

概要

本実験では、英文の単語から最初の 2 文字のみを抽出し、それを元に元の文章を推定するタスクを行う。例えば、"A first two character input method"という元の文章がある場合、それぞれの単語の最初の 2 文字を取り出して"Afitwchinme"という抽出文を生成し、この抽出文から元の文章を推定する。書籍の最終章を除く部分から学習したモデルを用いて、同一書籍の最終章においてタスクの評価実験を行う。また、異なる書籍の最終章以外の文章から学習したモデルを用いた実験も行う。このように学習データの異なるモデルでテストを行うことによって、同一書籍で学習したモデルと他書籍で学習したモデルの間には、どれほど精度の差が生じるのかを調べる。学習では、書籍中に文章の単語の最初の 2 文字と元の単語のペアを学習させた。例えば、"A first two character input method"では、"A/A fi/first tw/two ch/character in/input me/method"を学習する。学習や推定には、KyTea[1]を使用した。また、書籍のデータセットには Project Gutenberg [2]のものを使用した。実験の結果、同一書籍を用いて学習したモデルでは、accuracy が 41.50%となり、他書籍を用いて学習したモデルでは、accuracy が 36.79%となった。差は 4.71 ポイントとなり、同一書籍を用いて学習したモデルの方が accuracy が高くなった。以上の結果から、モデルは学習データの文脈に依存しており、同一書籍を用いて学習したモデルの方が他書籍を用いて学習したモデルのよりも高い精度であることがわかる。

手法

Project Gutenberg の書籍からテスト用データと学習用データを取得する。テスト用データは書籍の最終章のみを抽出したものを使用する。抽出と同時にデータを図 1 のように文章ごと改行で分け、文章に含まれる単語の最初の 2 文字だけを抽出する。学習用データを用意するための前処理として、図 2 のように書籍の本文とは関係ない部分と最終章を除く。その後、文章から図 1 のような単語の最初の 2 文字と元の単語のペアからなる学習用データを作成する。この学習用データを、KyTea を用いて学習し、モデルを作成した。作成したモデルは二つあり、一方はテスト用データと同一書籍から学習したモデルであり、他方はテスト用データとは異なる書籍から学習したモデルである。これら二つのモデルで一つの書籍のテストデータを対象に実験を行いその結果を比較する。評価指標は各文章の accuracy と文章全体の accuracy を平均したものである。accuracy の算出は文章ごとに行い、式 (1) を用いて算出した。例えば、元の文章が "What agonizing fondness did I fell for them" であり、推定文章が "What ago for dishonesty I feared for them" である場合は、推定単語数が 8 であり、推定単語と元単語が合致する数は What, I, for, them の 4 であるため、accuracy は 50%となる。

$$\text{accuracy} = \frac{\text{推定単語と元単語が合致する数}}{\text{推定単語数}} \times 100 \quad (1)$$

```
Iwahuawbyfurealenmewistancoitmomyfeanalmetobecaancaatpewhotdeordewohabemyp
MyfirewatoquGefomycowhwhIwahaanbewadetomenoinmyadbeha
Iprmywiasuofmotowiafejewhabetomymoande
Annomywabewhartocebuwili
Ihatravapooftheaanhaenalthhawtrindeanbacoarwotome
HoIhaliIhaknmatihaiIstmyfaliupthsaplanprfode
BurekemealIdanodianlemyadinbe
WhIquGemyfilawatogasoclbwhImitrthstofmyfien
BumyplwaunanIwamahorothcoofthtounwhpaIshpu
AsniapIfomyatthenofthcewhWiElanmyfare
Ienitanapthtowhmathgr
Evwaslexthleofthtrwhweagbythwithniwanedaaanthscwohabesoanafevtoanunob
Thspofthdesetoflarantocaashwhwafebunosearthheofthmo
Thdegrwthschaatfiexqugawatoraande
ThwedeanIlithmualliantodehIludroumyweex
IknonthgrankitheaanwiquillexBythsaeaonwhIknbythshthwanemebythdeanetgrthIfeIs
FothpuIwiprmylitoexthderewiIagbethsuantrthgrheofeawhotshvafrmyeyfo
AnIcaonyospoftdeanonyowamiofvetoaiancomeinmywo
LethcuanhemodrdeofaglehifethdethnotomeIhabemyadwisoananawwhalasmeththshofmy
```

図1 テスト用データ

```
The Project Gutenberg eBook of Pride and prejudice, by Jane Austen

This eBook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this eBook or online at
www.gutenberg.org. If you are not located in the United States, you
will have to check the laws of the country where you are located before
using this eBook.

Title: Pride and prejudice
Author: Jane Austen
Release Date: November 12, 2022 [eBook #1342]
Language: English
Produced by: Chuck Greif and the Online Distributed Proofreading Team at
http://www.pgdp.net (This file was produced from images
available at The Internet Archive)

*** START OF THE PROJECT GUTENBERG EBOOK PRIDE AND PREJUDICE ***

[Illustration:
GEORGE ALLEN
PUBLISHER
156 CHARING CROSS ROAD
LONDON
RUSKIN HOUSE
]
[Illustration:
_Reading Jane's Letters._ _Chap 34._
]
```

図2 書籍の本文とは関係ない部分

```
Sa/Saville En/England Yo/You wi/will re/rejoice to/to he/hear th/that no/no
I/I ar/arrived he/here ye/yesterday an/and my/my fi/first ta/task is/is to/
I/I am/am al/already fa/far no/north of/of Lo/London an/and as/as I/I wa/wa
Do/Do yo/you un/understand th/this fe/feeling
Th/This br/breeze wh/which ha/has tr/travelled fr/from th/the re/regions to/
In/Inspired by/by th/this wi/wind of/of pr/promise my/my da/daydreams be/
I/I tr/try in/in va/vain to/to be/be pe/persuaded th/that th/the po/pole is
Th/There Ma/Margaret th/the su/sun is/is fo/forever vi/visible it/its br/br
Th/There fo/for wi/with yo/your le/leave my/my si/sister I/I wi/will pu/put
It/Its pr/productions an/and fe/features ma/may be/be wi/without ex/example
Wh/What ma/may no/not be/be ex/expected in/in a/a co/country of/of et/etern
I/I ma/may th/there di/discover th/the wo/wondrous po/power wh/which at/att
I/I sh/shall sa/satiate my/my ar/ardent cu/curiosity wi/with th/the si/sigh
Th/These ar/are my/my en/enticements an/and th/they ar/are su/sufficient to/
Bu/But su/supposing al/all th/these co/conjectures to/to be/be fa/false yo/
Th/These re/reflections ha/have di/dispelled th/the ag/agitation wi/with wh
Th/This ex/expedition ha/has be/been th/the fa/favourite dr/dream of/of my/
I/I ha/have re/read wi/with ar/ardour th/the ac/accounts of/of th/the va/va
Yo/You ma/may re/remember th/that a/a hi/history of/of al/all th/the vo/voy
My/My ed/education wa/was ne/neglected ye/yet I/I wa/was pa/passionately fo
Th/These vo/volumes we/were my/my st/study da/day an/and ni/night an/and my
Th/These vi/visions fa/faded wh/when I/I pe/perused fo/for th/the fi/first
```

図3 書籍から作成された学習用データ

データセット

データセットには、Project Gutenberg で公開されている書籍を用いた。Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (<https://www.gutenberg.org/ebooks/84>) と Pride and Prejudice by Jane Austen (<https://www.gutenberg.org/ebooks/1342>) を選んだ。前者からテスト用データを作成し、両者からモデルを作成した。

実験評価

実験の結果は表 1 のようになった。また、各階級の accuracy のヒストグラムは図 4 のようになった。これらから、同一書籍を用いて学習したモデルでは、accuracy が 41.50% となり、他書籍を用いて学習したモデルでは、accuracy が 36.79% となった。差は 4.71 ポイントとなり、同一書籍を用いて学習したモデルの方が accuracy が高くなった。以上の結果から、モデルは学習データの文脈に依存しており、同一書籍を用いて学習したモデルの方が他書籍を用いて学習したモデルのよりも高い精度であることがわかる。

また、本実験では、“A first two character input method”という元文章からは、“Afitwchinme”というテストデータ用の文章が生成され、このように生成された文章から単語がどこで区切れているのかも推定させている。この例では“A fi tw ch in me”と区切ることができれば、推定が成功しているが、“Af i tw ch in me”と推定すれば失敗である。このように、“I”や“a”などの一文字からなる単語が含まれているかも学習させた。単語“I”から始まる文章が多いため、文字“I”から始まる文章の推定は表 2 からわかるようにバイアスがかかっており、単語“I”から始まると推定する。また、本実験では、文章中の単語の最初の 2 文字とその単語のペアのみを学習しただけである。その単語の品詞や時制などは学習していないため、文法規則を利用した推定が行えていない。表 3 に文法規則を利用して推定が行えていない例をしめす。この例では、過去の疑問形の文章であるため、動詞は原型の“cling”となると文法の時制について学習がされていれば、推定できたと考えられる。このことから、品詞や時制のアノテーションを加えることによって、さらなる精度の向上が見込まれると考えらる。

表 1 実験結果

テスト書籍	モデル書籍	accuracy (%)
Frankenstein	Frankenstein	41.50
Frankenstein	Pride and Prejudice	36.79

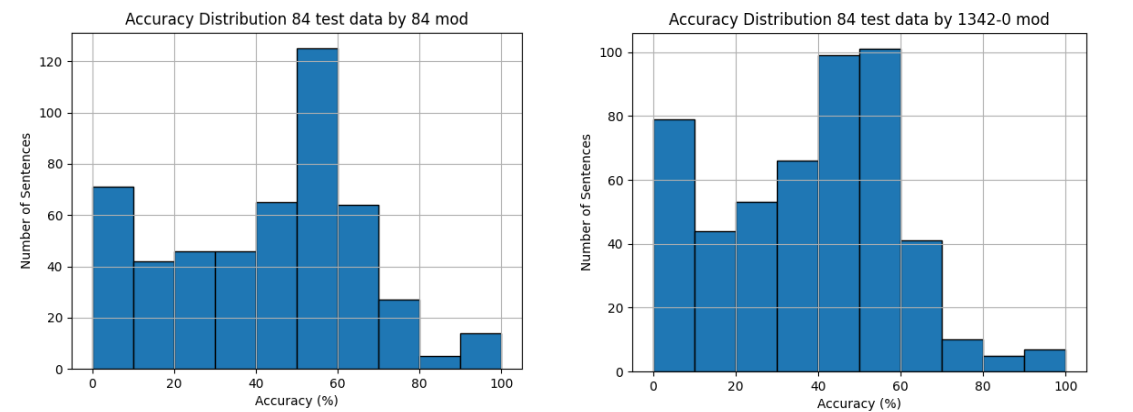


図 4 モデル書籍 Frankenstein (左) モデル書籍 Pride and Prejudice (右)

表 2 文字“I”から始まる文章は単語”I”から始まるバイアスの例

元文章	推定文章
In other places human beings …	I n other place human being…

表 3 時制の文法規則を利用して推定が行えていない例

元文章	推定文章
How did I cling to…	How did I clung to…

まとめ

本実験では、文章に含まれる単語の最初の 2 文字だけを抽出してそこから元の文章を推定するタスクに取り組んだ。学習データとテストデータの書籍が一致する場合には、accuracy の向上が見られた。このことから、モデルは学習データの文脈に依存していることが考えられる。また、元文章と推定文章を比較してみると、”I”に対するバイアスや文法を考慮した推定が十分に行えていないことが考えられる。このことから、品詞や時制のアノテーションを付け加えたデータを用いて学習をすることによって、accuracy が向上すると考えられる。

参考文献

[1] Graham Neubig, 笹田哲郎, 森信介. "KyTea". 2014.
<https://www.phontron.com/kytea/index-ja.html>, (参照 2024-05-29)

[2] Project Gutenberg. "Gutenberg". 2024. <https://www.gutenberg.org/>, (参照 2024-05-28)