

進捗報告

1 今週やったこと

- Transformer や BERT についての記事を読んだ
- GPT-2 の loss について調べた
- OpenAI の GPT-3 の API について

2 Transformer や BERT についての記事を読んだ

先生に教えていただいた Transformer や BERT, sentence-BERT 日本語版についての記事を読み、記事の中にあるプログラムを動かしてみた。次は BERT に慣れるためにも実際に使ってみたい。

3 GPT-2 の loss について

GPT-2 について調べていたところ、「ユーザが共感できる悩みの対話コンテンツ生成 [1]」という論文を見つけた。この論文の内容について詳しくはまだ見れていないものの、「GPT を用いたファインチューニングは次の単語を予測する言語モデルタスクとファインチューニングしたい固有タスクのロスを最適化する。」という記述があり、loss とはこの「次の単語を予測する言語モデルタスクとファインチューニングしたい固有タスクのロス」のことだろうと考えた。

次はこの論文に目を通し、内容の理解をしたい。

4 OpenAI GPT-3 API について

OpenAI の GPT-3 の API を動かした。

GPT-3 の engine には

- Davinci(davinci)

デフォルトでオススメされているエンジン。複雑な文章生成などに向いている。

- Curie(curie)

より複雑な文章については Davinci の方が向いているようだが、Curie はセンチメント分析や翻訳などに向いているとのこと。

- Babbage(babbage)

簡単なテキスト分類に向いている。

- Ada(ada)

一番速いモデルで単純な分類などに向いているとのこと。

の 4 つがあるが、今回は特に目的のない文章生成であったため、デフォルトでオススメされている Davinci を用いた。

GPT-3 を用いて文章の生成をするにあたって、どのような文章を生成したいのかを GPT-3 に入力する文章で伝える「プロンプトデザイン」という考えが重要であるということを知った。

これは例えば、ただ GPT-3 に "4" と入力しただけでは出力は "4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20" となってしまうが、
"1 1

2 4

3 9

4 ”

と入力した場合に出力が

”1 1

2 4

3 9

4 16”

となり、ちゃんと左側の数を二乗したものが出力される、といった形のものである。

つまり、例えば会話を生成させたいときに ”Hello, who are you?” とだけ入力すれば

Hello, who are you? ”

“ I ’ m a friend of Mary Ann ’ s. Can I come in? ”

“ I guess so. Just don ’ t touch anything, please. ”

“ I won ’ t touch anything. I promise. ”

She led me into the living room, which still looked like a bomb had hit it, but at least she had cleaned up the blood. I sat on a couch that should

という出力となり、今回は Mary の友人との会話になった（状況はよくわからないが事件だろうか、続く文章によると視点人物はこの Mary の友人のようだ）が、ここで入力を

”Human: Hello, who are you?

Devil:”

といった形にすると、出力は

Human: Hello, who are you?

Devil: I am the Devil!

I’m the guy that’s gonna send you to Hell!

Human: I’m going to go to Heaven.

Devil: I’ll get you!

Human: Don’t you know that God loves me?

Devil: I don’t care.

I’m gonna send you to Hell.

Human: God loves everybody.

Devil: He does not love you!

He loves the priest and nuns, but not you.

And besides

といった形で上の例のような会話文ではなく、自分の指定した形式の会話文が、それもちんちんと人間と悪魔らしい会話が生成された。

会話文の他にも、

”Three ways to defeat the devil.

1. Sprinkle holy water on.

2.”

といった形の入力であれば出力文は

Three ways to defeat the devil.

1. Sprinkle holy water on.

2. Use the sign of the cross on.

3. Say the name of the Lord Jesus Christ on.

”A Prayer to St. Anthony”:

Saint Anthony,

Sweet Saint Anthony,

pray for us,

That we may be made worthy of the promises of Christ.

Let us place all our trust in thy goodness, O holy father, and to make known to thee our needs and desires.

Be our guide in the way of salvation and show us how to avoid the

とちゃんと悪魔を倒す方法が 3 つ提示され、それらしく St. Anthony への祈り文まで生成された。なお、St. Anthony という聖人はいるようだがこの祈り文と同じ文章は検索してもヒットしなかった。

そして、

”Steps to defeat the devil.

1. prepare a crucifix.

2.”

という入力を行うと、

Steps to defeat the devil.

1. prepare a crucifix.

2. On the top of the cross so that the body of Christ is facing you and the head is facing the devil.

3. Pray the rosary while you hold it in front of you.

4. The devil will come and try to take it from your hands, but if you do not give it up, he will become frustrated and bring a lot of temptations to distract you.

5. While you are praying, he will bring temptations to make you stop, such

(実際何をやっているのかはおいておいて) ちゃんと手順らしきものが生成されたことから、文章の本題に入る前の説明、地の文のようなものを入れることにもちゃんと意味があるということがわかった。

このように GPT-3 においてはプロンプトデザインは重要なものであったが、では B3 実験のときに用いた GPT-2 でももしかしたらそうだったのではないかと、思い、次は GPT-2 でも同様の検証をしたいと考えた。

参考文献

- [1] 橋口 友哉, 山本 岳洋, 藤田 澄男, and 大島 裕明, ユーザが共感できる悩みの対話コンテンツ生成, (2021.3.1).