

目大生産工(院) ○岡崎 優佑

あり、ここで間違っただことが分かる。1回間違えばその後はすべて間違ふ。また、復号に失敗しているため取り出された情報も間違っただものになってしまっている。このように嶋津の方法で復号を行うと復号に失敗してしまうことがある。

5 RNNの評価

今回用いるRNNは、学習データから次に出てくるアルファベットの確率を学習し、その学習結果から、入力したアルファベットの次に出てくると予想されるアルファベットを出力するものである。このRNNを以下のように評価した。

約7500字の学習データ(英文のテキストファイル)3)を25字毎に1ユニットに分け、1ユニットの学習を1回とし、RNNに10万回学習させる。その学習結果を用いて性能を評価する。評価時の入力は一節で用いた文である。この文に対してアルファベットの生成確率順にアルファベットを表示させ、アルファベットの順位と入力した1文を比較する。その後、生成されるべきアルファベットの確率順位をヒストグラムとして表す。

My Dear Huret, You have given me an attack of vertigo. I have been reading your biography of our illustrious friend. Its rapid, nervous style, its accumulation of dates and facts, its hurried rush of scenery and events flying past as though seen from an express train, all help to attain what I imagine must have been your object-to give the reader vertigo. I have got it. I knew all these things, but I had forgotten them. They are so many that no one even attempts to reckon them up. We are

図3 学習データの一部

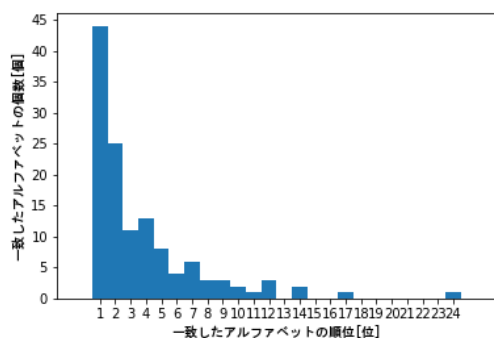


図4 シミュレーション結果

シミュレーション結果は図4のようになる。横軸は一致したアルファベットの順位で、縦軸はその度数である。図4を見ると、1位のアルファベットが最も多く、1位ではないアルファベットも10位までにほとんど入っていることが分かる。この結果から、学習データ内から抜き出した文を入力として用いると、かなり高い確率で入力文と一致するアルファベットが生成されると言える。

これを用いて復号を行うとどうなるか検証する。図2より間違っただアルファベットが最初に出てくるのはfairysClのsである。このsは48字目であるので、RNNで48字目を予想した時のアルファベットの順位をしてみる。図5は48字目のアルファベットの順位の高いものを左から順に並べたものである。これを見ると、正しいアルファベットであるlが間違っただ復号された文のsよりも順位が高いことが分かる。よって、これを用いるとfairylまでは復号に成功すると考えられる。

l . l . t : d ? ! o x i c e s a B - p g h n r l

図5 48字目のアルファベットの順位

6 まとめ

RNNに対して、学習データ内の1文を用いてテキスト生成を行った場合にアルファベットの確率順位を出すと、上位に入力として用いた英文のアルファベットがくることが分かった。これを再圧縮の復号に用いれば、少なくともアルファベットのみを正否判定に用いて復号を行った場合よりは復号に成功する確率が高いと期待できる。

今後の課題として、確率順位が1位ではないアルファベットがあり、復号時に正解のアルファベットよりも順位の高いアルファベットが復号された場合に復号に失敗するので、これを改善する必要がある。また、今回のシミュレーションと同じ方法で学習させたRNNを用いて実際に復号を行った場合、付加情報量をどの程度まで増やせることができるのか、学習は10万回で十分かどうか、10万回よりも少ない学習回数でシミュレーションを行うとどうなるかなどが課題として挙げられる。

「参考文献」

- 1) 滋田康平, 「算術符号化されたファイルに対する情報の付加」日本大学生産工学部卒業論文 (2014)
- 2) 嶋津孝一, 「オフライン情報付加型再圧縮方式の性能改善の研究」日本大学生産工学部卒業論文 (2016)
- 3) Jules Huret, "Sarah Bernhardt" (1899)