Spanish to English Translation in Patient-Facing Medical Settings

University of California, Berkeley
School of Information

Hassan Saad
hassansaadca@berkeley.edu

Courtney Smith
smith.courtney@berkeley.edu

Grant Wilson
grant.wilson@berkeley.edu

Abstract

We address the process of building a model to translate Spanish to English in the biomedical domain. We start with an existing model sourced from HuggingFace and we fine-tune on a text corpus from the Institute of Formal and Applied Linguistics (UFAL).[1] We report the evaluation metrics both in terms of the corpus BLEU scores and a comparison to translations completed by a professional Spanish-English medical interpreter.

1 Introduction

There are many people in the United States for whom English is a second language, which may hinder their ability to get access to essential services like medical care. In addition to difficulties inherent in interpreting everyday language, industry-specific terminology adds another degree of complexity between native English and native Spanish speakers.

While medical interpreters do exist to assist with these tasks, they are usually short staffed, and their time is best used in live (active) translation scenarios rather than document-based (passive) translation tasks.

Our model will allow for the full translation from Spanish to English in the written form to assist with the passive element.

We envision a tool that will allow Spanish-speaking patients or family members to type in a phrase which can then be interpreted to communicate with a medical staff member.

2 Related Work

Earlier work in this area focused on a word- or phrase-based statistical machine translation approach that leveraged parallel terminologies. In recent years, methods have shifted towards neural machine translation (NMT) which is better suited for longer sentences or paragraphs.[2] Previous work training models primarily on in-domain texts for specific use cases have led to poor performance when the model comes across out-of-domain content.[3] For this reason, we chose to leverage the UFAL corpus to fine-tune a Marian model that had been pre-trained extensively on general, "everyday" texts from the OPUS corpus in several different languages.

Marian is an open-source model for NMT developed at the Uni-

versity of Edinburgh in collaboration with Microsoft. While originally built in C++, Natural Language researchers at the University of Helsinki developed transformers for several different language combinations and open-sourced these via HuggingFace.[4] The MarianMT model we used as a baseline was specifically trained for Spanish to English translation. Few (if any) of the training corpora included text from the medical domain.

3 Data

The UFAL corpus provides access to a 13.1 GB dataset of text pairs between English and several other languages (e.g. French, Spanish, Hungarian, Romanian, etc.). Over 430M text pairs exist across these languages, and roughly 10M of them consist of text that was extracted from medical documents. Because the MarianMT model is already trained for general translation tasks, we focused on using just the medical-related data in our fine-tuning stage. This yields 791,000 tab-separated text pairs specific to medical Spanish language out of a total 91M Spanish/English text pairs.

Corpora	es-en
ECDC	2,357
EMEA (OpenSubtitles)	487,901
EMEA (new crawl)	-
Medical Web Crawl	148,982
Subtitles	151,675
Parallel Segments	**790,915**
Total Words (es/en)	**9M/10M**

Table 1: Breakdown of dataset sources

4 Implementation

Data

The texts collected from medical resources included tags for the source and the specific corpus, either *medical_corpus* or *general_corpus*. Only lines with the *medical_corpus* tag were used for fine-tuning. We started by first filtering the 91M sentence pairs by corpus tag and writing our samples to a text file. After this step was completed, we separated the data into train and test sets with a 90%/10% split, which yielded 79,091 pairs in our test set and 711,824 in our train/validation set.

Limited pre-processing was done on the data. We removed from our training data 2,800 samples for which the source text was blank and 3,000 samples for which the target text was blank. Due to the size of the dataset, it was not feasible to confirm the correctness/validity of each translated sentence pair.

There is limited documentation related to the MarianMT model set that uses the TensorFlow API, so we implemented PyTorch to create our own training script. In addition, most of the documentation entails the use of pre-loaded datasets which are in a convenient format for feeding into sequence to sequence models, and that do not reflect practical data source formats. We therefore had to build a child class from the torch.utils.data.Dataset object in order to extract the relevant portions of the English and Spanish text from our data file and combine them into one tokenized sample. For each sample, we created a dictionary consisting of 3 tensors: the `input_ids`

and `attention_mask` from the Spanish tokenization, and the `input_ids` from the English tokenization, which we renamed `label`.

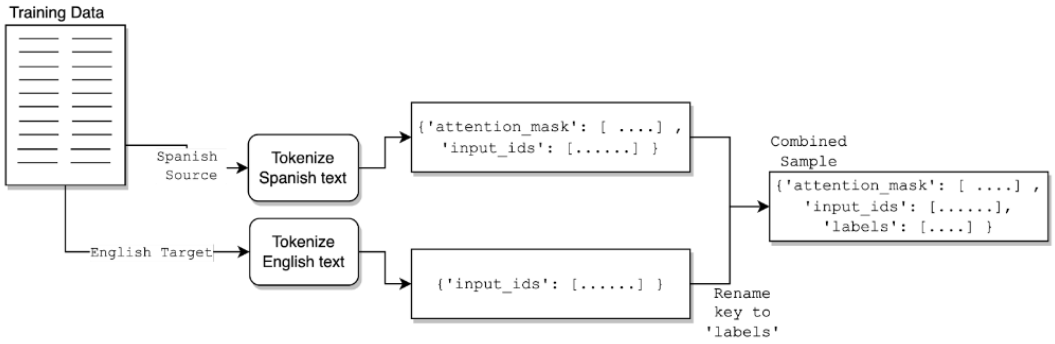

Figure 1: Implementation of PyTorch Dataset object to create training samples for each line of text

Model

The MarianMT transformer consists of an encoder-decoder stack with 6 layers in each component. Similar to T5, it uses a self attention layer on the encoder side and a masked self-attention layer on the decoder side to act as an autoregressive language model when converting from embeddings to text. Where T5 is trained on various different tasks including question-answering, translation, etc., our fine-tuning only focused on a translation-based training strategy.

We chose to use MarianMT over T5 due to the specific use case with which it was originally trained. Rather than using a more general model like T5 that allows for translations between several languages, we selected one of the many available MarianMT models that was specific to Spanish-English translation.

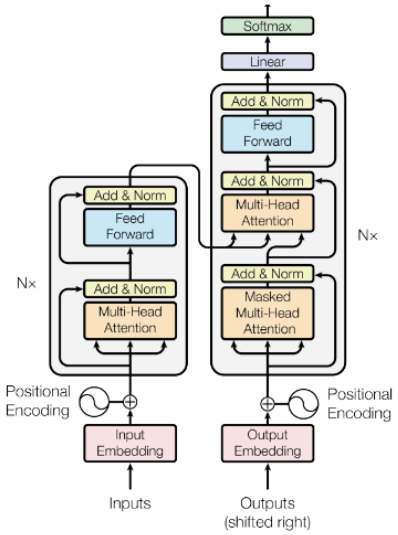

Figure 2: Transformer Architecture

Due to the large number of samples in our dataset, we limited the number of training epochs to 1. We

implemented a batch size of 8 for both training and evaluation, with a learning rate of 2e-5. For the first trial run of training, we applied a patience factor of 10, however training still completed across the entire epoch. This patience parameter was therefore removed in the final training iteration.

Lastly, we did not make any additions to the model vocabulary during tuning. Any rare medical terminology not in the baseline model vocabulary was therefore tokenized in the same manner when running inference with the fine-tuned model.

Given the size of the training data, we utilized an Amazon Web Services "g4dn.4xlarge" GPU-enabled compute instance to conduct the training for the PyTorch model. Our training time took 2 hours and 47 minutes while evaluation took 5 hours and 22 minutes. During the training process, we saved checkpoints every 50 samples to use for model testing and validating our BLEU scores between baseline and fine-tuned models.

Model Evaluation Strategy
Our plan for evaluating the translations consists of two steps:

1. Using the BLEU metric for both the baseline and fine-tuned MarianMT models across the entire test corpus

2. Providing several samples to a professional Spanish/English medical interpreter and comparing results (sentence-level BLEU scores) for the corresponding baseline and fine-tuned translations

5 Results

Overall BLEU Scores
The trained model provides a noticeable increase in the BLEU score relative to the baseline model when applied to the test set of 79,000 samples. The baseline MarianMT model already performs quite well, achieving a BLEU score of 52.39 across the test data, while our fine-tuned model improves on this with a BLEU score of 56.43.

Model	BLEU Score (Test Corpus)
Baseline	52.3942
Fine-Tuned	56.4263

Table 2: Comparison of BLEU Scores

While these results are promising, we wanted to use an additional strategy aside from the BLEU score to understand the interpretability and fluidity of our translations. We therefore reached out to a professional medical interpreter at Stanford Hospital who specializes in Spanish/English translations.

Professional Interpretation of Subset of Test Samples
Because creating written translations is a lengthy process, we limited this sample size to 20 sentences. We did not pick these samples randomly due to the fact that some of the data (even in the medical corpus) consists of phrases that don't contain very technical medical terminology. To confirm model translations of medical text were successful, we selected relatively technical sentence pairs by hand for this portion of the evaluation, ensuring that they were complex enough to create variability between translations.

It's important to note that our professional interpreter was asked to limit how much the translations were restructured, while at the same time retaining the fluidity that is inherent to the English language. For example, one target translation in our corpus is:

"*Talk to your doctor about the possible risks of using this medication for your condition.*"

Our interpreter's translation of the original Spanish sentence is as follows:

"*Ask your doctor about possible risks associated with using this medication to treat your condition.*"

However, in a practical scenario, she would have omitted the last part of this sentence on the basis of it being implied, and she would have provided the following more concise translation:

"*Ask your doctor about possible risks associated with using this medication.*"

The interpreter noted that the English language in general requires the use of fewer words than Spanish. In addition, the context of the translation may affect whether to prioritize fluency over an exact translation. An official document may require a more literal translation, where an email or note could place more value on fluency, brevity, and a more natural English phrasing.

After getting these results from our interpreter, we also wanted to compare the translations to the baseline and tuned model results on the same 20 samples. We were pleased to see that our fine-tuned model achieved a higher BLEU score than the baseline for these selected sentences. In addition, we ran the source text samples through Google Translate, and our tuned model's results were better than Google's output as well. We provide a table in the appendix indicating the score for each model (and the interpreter's output) relative to the target translation.

One of the more interesting examples we found relates to the following target and output translations (sample 12 in the table shown in the appendix):

Target: "*If you stop taking CYMBALTA Do not stop taking your capsules without the advice of your doctor even if you feel better.*"

Baseline: "*If you stop taking CYMBALTA Do not stop taking your capsules without your doctor's advice even if you are feeling better.*"

Fine-Tuned: "*If you stop taking CYMBALTA Do not stop taking your capsules without the advice of your doctor even if you feel better.*"

Interpreter: "*If you decide you are ready to stop taking CYMBALTA, do not stop taking your capsules without consulting your doctor, even if you feel better.*"

We note that the fluidity and logic of the professional interpretation is likely the best out of all examples. However, compared to the baseline and fine-tuned models, it

performs very poorly as measured by the BLEU score:

Model	BLEU Score
Baseline	59.33
Fine-Tuned	100.00
Human Interpreter	41.03

Table 3: Comparison of BLEU Scores: Models to Human Interpretation

In addition, the high BLEU score for the fine-tuned model output made us concerned regarding data leakage. After tuning, the model opts for the use of "without the advice of your doctor" rather than "without your doctor's advice". After examining our data further, however, we confirmed there was no leakage of the specific sentence pair between the train and test sets. We observed that the train dataset contains 13 instances of the phrase "without the advice of your doctor" and only 1 of "without your doctor's advice" in the targets. Since we're clearly creating our new sentence based on similar phrases in the training set, we'll want to be careful that this isn't evidence of overfitting. The term frequency relative to the size of the training set (less than .0001%) suggests that this may just be a rare term, but we should still give attention to use cases like this that may create issues in model generalizability. Given the data size and BLEU score, future work on this use case should focus on collecting more data and being conscious of potential overfitting.

6 Conclusion

Translating between Spanish and English is a difficult enough task without considering the context of medical terminology. To ensure both patient safety and proper access to healthcare, we have to be mindful of those difficulties as they could literally be the difference between life and death. Our Fine-Tuned MarianMT ES-to-EN model lays out an approach for significantly reducing or eliminating those difficulties by fine-tuning our translation models for medical contexts that can conduct translations without requiring the availability of a human translator.

7 Appendix

Small-sample translations and comparisons across baseline model, fine-tuned model, Google translate, and human interpreter. Example translations can be found here.

Sentence-level BLEU Scores:

Sentence Number	Baseline MarianMT	Fine-Tuned MarianMT	Google Translate	Interpreter
1	39.33	53.52	41.88	29.94
2	53.73	53.73	53.73	53.73
3	100.00	100.00	100.00	64.35
4	38.30	51.56	61.43	52.26
5	40.13	38.16	48.30	33.97
6	61.48	86.12	68.65	23.18
7	48.89	48.89	42.38	37.68
8	79.11	79.11	79.11	56.59
9	68.90	68.90	75.42	72.98
10	32.90	32.90	32.90	32.90
11	62.34	62.34	56.88	40.24
12	59.33	100.00	70.92	41.03
13	57.75	76.36	73.86	37.82
14	70.86	65.67	88.44	17.05
15	44.81	33.45	27.21	22.61
16	45.79	66.06	41.37	27.88
17	62.07	62.07	53.33	46.97
18	72.17	72.17	64.32	26.61
19	10.89	17.96	2.84	9.24
20	20.66	20.66	24.55	32.67
Average	**53.47**	**59.48**	**55.38**	**37.99**

Table 4: Samples of BLEU Scores - Various Models to Human Interpretation

8 References

1. Institute of Formal and Applied Linguistics. UFAL Medical Corpus 1.0. 2022.
   https://ufal.mff.cuni.cz/ufal_medical_corpus

2. Konstantinos Skianis et al. *Evaluation of Machine Translation Methods applied to Medical Terminologies.*
   https://aclanthology.org/2020.louhi1.7.pdf

3. Junjie Hu et al. *Domain Adaptation of Neural Machine Translation by Lexicon Induction*
   https://arxiv.org/pdf/1906.00376.pdf

4. Marcin Junczys-Dowmun et al. *Marian: Fast Neural Machine Translation in C++*. April 2018. The University of Edinburgh.
   https://arxiv.org/pdf/1804.00344.pdf