Project 3 - Neural Machine Translation, Entity Recognition, and Auto-Summarization

Ziming Liu, *Member, IEEE,* Hunter Price, *Member, IEEE,* and Logan Wrinkle, *Member, IEEE*

***Abstract*—This project aims to conduct a practice on a variety of transformer-based natural language tasks. We explore Neural Machine Translation, Named Entity Recognition, and Auto-Summarization. This work uses sections from the Arthur Conan Doyle novel "The Hound of The Baskervilles." We show that transformers perform adequately well for each task.**

***Index Terms*—Hugging Face pipeline practice, Neural Machine Translation, Named Entity Recognition, Auto-Summarization.**

I. Introduction

In this study, we aim to conduct three tasks (1) Bi-directional Neural Machine Translation (NMT), (2) Named Entity Recognition (NER) and (3) Auto-Summarization based on sections from Arthur Conan Doyle's book "The Hound of the Baskervilles" we collected in the previous project.

The three tasks apply different pre-trained transformer models, respectively. They are:

1) Bi-directional Neural Machine Translation: M2M100
2) Named Entity Recognition: bert-base-NER
3) Auto-Summarization: T5 model

The three tasks provide us with a better understanding of the application of transformers and experience of implementing corresponding tasks in a real-world scenario.

II. Methodology

In this section, we present the approaches of conducting practice on three transformer-based natural language tasks: **(1) Bi-directional Neural Machine Translation (NMT)**, **(2) Named Entity Recognition**, and **(3) Auto-Summarization**.

A. Bi-directional Neural Machine Translation (NMT)

In the previous two projects we extracted and cleaned text from Sir Conan Doyle and chose five 500 token passages which we reused for this task. Despite previous cleaning of the text, we still needed to do some minor cleaning by removing newlines and extra spaces. Next we used the spaCy nlp function to break our passage into sentences. Once we had the sentences we could use the M2M100 transformer model to translate our sentences to Swahili and Spanish before immediately translating back to English. This double translation was necessary since none of us could speak our target languages. These sentences were then saved into files based on passage and language. In total this left us with 15 total files for the original, Spanish, and Swahili text for analysis.

B. Named Entity Recognition (NER)

In this project, we aimed to apply bert-base-NER to conduct the NER for the cleaned text from Sir Conan Doyle's novel. bert-base-NER is fine-tuned by the CoNLL-2003 NER database and has the ability to recognize four types of entities: location(LOC), organizations (ORG), person(PER), and Miscellaneous (MISC). 4 sections of text within 500 tokens were collected from "Hound of the Baskervilles" and used as input to the transformers.

Because the input texts were extracted from a famous crime novel, we naturally assumed that a NER database collected from Wikipedia would perform better than the CoNLL-2003 dataset. Therefore, we applied Wikiann, the English version, as the input to the bert-base-NER model to conduct the fine-tuning. Because of the label differences between the two datasets, we add two empty labels, B-MIS (Beginning of a miscellaneous entity) and I-MIS (Miscellaneous entity) in the Wikiann dataset.

Due to the varieties of entities in this analysis, we aimed to manually select entities in the 4 collected text inputs. The fine-tuned and pre-trained model were compared based on the number of missing recognized entities from their generated results. Since the fine-tuned model did not contain attention to Miscellaneous entities, we planned to compare their performances from three aspects: LOC, ORG, and PER. If the results missed part of the entity, for example, Sir Charles Baskervilles, the model only detected Charles Baskervilles, and the missed score would be counted as 0.5. Otherwise, it would count as 1.

C. Auto-Summarization

We used the T5 model and the hugging face library to perform auto-summarization. The input to this model consisted of 4 sections of text collected from Sir Conan Doyle's novel "Hound of the Baskervilles." The inputs describe details of the case, the murder, a clue, and the victim. These sections of text are described in more detail in Table III in the appendix. For each input, we manually generated a summary describing the critical details of the text, which is used as reference text when calculating the ROUGE score. Each section of text is given to the model. We then calculate the model's ROUGE score with the generated predictions and the references. In addition to calculating the ROUGE score for the T5 model, we also use the BART and Pegasus models for comparison.

III. Results

A. Bi-directional Neural Machine Translation (NMT)

Manually comparing the translated texts with the original gave our hypothesized results. The Spanish text was very close to the English with the differences being minor enough that the grammar and meaning was preserved albeit through some slight awkwardness. For example, the phrase ”I believe you have eyes in the back of your head.”, ended up as ”I think you have eyes on the back of your head.” The difference between these sentences is ”in” vs ”on” which conveys the same meaning although the phrasing is atypical. The Swahili translation was much less accurate, to the point of being unreadable. The same phrase from above translated to, ”I am very pleased to take you with Jerry.” This reinforces our hypothesis that more common and data-rich languages will have more accurate translations since Spanish performed better than Swahili.

Despite some flaws in the metric, we also calculated the Levenshtein distance and converted it to a percentage in order to have a more concrete result. The formula for this was 1-(Levenshtein distance / Max(originalLength/translatedLength). This meant that the more accurate translation (defined as 1 to 1 character matching) had a higher percentage score. After applying and averaging these scores across all the passages, we found that Spanish averaged .718, while Swahili averaged .505.

B. Named Entity Recognition

The appendix shows an overview of the observation of the pre-trained model’s output in Table I. Overall, the majority of the mistakes from the pre-trained model concentrate on the misrecognition of a person’s title, such as Mrs., Dr., and Sir. Moreover, interestingly, we found that word ”Times” was recognized as an organization. Based on the observation, we did not notice any entity as Miscellaneous, which provides us more confidence in the further comparison.

The observation result of the fine-tuned model is shown in Table II. After comparing the performance under different settings of the training model, we optimized the training model as: learning rate = 1e-4, number of training epochs = 3, batch size = 16, and weight decay = 0.01. Compare with Table I, it is obvious to notice that the fine-tuned model made fewer mistakes. Interestingly, the fine-tuned model recognized ”.” from ”Dr. Mortimer” as a name entity. This may be due to over-fitting. However, due to the large size of the dataset, it is hard to analyze. Overall, the comparison result is shown in Figure 1. The entity recognition performance of the fine-tuned model is better than the pre-trained one.

C. Auto-Summarization

The appendix shows an overview of the summarizations produced by the T5 model in Table IV. The first input, which contained details about the case, produced a summary that was fairly incoherent. It had seemingly unimportant information regarding the input text. The second input, a description of the murder, was slightly better; it contained pivotal details, but only regarding the first half of the input. The third input, a description of a clue, contained mostly correct grammar. Interestingly, it correctly used quotations that mimic how the author would introduce information. The information contained in this output, however, does not include important information about the text. Rather, it seems to have reordered 2 sentences from the input text. The final input, a description of the victim, contained mostly correct grammar. Similar to the previous example, it seemed to recite a few sentences of the input text. Overall the model produced readable text that included very few details regarding the input text.

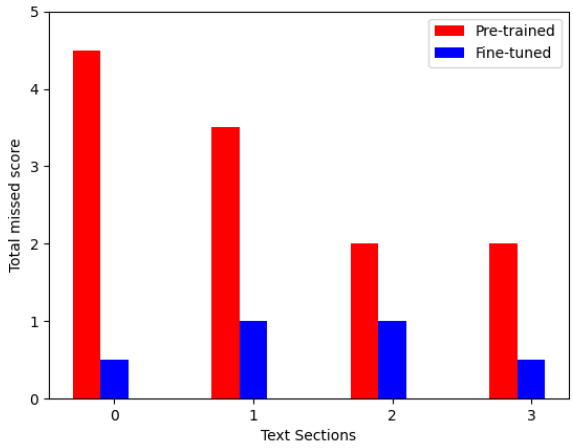

Fig. 1: Total the number of missed recognition between pre-trained and fine-tuned model.

Figure 2 shows the resulting rouge scores for T5, BART, and Pegasus. BART performed the best, surpassing the other two models in rouge1, rouge2, rougeL, and rougeLsum. T5 produced scores similar to BART but slightly worse. Pegasus performed the worst receiving the lowest score in all 4 ROUGE metrics.

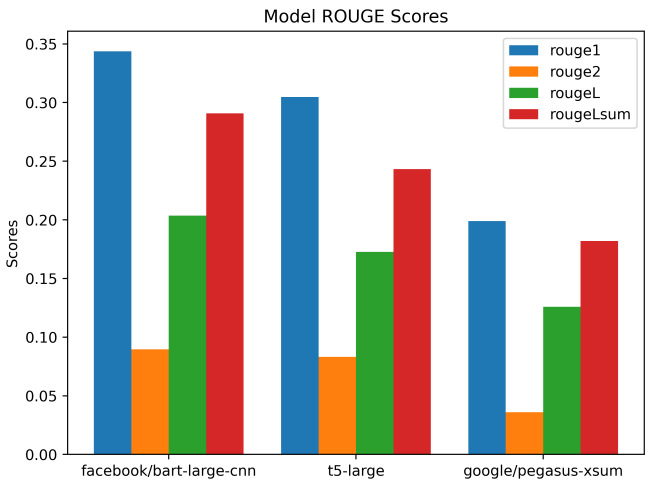

Fig. 2: ROUGE score comparisons across different models.

IV. Conclusion

In this work, we have explored the application of transformers to various Natural Language Processing tasks. We have shown that when transformers are applied to the task of Neural Machine Translation, the resulting translations show promise. Translations to and from Spanish, a language with a much larger collection of text, perform well with a Levenshtein distance of .718. Moreover, Swahili, which does not have a substantially large collection of text to train on, performs as expected with a Levenshtein distance of .505.

In the task of named entity recognition, we compared the pre-trained and fine-tuned NER model. The results show the fine-tuned model performance has a significant increase. In the future, we will generate our own dataset based on Arthur Conan Doyle's books and re-train the model to further improve the performance.

We have shown that the task of Auto-Summarization with transformers performs moderately well. The model was able to generate coherent text with decent grammar. Unfortunately, the summaries were not perfect, but they did produce a few key details from each prompt.

APPENDIX

V. NAMED ENTITY RECOGNITION

Section	Observation
1	6 ”Sir Charles” missed the recognition of the beginning of the person entity–”Sir” ”Barrymores”, which refers to the couple, missed person entity- ”s” Missed an entire location, ”Baskerville Hall”
2	2 ”Sir Charles” missed the recognition of the beginning of the person entity–”Sir” ”Mrs. Lyons” missed the recognition of the beginning of the person entity–”Mrs” ”Coombe Tracey” was recognized as two locations Missed an entire name, ”Grimpen Mire”
3	Recognized ”Gum” as a name. Recognized Times as an organization.
4	3 ”Sir Charles” missed the recognition of the beginning of the person entity–”Sir” ”Dr. Mortimer” missed the recognition of the beginning of the person entity–”Dr”

TABLE I: Observations on the outputs from pre-trained Named Entity Recognition

Section	Observation
1	Progressively defined ”.” from ”Dr. Mortimer” as an name entity.
2	”Mrs. Lyons” missed the recognition of the beginning of the person entity–”Mrs” ”Dr. Mortimer” missed the recognition of the beginning of the person entity–”Dr”
3	Recognized ”Gum” as a name.
4	”Sir Charles” missed the recognition of the beginning of the person entity–”Sir” ”Dr. Mortimer” missed the recognition of the beginning of the person entity–”Dr”

TABLE II: Observations on the outputs from fine-tuned Named Entity Recognition

VI. AUTO-SUMMARIZATION

Ground Truth		
#	Situation	Key Details
1	Details of the case.	Victim goes for a habitual walk Body found at the end of alley. Body found with no indication of violence. Body had facial distortion. Next of kin must be found for inheritance.
2	Description of the murder.	Perp found a way to lure victim. Perp pressures Mrs. Lyons write letter to victim. Perp gets and paints his hound and waits for victim. Hound pounces at victim. Victim dies from heart disease and terror.
3	Description of clue.	Describes perp took care to remove clues. Letter printed in rough characters. Perp tried to seem uneducated. Perp was careless and in a hurry Holmes questions why the perp was in a hurry.
4	Description of victim.	Describes that rumors surround victim’s death. Describes that there is no reason to suspect foul play or natural causes. Victim showed poor health conditions.

TABLE III: Overview of model inputs for auto-summarization

Prediction			
#	Mentioned Details	Grammar	Conclusion
1	Mentions victim has a habit. Mentions evidence with no detail. Mentions need to find victims heir.	Poor	Aptly described situation. Provided no details.
2	Mentions perp acquired influence over Mrs. Lyons. Mentions perp had Mrs. Lyons write a letter to victim.	Poor	Aptly describes first portion of text. Does not provide any details of second half of text.
3	Describes the victims attempt at seeming uneducated. Describes the rough characters.	Average	Describes there is no reason to suspect foul play. Mentions circumstances of victims death are not cleared up.
4	Describes there is no reason to suspect foul play. Mentions circumstances of victims death are not cleared up.	Average	Repeats a sentence of the input. Has a single key detail.

TABLE IV: Overview of model results for auto-summarization.