SSMT

Machine Translation with/without Rich Information

Sumanth Balaji - 2018114002
Monil Gokani - 2018114001

Introduction

The aim of the project is to build, train and test machine translation with/without Rich information for spoken and written text. Here, “Rich information” includes:

- Punctuation
- Pauses, Filled Pauses and Pet phrase annotations

In this report we collate our findings from different experiments and analysis exploring how presence of Rich information affects machine translation. We also make our codebase [1] public for future work and exploration. The datasets we have used are linked on the codebase.

Literature review

Not much literature exists in the problem statement that we are exploring. Of the few literature that is available, Using Prosody for Automatic Sentence Segmentation of Multi-Party Meetings [2] (which explored leveraging prosody for Chinese-English translation) was explored, but was not found to be very relevant to our project. A useful literature reference was Enriching machine-mediated speech-to-speech translation using contextual information [6] which explored integration of dialog act tags in statistical translation frameworks.

Datasets

We have primarily used two datasets, and processed them to produce 4 different sets for our purposes. The primary datasets used are:

1. IITB En-Hi parallel dataset: This dataset[5] is published by IIT Bombay. It contains approx. 1.6M sentence pairs between English and Hindi. We have used this to create two corpora, described below.
2. Swayam Dataset with speech annotations: This dataset was provided to us, structured as parallel transcripts of lectures from 3 topics. We processed and compiled into a single set of parallel files.

We created 4 datasets, which are described below.

1. **Raw Dataset <iitb_norm>:** We cleaned and filtered out undesirable instances from the IITB dataset to form our primary training dataset. Some of the filtering we did was to remove instances that were too short (< 7 words), and normalise illegal/unwanted characters (converting everything to utf-8 encoding). Statistics from the dataset:
    - Total instances after filtering: ~950k
    - Average instance length: ~21 tokens
    - Splits: 60-20-20 train-dev-test
2. **Punctuation and Case free dataset <iitb_nopunc>**: We processed the iitb_norm dataset and removed all punctuation from both sides of the dataset. Additionally, we lowercased the English sentences to remove capitalisation information. The statistics and splits for this dataset are identical to the iitb_norm dataset.
3. **Rich text annotated dataset from Swayam <swayam_rt>**: The swayam dataset was obtained in the form of paragraph-aligned parallel transcripts for lectures from 3 modules: AI, Biomolecular Structure Functions in Health and DIsease, and FMFS. We combined the individual files into a single paragraph level aligned dataset. This dataset contains annotations for speech in the English transcript. The full annotation guidelines can be found in the github repository. The most frequent labels are shown in the table.

    Important Rich text labels:

<HES></HES>	Used to mark hesitation in speech, like utterances such as *uhh* or *err*

<lp>	Denotes a pause of 1-2 seconds in the utterance
<AB></AB>	Marks abbreviations, such as *<AB> DNA </AB>*
<FW></FW>	For Foreign Words, such as *dahi*
<PET></PET>	To mark pet/filler phrases that are often used while speaking such as: *We have finished the third module <PET>right</PET> and we start the new chapter.*

Some statistics in the final dataset:

- Total instances after filtering: 2774
- Average instance length: ~89 tokens
- Splits - 80-10-10 train-dev-test
  - 2219 train instances
  - 278 dev instances
  - 277 test instances.

4. **Manually created sentence level test set from swayam_rt <man_test>:** Since there was a significant difference in the length of the input sequence in the swayam_rt set and the iitb based sets, it was difficult to meaningfully compare the output of the different models on the various test sets. Hence, we manually created a small test from the test set of swayam_rt that was aligned at the sentence level. We then processed these sentences to remove annotations and punctuations and create 3 input files and one reference file, each containing **117** instances:
   - **en.txt <en>**: this set contains sentences without any swayam_rt annotations, but retains punctuation and capitalisation information.
   - **en_nopunc.txt <en_nopunc>:** The same set of sentences, but lowercased and all punctuation removed.
   - **en_rt.txt <en_rt>:** the original sentences containing all the rich text annotations and punctuation information.
   - **hi.txt** : reference file for the expected translations.

   The average sentence length in these files was reduced to approximately **24 tokens per instance**, which was comparable to the length that the model was trained on,

and hence the outputs on these files were used for our analysis. The table below contains a sample sentence from each file:

Sample sentences from each file in <man_test>

en	So, you can evaporate the organic solvent by flushing uhh a stream of nitrogen gas into the tube.
en_nopunc	so you can evaporate the organic solvent by flushing uhh a stream of nitrogen gas into the tube
en_rt	So, you can evaporate the organic solvent by flushing <HES>uhh</HES> a stream of nitrogen gas into the tube.
hi	तो, आप ट्यूब में नाइट्रोजन गैस की एक धारा प्रवाहित करके कार्बनिक विलायक का वाष्पीकरण कर सकते हैं।

Models

We produce 3 models:

- **Raw <base>**:
  - This is a transformer trained on <iitb_norm> dataset.
- **NoPunc <nopunc>**:
  - This is a transformer trained on <iitb_nopunc> dataset, where punctuation information was removed
- **Rich Information <base + rt>**:
  - This is a transformer that is produced by fine tuning the existing *base* transformer on the Rich Information dataset.

The models were trained with inspiration from transformer training scripts by Ai4Bharat indicTrans[2] and makes use of the fairseq seq2seq toolkit[3] . Data for the models training is preprocessed using BPE. The hyperparameters for training is shown in table below:

Parameter	Value
arch	transformer
optimizer	Adam (betas: 0.9 and 0.98)

clip-norm	1.0
Learning rate	5e-4
Learning rate scheduler	inverse_sqrt
Warmup initiate learning rate	1e-7
dropout	0.3
Weight decay	0.0001
Label smoothing	0.1
warmup-updates	4000

Experiments and Results

For experimentation, we fed each of the three input files in the *man_test* dataset to the three models that we had trained, to obtain a set of 9 outputs. We obtained **BLEU** scores for each of these using the sacrebleu[7] module. The scores can be seen in the following table:

Input⇒ Model⇓	en	en_nopunc	en_rt
base	21.0888	19.0347	18.3616
nopunc	13.7366	18.5659	13.0138
base + rt	**23.4013**	**21.2971**	**22.6571**

The scores, though indicative of overall model performance, do not give a very accurate picture of the actual performance of the model. For that purpose, we have qualitatively analysed instances from the outputs, which are discussed in the next section. However, based on the scores themselves, we can the following observations:

1. The *nopunc* model performs decisively worse on all inputs than the other two models. This is along expected lines, since punctuations and capitalisation is used to encode a lot of information in the English language. The *nopunc* model therefore misses out on a lot of this information and is not able to produce satisfactory

output, especially on the *en* and *en_rt* sets, since the added punctuations and annotations only act as noise to the model.

2. The *base* + *rt* model performs best on all three inputs. We believe this is because the *swayam_rt* set used to finetune it has significantly longer sequences in the input, and therefore the model performs better on longer sequences than the *base* model. (illustrated in next section).
3. The *base* + *rt* model scores lesser with the annotations in the input as compared to without. We believe this is simply the case of a lack of training data, since as we show in next section, it does manage to capture the more frequent tags well, but messes up when it encounters unfamiliar or an excessive number of tags. A larger training set with more instances of each tag would help mitigate this problem.

Qualitative Analysis

This section contains some examples illustrating observations made while going through outputs, especially those of *base* and *base+rt* on *en* and *en_rt*. (left entry is of the form <model>_pred_<inp>, i.e. prediction of <model> on input from file <inp>)

THE GOOD

1. Model generalises better on longer sentences after finetuning on *swayam_rt*. The *base* model here misses out on some of the context that the *base+rt* model is able to pick up.

Input: (en)	Once the lipid is extracted from a cell or a tissue uhh of an organism or from food samples, the weight of the lipid is measured by either weighing the remaining sample after the extraction with organic solvents or weighing the extracted lipid sample after evaporating the organic solvent.
base_pred_en	कार्बनिक विलायक को वाष्पित करने के बाद निष्कर्षण के बाद के शेष नमूने का भार या तो वियोचित सॉल्वैंट्स के साथ वज़न कर लिया जाता है।
base+rt_pred_en	लिपिड को एक कोशिका या किसी जीव के ऊतक यूएच से निकाला जाता है या खाद्य नमूनों से, लिपिड का वजन या तो कार्बनिक सॉल्वैंट्स के साथ निष्कर्षण के बाद के शेष नमूने को कम करके मापा जाता है या कार्बनिक विलायक को वाष्पन करने के बाद निकाला गया लिपिड के नमूने को वज़न कर दिया जाता है।

2. *base+rt* manages to retain foreign words thanks to the <FW> tag in the en_rt input, which was lost in the *base* model (it was replaced by a generic pronoun in the absence of the tag).

Input: (en_rt)	<lp> The oldest fermented milk of India, <FW>Dahi</FW>, still occupies a pivotal position not only in daily diet of people of this country but also in their customs and religious rituals.
base+rt_pred_en	यह न केवल देश के लोगों के दैनिक आहार में बल्कि अपने रीति-रिवाजों और धार्मिक अनुष्ठानों में भी महत्वपूर्ण स्थान रखता है।
base+rt_pred_en_rt	अब भी, दही इस देश के लोगों के दैनिक आहार में ही नहीं, बल्कि उनके रीति-रिवाजों और धार्मिक अनुष्ठानों में भी, एक महत्वपूर्ण स्थिति है।

3. Tags help to clarify certain words like 'uhh' which are otherwise recognised as acronyms by the model, and are thus transliterated instead.

Input: (en_rt)	So, in overall, <HES>uhh</HES> if you have <HES>uhh</HES> different methods, it is useful to get the <HES>uhh</HES> complex inform... <HES>uhh</HES> protein complexes and large <HES>umm</HES> assemblies, <HES>umm</HES> all these four methods are useful to get the information.
base+rt_pred_en	तो, कुल मिलाकर, यूएचएच में, यदि आपके पास विभिन्न विधियां हैं, तो यूएचएच कॉम्प्लेक्स को सूचित करना उपयोगी होगा..... प्रोटीन कॉम्प्लेक्स और बड़ी संख्या में, इन चार विधियां जानकारी प्राप्त करने के लिए उपयोगी हैं।
base+rt_pred_en_rt	कुल मिलाकर, यदि आपके पास विभिन्न विधियां हैं, तो यह जानकारी प्राप्त करने के लिए उपयोगी होता है।

THE BAD

4. Infrequent tags cause confusion in the model, such as the <*em*> tag in this case. The model predicts a better output without the tag in the input, than with it.

Input: (en_rt)	According to food regulatory authorities, twenty percent of carbon dioxide is sufficient to control the growth of microorganisms.
base+rt_pred_en	खाद्य नियामक प्राधिकरणों के लिए बीस प्रतिशत कार्बन डाईआक्साइड सूक्ष्मजीवों की वृद्धि को नियंत्रित करने के लिए पर्याप्त है।
base+rt_pred_en_rt	सूक्ष्मजीवों के विकास को नियंत्रित करने के लिए, कार्बन डाईआक्साइड का बीस प्रतिशत।
Hi (ref)	खाद्य नियामक अधिकारियों के अनुसार, कार्बन डाइऑक्साइड का बीस प्रतिशत सूक्ष्मजीवों की बढ़ोतरी को बाधित करने के लिये काफी है।

5. Domain specific words are hard to capture. Consistently replaced by similar words in the domain, possibly because the incorrect word had a lot more occurrences in the *iitb_norm* set that the model was trained on, than the correct one in the *swayam_rt* set that was used for finetuning. Consistently occuring across multiple examples.

Input: (en_rt)	Lexitropsines, lexitropsines are semi-synthetic ligands.
base+rt_pred_en	अन्तःएक्ट्रोपीट्रान, लेक्ट्रोपी, अर्धसिंथेटिक लिग्डस हैं।
base+rt_pred_en_rt	एक्ट्रोट्रोफिन, लेक्ट्रोफिन अर्धसिंथेटिक लिग्डस होते हैं।
Hi (ref)	लेक्सिट्रोप्सिन - लेक्सिट्रोप्सिन अर्ध-कृत्रिम लिगैंड्स होते हैं।

THE UGLY

6. Due to a lack of training data, the model fails utterly when too many RT tags are present in the input, to the point that the output is entirely non-sensical.

Input: (en_rt)	<FW>Dahi</FW> also serves as a raw material for preparation of related products such as <FW>lassi</FW>, <FW>chhash</FW>, <FW>shrikhand</FW>, etcetera.
base_pred_en	चिश्ती से संबंधित उत्पादों जैसे लैस्सी, चिश, झारखंड आदि की तैयारी के लिए कच्चे माल के

	रूप में भी कार्य करती है।
base+rt_pred_en	साथ ही, लैस्सी, चिश, श्रीखंड, आदि संबंधित उत्पादों की तैयारी के लिए कच्चे माल के रूप में भी कार्य करता है।
base+rt_pred_en_rt	इसके अलावा, संबंधित उत्पादों के लिए एक कच्चे माल के रूप में भी कार्य करता है, जैसे कि <whai> lai> lai </daint> chah> chah </deskht>, <haa> sauk> sauthor </deski>, sauthesko saom.
Hi (ref)	संबंधित उत्पादों जैसे लस्सी, छाछ, श्रीखंड आदि बनाने के लिए दही कच्चे माल के रूप में भी सहायक होता है।

7. Regardless of annotations, when the spoken form is highly irregular compared to standard text input such as having multiple grammatical inaccuracies, it is not possible to accurately translate it automatically.

Input: (en)	The same way doctor also, doctor would say that you won't eat hundreds of items and the patient will say no, I cannot live with, if you you just ban me.
Input: (en_rt)	The same way doctor also, doctor would say that you won't eat<lp> hundreds of items and the patient will say no, I cannot live with, if you you just ban me<lp>.
base_pred_en	इसी तरह डाक्टर भी कहते थे कि आप सैंकड़ों चीजें नहीं खायेंगे और रोगी नहीं कहेगा, अगर आप मुझ पर प्रतिबंध लगा रहे हैं, तो नौसिखिया के साथ नहीं रह सकता।
base+rt_pred_en	इसी तरह डाक्टर भी कहते थे कि आप सैंकड़ों चीजें नहीं खायेंगे और रोगी, नहीं, अगर आप मुझ पर प्रतिबंध लगाते हैं, तो वह मेरे साथ नहीं रह सकता।
base+rt_pred_en_rt	इसी तरह डाक्टर भी कहते हैं कि आप नहीं खा सकते हैं और रोगी नहीं, अगर आप सिर्फ मुझे प्रतिबंधित कर देते हैं, तो नहीं, साथ नहीं रह सकता।
Hi (ref)	उसी तरह जब डॉक्टर मरीज से कहते हैं कि आप ये सैकड़ों चीज़ें नहीं खा सकते और मरीज कहता है कि नहीं, अगर आप मेरे खाने की चीजों पर प्रतिबंध लगाते हैं, तो मैं नहीं जी सकता।

Future work

There is potential to extend this project by increasing the annotated dataset, perhaps by exploring ways to automatically annotate features such as hesitation and pet phrases.

Additionally, for text-to-text MT, it might be worthwhile to explore finetuning large models trained on traditional sentence aligned datasets with datasets that are aligned at a higher level, paragraph or perhaps even document level, to improve the model performance on longer sequences.

References

1. https://github.com/sumba101/MachineTranslation
2. https://link.springer.com/chapter/10.1007/11846406_79
3. https://github.com/AI4Bharat/indicTrans
4. https://github.com/facebookresearch/fairseq
5. https://www.cfilt.iitb.ac.in/iitb_parallel/
6. https://sail.usc.edu/publications/files/1-s2.0-S0885230811000428-main.pdf
7. https://pypi.org/project/sacrebleu/