, these models lack any length constraint to produce short or long outputs from the source text. In this paper, we propose a hierarchical approach to generate isometric translation on MUST-C dataset, we achieve a BERTscore of 0.85, a length ratio of 1.087, a BLEU score of 42.3, and a length range of 51.03%. On the blind dataset provided by the task organizers, we obtain a BERTscore of 0.80, a length ratio of 1.10 and a length range of 47.5%. We have made our code public here https://github.com/aakash0017/Machine-Translation-ISWLT.

1 Introduction

The ability to reach a worldwide audience is a critical aspect of audio-visual content localization. This automation necessitates source language speech translation and seamless integration of target language speech with the original visual information. The uniqueness of this task is to generate length-controlled outputs. A significant application of isometric translation is in automatic dubbing, where the most crucial part is to sync the length of translated subtitles with the audio of the source language. These types of translations give a holistic experience to the user while reading the translated sentences. This paper will explain our hierarchical architecture for generating such isometric outputs.

Initially, we experimented with a verbosity-controlled multi-task model. We used two prompt types: (i) task prompt and (ii) length prompt. The task prompt decides what task the model should perform. For example, an empty prompt means that the model will receive English inputs and generate translated French outputs, whereas "para" prompt means that the model will receive French input and generate paraphrased French sentences. Para prompt always accompanies a length prompt that ensures that the paraphrased output is of the desired length. To illustrate, if the initial translated output of the model falls short of the source text, we will append the prompt: "para long." This prompt will help the model paraphrase this generated output to an optimal length. We experimented with various combinations of this translate-paraphrasing approach. Finally, our two best architectures consist of two/three separately trained models for translation and paraphrasing. We have used Helsinki OPUS-MT and Google’s MT5 for machine translation & paraphrasing, respectively, while Google translation API for short-length sentences. We use MUST-C v1.2 FR and PAWS-X EN-FR datasets to train these models.

2 Shared Task Overview

This task entails creating translations that are similar in length to the source. The shared task’s outcome can help with the following issues: auto standardized dubbing to achieve coupling between the source and target speech, improved subtitling to fit the translated content into a specified video frame, layout constrained translation to control the generated text to fit in the document tables or database fields, and more general simultaneous speech translation for ease of reading or listening. Participants in the shared task can create text-to-text MT systems for languages such as German (De), French...
Background

Our approach towards controlling the output length of translated sequences is based on the recent advancement in the transformer architecture (16) towards multi-task training.

3.1 Transformer

With the advent of transfer learning techniques in NLP through transformer-based models like T5 (11) have become more unified & can convert all text-based language problems into text-to-text formats. Trained on Datasets like C4, these models have achieved state-of-the-art performances for text generation tasks like summarization, question-answering & machine translation, to be precise. At its core, these models constitute a sequence-to-sequence architecture that can process sequences using only attention & feed-forward networks—partitioned into Block of Encoders and Decoder, each of which comprises multi-headed attention.

3.2 Few shot learning

As described in Brown et al. (2), fine-tuning a model for machine translation using a pre-trained model has been the most common approach in recent years, which involves updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used. The main disadvantages are the need for a new giant dataset for every task, the potential for poor generalization out-of-distribution, and the potential to exploit spurious features of the training data, potentially resulting in an unfair comparison with human performance. However, on the contrary, few-shot learning refers to the setting where the model is given a few demonstrations of the task at inference time. This works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion.

4 System Overview

In this section, we will explain our architecture in detail. As mentioned in the above sections, we implement a hierarchical architecture consisting of 3 separate modules. Our model is a complex fusion of two distinct functionalities, resulting in a differentiated pipeline that adds to improved performance for text generation tasks. The entirety of the model is fragmented into neural machine translation and a text paraphrasing system. While the former converts text from the source (En) to target (Fr) language, the latter, which is trained independently of the NMT model, assists in deforming the generated text into a more useful form specific to the task. Additionally, we are also using Google’s translation API for short-length sentences.

4.1 Translation Module

This module uses the state-of-the-art transformer-based neural machine translation model Helsinki OPUS-MT (15). The model is pre-trained using the MarianMT framework (5), a stable production-ready NMT toolbox with efficient training and decoding capabilities, and is trained on freely available parallel corpora collected in the large bitext repository OPUS (14). The pre-trained version of the OPUS-MT model has six self-attentive layers in both the encoder and decoder networks and eight attention heads in each layer. We use verbosity control during fine-tuning. While training, we use three prompts: "long," "short," and "normal." These prompts are defined by the Length-Ratio (LR) between the source and target texts. These prompts are appended to the input text, thus, allowing the model to recognize and differentiate key attributes governed by the Length Compliance (LC) matrix. The exact range of the ideal LR ratio is mentioned in the equation 1.

\[
  f(x) = \begin{cases} 
  \text{short}, & LR < 0.95 \\
  \text{normal}, & 0.95 \leq LR \leq 1.10 \\
  \text{long}, & LR > 1.10 
  \end{cases}
\]

(1)

\[
  f'(x) = \begin{cases} 
  \text{para long}, & LR < 0.95 \\
  \text{para short}, & LR > 1.10 
  \end{cases}
\]

(2)

We experimented the OPUS-MT model on two different datasets: WMT (1) and MUST-C (4). After experimentation, we decided to use MUST-C as it gave the most optimal results. OPUS-MT model, however, does not have any length-control mechanism. To fine-tune the model for isometric translation, we use the previously mentioned
And that might seem a bit surprising, because my full-time work at the foundation is mostly about vaccines and seeds, about the things that we need to invent and deliver to help the poorest two billion live better lives.

Le climat se détériore, ce qui signifie qu’il y aura de nombreuses années où leurs cultures ne pousseront pas. Il y aura trop de pluie, ou pas assez de pluie.

So, the climate changes will be terrible for them.

Les changements climatiques seront terribles pour eux.

Table 1: Examples from MUST-C dataset. Here SL is source length, TL is target length and LR is length ratio that is calculated by TL/SL. Isometric sentences are those, whose LR ratio lies within 0.95-1.10.

<table>
<thead>
<tr>
<th>Source Text (EN)</th>
<th>Target Text (FR)</th>
<th>SL</th>
<th>TL</th>
<th>LR</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>And that might seem a bit surprising, because my full-time work at the foundation is mostly about vaccines and seeds, about the things that we need to invent and deliver to help the poorest two billion live better lives.</td>
<td>Et cela peut sembler un peu surprenant parce que mon travail à temps plein à la Fondation concerne plutôt les vaccins et les semences, les choses que nous devons inventer et distribuer pour aider les deux milliards des plus pauvres à vivre mieux.</td>
<td>226</td>
<td>256</td>
<td>1.13274</td>
<td>Not Isometric</td>
</tr>
<tr>
<td>The climate getting worse means that many years, their crops won’t grow: there will be too much rain, not enough rain; things will change in ways their fragile environment simply can’t support.</td>
<td>Le climat se détériore, ce qui signifie qu’il y aura de nombreuses années où leurs cultures ne pousseront pas. Il y aura trop de pluie, ou pas assez de pluie.</td>
<td>199</td>
<td>162</td>
<td>0.8140</td>
<td>Not Isometric</td>
</tr>
<tr>
<td>So, the climate changes will be terrible for them.</td>
<td>Les changements climatiques seront terribles pour eux.</td>
<td>50</td>
<td>54</td>
<td>1.08</td>
<td>Isometric</td>
</tr>
</tbody>
</table>

Figure 1: Architectural representation of the flow of our pipeline. The first block in the figure represents the OPUS-MT model that we use for EN-FR translation. The right part in the diagram showcase the 2 paraphrasing models used: Google MT5 fine tuned and Google Translation API. Based on the condition we decide which model to use after translation.

4.2 Paraphrasing & Length Correction

According to Zhao et al. (21) the main goal of sentence paraphrasing is to improve the clarity of a sentence by using different wording that conveys the same meaning. For this task, we are fine-tuning Google’s MT5 model (18) on PAWS-X French dataset (19) to leverage the functionality of Text paraphrasing. We have fabricated the use of the prompt engineering approach (7) (12) to enable the model to recognize the paraphrasing task as well as modify its parameter based on the argument to generate isometric text. Manually engineered prompts are appended during training for both of the models, as mentioned earlier, based on the source and target text; however, during testing, the prompt for each input sentence is modified based on the conditional task of isometric text generation (see Figure 2).

5 Experimental Setup

During the experimentation, we used three datasets: 1) WMT, 2) MUST-C 3) PAWS-X. Table 3 shows the exact train/test/dev split of all the three datasets. Also, the task provides us with a blind dataset for each language pair. Particularly En-Fr pairs in the blind consisted of very few characters per sentence. After experimentation, we found that our model was not performing well for sentences with less than five words. To solve this issue, we used Google Translator API, which improved the length ratio and length constraint significantly.

We experimented with various approaches that involved multi-task training and hierarchical archi-
<table>
<thead>
<tr>
<th>Model</th>
<th>MUST-C Fr</th>
<th>Blind En-Fr</th>
<th>MUST-C Fr</th>
<th>Blind En-Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BERT Score</td>
<td>Length Compliance</td>
<td>BERT Score</td>
<td>Length Compliance</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>Length Ratio</td>
</tr>
<tr>
<td>System 1</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
<td>1.11</td>
</tr>
<tr>
<td>System 2</td>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
<td>1.08</td>
</tr>
<tr>
<td>System 3</td>
<td>0.86</td>
<td>0.85</td>
<td>0.85</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Table 2: prediction on MUST-C v1.2 En-Fr and blind dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MUST-C</th>
<th>PAWS-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>en-fr</td>
<td>fr-fr</td>
</tr>
<tr>
<td>Train</td>
<td>275086</td>
<td>49401</td>
</tr>
<tr>
<td>Validation</td>
<td>1413</td>
<td>2000</td>
</tr>
<tr>
<td>Test</td>
<td>2633</td>
<td>2000</td>
</tr>
</tbody>
</table>

Table 3: description of various datasets used during the experimentation.

Figure 2: Multi task model architecture of updating parameters according to the prompts supplied

5.1 Evaluation Measures

This task is evaluated on two parameters. The first is the quality of translation, and the second is the length constraint. We use BERTscore and BLEUscore for qualitative analysis of the translated sentences and Length Compliance matrix for the isometric constraint. Table 1 in appendix 7 shows a detailed overview of how Length Compliance matrix works. We can see that the optimal predictions lie within the LR range of 0.95 and 1.10.

6 Result and Analysis

As shown in Table 2, system three has gained a substantial increase in overall Length compliance metrics. However, the BERT Score has depleted by a factor of 0.5. The Length Ratio for the OPUS-MT system is 1.085, close to the ideal value in isometric translation. The isometric translation aims to generate the length ratio between 0.95 and 1.10, i.e., considering the +10% shift in the characters. We can achieve it through two systems, with system-

---

1https://simpletransformers.ai/
2https://github.com/huggingface/transformers
Algorithm 1 Algorithm for our pipeline

1. Variables
   - $S$ Source text [train]
   - $T$ Target text [train]
   - $S_t$ Source text [test]

2. Pre-Processing
   - **procedure** GENERATE-LENGTH-PROMPT($S, T$)
     - for $i \leftarrow 1$ to $S$
       - prompt $\leftarrow f(S, T)$
       - $S'_i \leftarrow$ prompt $+ S_i$
     - end for
   - end procedure
   - $S'_t \leftarrow$ normal $+ S_t$  \(\triangleright\) process test-data

3. Neural Machine Translation
   - **procedure** TRAIN-MT-MODEL($S'_t, T$)
     - input-ids, attention-mask, labels $\leftarrow$ Tokenizer
     - translation-model $\leftarrow$ Model("OPUS-MT-en-fr")
     - loss-function $\leftarrow$ criterion() \(\triangleright\) cross entropy loss
     - translation-model.train(input-ids, attention-mask, labels, loss-function)
   - end procedure
   - $T_p \leftarrow$ translation-model.predict($S'_t$) \(\triangleright\) follow step 3

4. Text Paraphrasing
   - Train MT5 model on PAWS-X dataset
   - **procedure** GENERATE-TASK-PROMPT
     - for $i \leftarrow 1$ to $S'_t$
       - prompt $\leftarrow f(S'_t, T_p)$ \(\triangleright\) Eq. 1
       - if prompt $\neq$ normal then
         - para_prompt $\leftarrow f'(S'_t, T_p)$ \(\triangleright\) Eq. 2
         - $T'_p \leftarrow$ para_prompt $+ T_p$
       - else continue
     - end if
   - end for
   - end procedure
   - $O \leftarrow$ paraphrase-model.predict($T'_p$) \(\triangleright\) final output

1 achieving a length ratio of 0.85 and system-2 achieving 0.87.

Secondly, the length range represents the percentage of total translated sentences falling under the ideal length ratios. Two of our suggested models are close to 50%, suggesting that almost half of the predictions are isometric with high BLEUscore and BERTscore. The decrease in the BERTscore of system 3 is that the model loses essential information while predicting the output. From various examples, we can see that verbosity control can sometimes lead to abrupt shortening of results, where the model skips words after a specific limit.

Along with length compliance metrics, outputs are evaluated for their adequacy and quality of translation. This task emphasizes more towards BERTscore rather than BLEUscore. When the length of source and target varies, BLEUscore does not adapt well; however, BERTscore can evaluate based on semantics. The challenge is to translate the source text to the target language with ideal length compliance while also maintaining the semantic meaning of the output.

While our suggested models are also perform-
ing equally well on the blind dataset provided by the organizer, however, a significant dip can be seen with the Length ratio & BERT score for the predicted outputs. The reason being is that the blind data covers a versatile range of source input with a word count ranging from 1 to 44. A significant issue in our implementation of system-1 and system-2 is that the PAWS dataset has an average length of 10–15 words and cannot provide a range of training examples with a short total token/word count. Thus, while predicting the model performs rather poorly for short-length examples, we have employed Google Translate API. However, for some instances within the 5-8 word count, the model can still not convert the input sequence to its target language ("French") counterpart.

Our experiments with the Google MT5 model, which is fine-tuned for machine translation and text paraphrasing, have shown considerable promise. However, it still needs rigorous experimentation and hyper-parameter tuning. In addition to quantitative, we vouch for qualitative analysis of our results in Table: 4. The table 4 describes the correct output corresponding to isometric source-target text. As shown in the fourth row of the Table, our system can precisely shorten the length of translated text while retaining semantical similarity. Secondly, as set out in the second and third row of the Table, few phrases in the English & French vocabulary do not align lexically together; thus, the model partitions the source text and translates each word separately.

7 Conclusion & Future Work

In this work, we propose a hierarchical MT approach, using prompt engineering to attribute the OPUS-MT and MT5 paraphrasing model. We evaluate the proposed approach in the Isometric machine translation case, where translated text is expected to match the source length to synchronize the source and target text. Our finding shows that though the model has been trained precisely for generating constrained output, However, a lot of improvements can be employed to produce more optimal results. Firstly, the paraphrasing model could not generalize for short sentences (i.e., LR < 0.95). Secondly, the MUST-C dataset has an unequal distribution of instances for all three categories of length ranges, which imposes an uncertain suspicion over the model predictions. Moreover, our finding shows that the proposed approach can perform better than Lakew et al. (6), length aware positional encoding based NMT approach.

References


I just came back from a community that holds the secret to human survival.

Je viens de revenir d’une communauté qui détient le secret de la survie de l’humanité.

Le gentil geste qu’elle a remarqué parmi tous les autres : quelqu’un lui avait même amené une paire de chaussures.

Si vous avez quelque chose à donner, donnez-le maintenant.

Servez de la nourriture dans une soupe populaire, nettoyez un parc dans votre quartier, soyez un mentor.

Vouici le monde des bonobos sauvages dans la jungle du Congo.

This is the world of wild bonobos in the jungles of Congo.

Table 4: Predicted Results from MUST-C dataset. Here SL is source length, TL is target length, PL is predicted length and LR is length ratio that is calculated by PL/SL. Isometric sentences are those, whose LR ratio lies within 0.95-1.10

---


