

文章自動生成の主な手法における非構造データ量と処理時間についての一考察

A Consideration on Unstructured Data Amount and Processing Time in Main Method of Automatic Text Generation

太田 博三¹

Hiromitsu OTA¹

¹放送大学 教養学部

¹Faculty of Liberal Arts, The Open University of Japan

Abstract: Recent developments in deep learning have been remarkable, from the field of image processing to the field of speech recognition and natural language processing has been penetrated and developed. In this study, we first picked up the following three main approaches to implement sentence generation. 1) Markov chain, 2) automatic summary, 3) sentence generation by deep learning (RNN / LSTM / GAN). As a subject, it was commonly seen that the connection between sentence and sentence was unnatural. We tried the connection between natural sentences and sentences which are also applicable in practice by the above three methods and considered countermeasures.

1. はじめに

1.1. 自然言語処理の研究区分

佐藤[1]は自然言語処理を、解析系と生成系とに分けている。解析系の研究とは、例えば Amazon のレビューなどのテキストが入力となり、それをポジティブ・ニュートラル・ネガティブなどに判別し、出力する。

一方、生成系の研究とは、逆で、入力にポジティブなどと判別された情報とは限らない。出力はテキストである。ここで入力となる情報には、ある基準を設ける必要が出てくる。また機械翻訳のように入力と出力の情報が対価である場合は変換系となる。

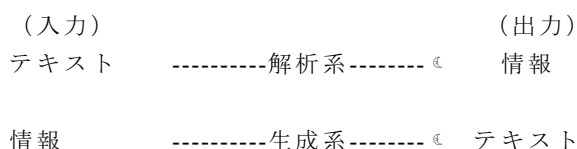


図 1.1 解析系と生成系

1.2. 文章自動生成のタスク設定

筆者は星新一のショートショート AI 小説家に憧れ、文章のジャンルを指定し、キーワードを指定する文章が自動生成される業務アプリ開発を目的とした。

主な仕様は下記の 2 点である。

- 過去の文章の引用ではなく、盗作や剽窃、著作権侵害に当たらない様にする。
- 300-500 文字の自然な文章であること。
- 言い換えにより、元の文章との類似度が低くなること。

また、昨今のニューラルネットワークの発展においても、ゴッホ風の画像やモーツァルト風の音楽まで生成できるが、著作権の話は十分になされていないのが現状である[2]。

1.3 文章自動生成の注目度

自動要約や文章自動生成のコンテスト (E2E NLG Challenge

<http://www.macs.hw.ac.uk/InteractionLab/E2E/>) が毎年、欧米を中心に開催されており、世界的に盛んである。

ディープラーニングの発展、特にリカレントニューラルネットワーク (RNN) やその発展系の LSTM そして敵対的生成モデル (GAN) などによって勢いづけられている。また、文書自動要約 (Text Summarization) も 10 年以上前から盛んに行われており、文章自動生成は文章自動要約と重なり合う部分もある。

*連絡先: 太田 博三 (放送大学教養学部)
〒112-0012 東京都文京区大塚 3-29-1
Email: otanet123@gmail.com

2. 本研究で用いた手法

2.1 各手法についての概観

文章自動生成を大きな枠で捉えるならば、次の3つの手法が考えられる。

1. マルコフ連鎖による文生成.
2. 自動要約/ 文圧縮による文章自動生成.
3. リカレントニューラルネットワーク/ LSTM による文章自動生成.

この他にも制御文によるフレームワークを用いた文章自動生成などもある。

2.2 マルコフ連鎖による文生成

マルコフ性 (Markov property) とは、次の状態が過去の状態に依存せず現在の状態のみによって決まる性質のことである。マルコフ性が存在する場合、状態が $\{q_0, q_1, q_2, q_3, \dots, q_{n-1}\}$ の n 通りを取るような状態遷移において、現在の状態が q_i であった時に次の状態 q_j に遷移する確率は純粋に次の状態と現在の状態のみで記述され、 $P(q_j | q_i)$ で決定される。同様に、状態遷移した順に並べた順序列 $\{a_0, a_1, a_2, \dots, a_{m-1}\}$ の生成確率は $\prod_{i=1}^m P(a_i | a_{i-1})$ と表すことができる。このようなマルコフ性を備えた確率過程を総称してマルコフ過程 (Markov/Markovian process) と呼ぶ。その中でも状態空間が離散集合を採る (つまり取りうる状態を示す値が連続的でなく離散的である) ものを特にマルコフ連鎖と呼ぶ[3]。マルコフ連鎖による文生成の例を示す。

{今日は、いい天気、です、.}という状態の集合があったとする。

「今日は」という状態の次に「です」という状態がくる確率は $P(\text{です} | \text{今日は})$ で表される。 $P(\text{今日は} | \text{今日は})$, $P(\text{いい天気} | \text{今日は})$, $P(\text{です} | \text{今日は})$, $P(. | \text{今日は})$ の4つのうち、最も高い確率をもつのは $P(\text{いい天気} | \text{今日は})$ であるはずである。確率的に「いい天気」へと状態が遷移すると、「今日は いい天気」という文が生成される。さらにその次の状態は $P(\text{今日は} | \text{いい天気})$, $P(\text{いい天気} | \text{いい天気})$, $P(\text{です} | \text{いい天気})$, $P(. | \text{いい天気})$ の4つを比較して決定される。確率が十分に正確であれば、「今日は いい天気 です .」という文の生成確率が最も高くなり、結果的にこの並びが一番選ばれやすくなる。」という遷移が発生した回数) / (「なんとか」という状態になった回数) で求められる。この確率の良し悪しで生成された文の良し悪しが決まる。

実際の文生成には状態として文節ではなく「形態素」と呼ばれる単語のようなものが用いられることが多い

ほか、直前の1個ではなく、4個までを考慮した高階マルコフ連鎖を使うことが多い。N-gram モデルと呼ばれる。

2.3 自動要約による文章自動生成

自動要約の古典的な H.P. Luhn [4] は、テキスト中の重要な文を抜き出し、それを出現順に並べることによってそのテキストを読むべきか否かを判定するといったスクリーニングのための要約が自動生成できることを示した。つまり、自動抄録に似ており、「理解し、再構成し、文章生成」というのではなく、「理解する箇所が重要部に近似する」と割り切って考えたものである。重要語の決定には、単語の頻度を用いるなど、現在の自動要約の流れは、H.P. Luhn の影響が少なくない。

また、ニューラルネットの文圧縮の研究も進んでおり、seq-to-seq モデルでは ROUGE スコアの低下はモデルへの入力文が長すぎると新聞記事のヘッドライン生成が劣化する問題点がある。Attention の付いていない encoder-decoder model を使用し、encoder には片方向 LSTM を適用し、最適化には adam を用い、出力時には beam-search を用いるなどが良い結果が出ているとされている[5]。さらに文抽出手法を強化学習にしたテキスト自動要約手法もの研究も行われている[6]。

2.4 リカレントニューラルネットワーク

(RNN)/LSTM/GAN による文章自動生成

Andrej Karpathy の char-rnn による tiny shakespeare[7] が有名である。詳細は述べないが、今までの単語列として、もっともらしい次の単語を予測することを Long short-term memory(LSTM)が担うもので、RNN の拡張として、1995 年に登場した時系列データに対するモデルまたは構造の一種である。しかし文章自動生成においては、LSTM の L が決して字面通り Long ではない。例えば Epoch が 100 を超えないとまともな一文にならないなどの問題がある。GPU の使用も取り入れるなど、学習には非常に手間暇を要する。Epoch が 2 桁であると、生成される文章が同じ句などの表現が出てくるなどの症状が見受けられ、大半はこの様な学習に悩まされることになるため、工夫が必要になる。

3. 実験結果 ([7])

3.1 各手法の実験概要

本研究ではファクタ定義は次のように定めた。ファクタ定義として、「文章自動生成とは、特定のジャンルにおいて過去の記事を学習データとして、300-500

文字前後の文章を自動生成する」ために、さかのぼって倍以上の文書量が必要となった。ここでは文と文とのつながりに留意した。

・手法一覧（再定義）：

- 1) マルコフ連鎖及び Doc2Vec による文章自動生成.
- 2) 単語出現頻度に基づく文章要約.
- 3) RNN/ LSTM による文章自動生成.

※1)での Doc2vec はマルコフ連鎖によって生成された複数の文章の類似度を計り、近いものどおしを結合するために用いた。

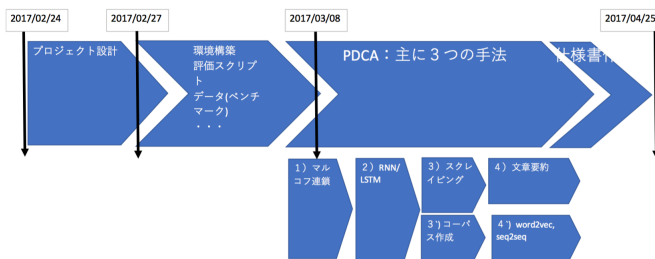


図 3.1.1 PJ フロー図

工程数	作業内容/項目	作業詳細	備考	2月20日	2月21日	2月22日	2月23日	2月24日
5	全体像の把握	論文調査	※マルコフ連鎖の精度は低					
14	準備 (環境構築など)							
2		Python開発環境構築						
2		theano/ keras/						
2		Chainer/ Tensorflow/D4J						
2		LexRank/ TextRank						
2		word2vec/doc2vecによる単語類似度算出						
2		tensorflow/ seq2seq						
1		文章生成スクリプト作成						
1		文章の評価スクリプト作成						
24	イテレーション							
7		1)マルコフ連鎖とDoc2vecによる文章の自動生成	1)スクリプト確認, 2)文章生成, 3)評価のアンケート, 4)解釈					
7		2)Luhnによる文章要約	1)スクリプト確認, 2)文章生成, 3)評価のアンケート, 4)解釈					
7		3)keras(RNN/ LSTM)による文章の自動生成	1)スクリプト確認, 2)文章生成, 3)評価のアンケート, 4)解釈					
3		※ a)tensorflow/ seq2seqRNNによる文章自動要約	1)スクリプト確認, 2)文章生成, 3)評価のアンケート, 4)解釈					
2	報告書/ 仕様書作成							
1	納品							

図 3.1.2 PJ スケジュール

以下に用いたデータセットの詳細について次の表に示す。

表 3.1.1 文書データの容量と文字数

文書データ名	容量	文字数
暮らしと健康雑学.txt	463KB	150235文字
ドクターズ_オーガニックコスメ.txt	200KB	65403文字
社説 (毎日新聞社)	490KB	336817文字
社説 (朝日新聞社)	1MB	159435文字
百貨店 (yahoo)	564KB	187285文字

・評価手法：

実務に役立つものを重視するため、まずは人手による評価に委ねることにした。リッカードの尺度評価を用いた。

次に評価に用いた各手法の文章生成の手順を示す。

- 1) マルコフ連鎖及び Doc2Vec による文章自動生成,
 1. 文章を単語に形態素に分解する,
 2. 単語の前後の結びつきを辞書に登録する,
 3. 辞書を利用してランダムに作文した.

※文章の長さは何文かを指定できるスクリプトを用いた。

4. Doc2vec/ Gensim による文書の類似度を計算
5. 文書間の類似度の高い数値の文書を求める
6. 類似度の近い文書を結合し、合計で 500 文字の文書とした。

2) 単語出現頻度に基づく文章要約, ここでは、H.P. Luhn(1958)による要約アルゴリズムを基に簡略化したものを用いた。

1. 形態素に分解し、各段落で単語の一覧を作成する.
2. 段落内で、もっとも多くの単語を含む文を探し、ランキングにする.
3. ランキング順に表示する.

3) RNN/ LSTM による文章自動生成

Recurrent Neural Network(RNN)の一種の Long Term Short Term Memory(LSTM)による文書生成である。RNN はニューラルネットワークを再帰的に扱えるようにしたもので、時系列モデルの解析を可能にしたものであるとされている。LSTM は RNN を改良したものであり、長期的に記憶を保存するためにブロック (ゲート) を採用したものである。

例えば、アルファベット順で「ABC」と来たら、「D」が来る可能性が高いようにしたものである。LSTM による文書自動生成は当然であるが、形態素解析を行わない。

※ エポック数は初期値を 60 とした。テキストの記憶は 20 とした。理論的には、このエポック数が大きければ大きいほど文書生成の精度が高くなるはならないと考えられるが、元データの大きさによっても影響されると考え大きめに取った。

3.2 各手法と好ましいと思われる文字数

憶測の範囲に過ぎません。定量化できればと試行錯誤中である。

- 1) マルコフ連鎖と Doc2vec による文章の自動生成: 100-200 字程度の文書
- 2) keras(RNN/ LSTM)による文章の自動生成: 5000 文字以上の文書
- 3) Luhn による文章要約: 1000 字以上
- 4) LexRank/ TextRank による文章要約: 300-400 文字
- 5) 文圧縮による文章要約: 10000 文字以上の文書
- 6) tensorflow/ seq2seq による文章自動要約: 10 万文字以上

4. 実験結果 ([8])

4.1 実験で用いた各手法の長所・短所

1. マルコフ連鎖 (形態素解析→辞書作成→文生成)

・メリット: 文章自動生成に時間を要さない. 極めて短い時間で文章自動生成が可能であること.

・デメリット: 文と文とのつながりが自然でない.

2. 自動要約 (頻出キーワード→それを含む文→昇順に並べ並べ返す)

・メリット: 文と文とのつながりが不自然でないことが多い.

・デメリット: 圧縮されるため, ある一定の学習コーパスが必要となること, リアルタイムには作れないこと. 元の文章のままであり, そのままでは使えないこと.

4. LSTM: (日本語コーパスの品質が良ければの条件つき)

・メリット: 学習回数によっては可能性があること.

・デメリット: 莫大なコーパスと学習時間が必要.

4.2 実験で用いた文章と評価結果

(その1)

[実験: 2パターンでの検証結果]

1) 1. (元データ)

文書データ名	容量	文字数	URL
暮らしと健康雑学.txt	463KB	150235文字	http://archives.mag2.com/0000252795/

マルコフ連鎖による生成文章の事業者評価とその生成文章を以下に示す.

4.2.1 マルコフ連鎖と明示した場合の2つの文章の評価結果

- ・元の記事のままではない点の評価できる.
- ・元の記事が長い記事であれば, 文と文のつながりはよくなかった.

4.2.2 マルコフ連鎖による生成された文章例 (2つ)

マルコフ連鎖による生成文章例として, 10個の文章と15個の文章とでランダムに文章を生成したものである.

4.2.2.1 マルコフ連鎖による生成された文章例その1

化粧水などで泡立ててくださいね. そのまま化粧水を防腐剤, 安定剤や酸化亜鉛などの影響で, かさつかず

しっとりとししました. ドクターズオーガニックの無添加化粧品の材料とコールドプロセス法によって製造することがわかります (新陳代謝が正常に導きます. またお手元に届くよう, 一刻も早い安全性と, とても良い下地クリームにした作りになっていますが, 抗菌仕様製品では次のような選び方が推奨されています. 毎日身体を動かすことで, 自分達の基礎化粧品の中身を確かめてくださいね. 無添加石鹸で... 油分を十分に補って, 盲信は禁物です. ナイロンタオルやボディケア用品の基材としてよく使われています. という具合に自分の肌トラブルを起こさない, これは現実的に引き締まった感じがするかもしれませんよね? 乾燥が気になることもあります. 刺激感などの添加があります. (361文字)

4.2.2.2 マルコフ連鎖による生成された文章例その2

例えばクモノスカビの大きさを保ち, かさかさ, ボロボロに. 一体何が無添加石鹸を泡立て, 泡の保持のため, ニキビ肌, 敏感にしましょう. 「今日はとても不自然だと悪化します. また, 脂肪を燃焼させ除去することフットクリームの全成分しかし, 同時に流れは細胞から二酸化炭素や老廃物を回収してください. シアバター 1,800円 (約70g) ご購入はこちらから」フットクリームの全成分漢方薬としては, 洗浄感の良いハンドクリームには皮膚がんの原因は消毒も殺菌して作られますが, 散乱剤②の2種類がありませんが, すでにたくさんの化粧水と, やわらかくて, たっぷりのお湯でお肌は, 天然の成分かつ健康法として推奨された安全性はまだまだ不確かです. でも, 必ずしも石鹸が必要なわけでも, 瞬間湯沸かし器などで泡立ててください. 乾燥肌対策のため植物性油脂であっても, きれいな水だけで数百種類以上の化学成分を毎日肌にのせたりすることが大事です. 原料への安全性はありません. この後, あるいはお風呂で体を温める効果のある人々 (活字関係) からは常温で固体のためのスキンケア (455文字)

4.3 実験で用いた文章と混合した手法の評価結果 (その2)

以下の文章が自然であるかに留意し, 5段階評価を行った. ※評価尺度は次の通りです. (自然な日本語) 5-4-3-2-1 (機械的な日本語) また, 気がついた問題や箇所は下線のスペースに記入してください.

b (一回目の言い換え後 448 文字)

横浜市の求職実態を把握しよう。都会に隣接した大型住宅地とお店が並ぶ地域の横浜市。神奈川県を中心に、県内で一番の都市として伝えられているのが横浜市です。政治と経済の中心部は、横浜市中区や西区に集約されています。馬車道や山下公園、横浜中華街なども近くに存在するため、観光地として知られています。横浜港に面してホテルや経済施設、商業施設が建ち並ぶ横浜みなとみらい 21 も、この地域に含まれます。横浜市は黒船来航といった伝統的な事実もあり、西洋の建造や帰国子女の学校、海外旅行者を多くみるでしょう。横浜駅を軸に広がるダウンタウンや観光地では、レストランやさまざまなお店が並んでいます。私鉄や都営地下鉄が多くあることから、地域によっては移動が楽で、都心の大型住宅地としても有名です。横浜市には、カレッジの施設も多いことから、大学の近郊や通学しやすい点で多くのアルバイトを見つけることが可能でしょう。経験のない人から始められる職業、技術が習得できるものなど、自分に適したアルバイトを見つけることができます。

c (2回目の言い換え後 405 文字)

横浜市で求職実態を把握しよう。都会の隣接した大型住宅地とお店の並ぶ地域の横浜市。神奈川県が中心でもあり、県内の一番の都市として伝えられているのは横浜市です。政治や経済の中心部が、横浜市中区と西区へ集約できます。馬車道と山下公園、横浜中華街などが近くへ存在することで、観光地として知られています。横浜港に面してホテルと経済施設、商業施設の横浜みなとみらい 21 が、この地域に含んでいます。横浜市の黒船来航といった伝統的な事実があり、西洋の建造と海外旅行者が多くみられるでしょう。横浜駅に広がる行楽地で、食堂と多くのショップがあります。鉄道がたくさんあることから、場所によって、移動が容易で、都心のベットタウンとして人気があります。横浜市では、大学の施設も多く、大学周辺や通学面でたくさんのアルバイトが見つかるでしょう。未経験から始められるジョブやスキルがマスターできるものを、自分に合ったアルバイトを見つけられます。

4.4.2 n-gram(n = 1-5)での定量化と言い換え回数について

以下のように定義した。

- a: 元の文章,
- b: a を言い換えた文章,
- c: b を言い換えた文章

- a から b への言い換え総数: 56回
- b から c への言い換え総数: 38回

• a と c の類似度の比較

- 2-gram: 1.151
- 3-gram: 0.582
- 4-gram: 0.506
- 5-gram: 0.388

• b と c の類似度の比較

- 2-gram: 1.386
- 3-gram: 0.798
- 4-gram: 0.3171
- 5-gram: 0.2075

4.4.3 考察結果

3-gram, 4-gram, 5-gram と言い換え回数と類似性との関係は負の関係にあり, n が 5 に近づくほど、言い換え回数が大きく増大すると考えられる。

5. まとめ

文と文のつながりについては、自動要約との関連や文と文とのつながりを entity-grid model[11]を用いて局所的なつながりの良さを表現するなどの談話構造解析[9][10]があるが、手動で行う判断を自動化することが可能か試行錯誤中である。Sentence ordering なども検討したいと考えている。またディープラーニングを用いた方策としては、敵対的生成ネットワーク (Generative Adversarial Network: GAN)による精度向上も精度向上が期待され、実験中である。今のところは完全自動化ではなく、人手を含めざる負えなく、主に制御文による文章自動生成が無難と思われる。

文 献

- [1] 佐藤理史 コンピューターが小説を書く日. 日本経済新聞出版社, 2016
- [2] Leon A. Gatys et al. A Neural Algorithm of Artistic Style, 2015
- [3] Wikipedia “<https://ja.wikipedia.org/wiki/マルコフ連鎖>”
- [4] H. P. Luhn. The Automatic Creation of Literature, IBM Journal, 1958
- [5] 長谷川, 平尾, 奥村, 永田. 文圧縮を活用したヘッドライン生成, 言語処理学会, 第 23 回年次大会発表論文集, 2017
- [6] 太田. 文章自動生成の事前調査報告書. 2017
- [7] 太田. 文章自動生成の最終調査報告書. 2017
- [8] 笹野, 飯田. 文脈解析, 自然言語処理シリーズ 10, コロナ社, 2017
- [9] 黒橋. 自然言語処理, 放送大学教材, 2016
- [10] 横野, 奥村. テキスト結束性を考慮した entity grid に基づく局所的な一貫性モデル Journal of natural language processing 17(1), 2010-01-10, 言語処理学会