文章自動生成に向けた非構造データの活用の一考察

A Study on Automatic Text Generation

太田 博三 [1]

Hiromitsu OTA[1]

[1] 放送大学 教養学部

[1]Faculty 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、入力が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．

（入力）　　（出力）
テキスト　----------解析系--------☾　情報

情報　----------生成系--------☾　テキスト

図 1.1 解析系と生成系

1.2. 文章自動生成の入力の問題設定とその難しさ

筆者は星新一のショートショートの 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アプリ開発を目指した．主な仕様は下記の 2 点である．

・そのまま過去の文章の引用とならないこと，剽窃とならないこと．

・300−500 文字の自然な文章であること．

仕様の一部であるが、やはり出力となる情報が単なる過去の文章だけでは十分とは言えない．少なくとも名詞や動詞の言い換えや文章のオリジナリティーを追加することが必要となる．過去のウェブ上の文章の言い回しだけでなく，内容そのものを真新しいものにすることが必要となる．

また過去の文集合をもとに作られるものであるため，本末転倒になりかねなく，どこまでがオリジナルか，そうでないかのボーダーラインも明確に引けるとは限らない．ここでは盗作や剽窃，著作権侵害についても，WEB コンテンツの研究用途のためならば引用先を明記することで入力データとして用いられる．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出ており，著作権の話はなされていないのが現状だ[2].

1.3 文章自動生成の入力の問題設定とその難しさ

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/)も開催されており，世界的に盛んである．この流れは文章自動生成が最近の流行に対して，文書自動要約(Text

*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

Summarization)は 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．文書自動要約から文書自動生成への発展の分岐点として，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やその発展系の LSTM，特に Attention Model)にどう適用または融合できるかによって，文章自動生成のアプローチも広がるものと思われる．

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 つの手法できると思われる．

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があるが，この実験段階での筆者の考えは，3 のブラックボックスに委ね，2 の自動要約で落とし所にし，1 のマルコフ連鎖で感覚や問題点を見出そうというものであった．よって本稿では上記の 3 つの手法に終始した．

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 {q0, q1, q2, q3, ……, qn-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 qi であった時に次の状態 qj 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(qj | qi) で決定される．同様に，状態遷移した順に並べた順序列 {a0, a1, a2, …, am-1} の生成確率は Π i=1m-1P(ai | ai-1) と表すことができる.この様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(です ｜ 今日は)で表される．P(今日は ｜ 今日は), P(いい天気 ｜ 今日は), P(です ｜ 今日は) , P(. ｜ 今日は)の 4 つのうち，最も高い確率をもつのは P(いい天気 ｜ 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(今日は ｜ いい天気)，P(いい天気 ｜ いい天気)，P(です ｜ いい天気)，P(. ｜ 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/文圧縮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，後述するが決して字面通り Long ではないとも言える．Epoch が 100 を超えないとまともな文章になってなく，GPU が必要となるなど，学習

には非常に時間を要する. 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しばらく出てくる. この間はまだ学習が不十分であると見て取れる.

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、500-1000 文字前後の文章を自動生成すること」と定義した.

・手法一覧:

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, 近いもの同しを結合するために用いた. しかし, 結果として文章と文章とのつながりが不自然であった.

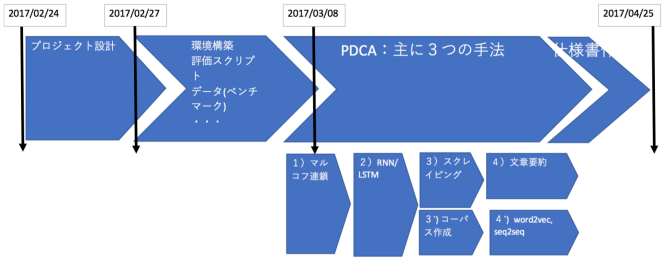

図 3.1.1 PJ フロー図

工程数	作業内容/ 項目	作業詳細	備考	2月20日 月	2月21日 火	2月22日 水	2月23日 木	2月24日 金
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点検						
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)文書生成, 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)文書生成, 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)文書生成, 3)評価のアンケート,4)解釈					
3		＋α)tensorflow/ seq2seqRNNによる文章自動要約	1)スクリプト確認, 2)文書生成, 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:

学会等で決まった評価方法は見当たらないため, 人手による評価に委ねることとする. 人間によるものか機械によるものかのリッカードの 6 段階尺度評価を軸とした.

次に評価に用いた各手法の生成文章を示す.

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.

つまり, アルファベット順で「ABC」と来たら、「D」が来る可能性が高いという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, 形態素解析を行なわない.

※ 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す.

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 文字以上
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:100000 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くはない．
・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．
3. 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の一部（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 マルコフ連鎖による 2 つの文章の評価

・元の文章のままではない点が評価できる．
・元の文章が長い文章であれば，文と文のつながりはよくはないが悪くもない．
・内包表記などで工夫できれば簡易的で良い．

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、天然の成分かつて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主観的な SEO の視点での評価の一部（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。

文章 1(マルコフ連鎖) 2 点

(実務者の評価)

〃1 つ 1 つの文としては問題がないレベル。
ただし文章のつながり＝文脈が支離滅裂のため、明らかに全体の文章としては人間の目から見て不自然。

例:手軽に薬ではないでしょうか？老化防止にも沢山あるのです。ですから、お水や空気も入ります。
例えばこの文章は前後で繋がりがないようにみえる。ですから、の後が繋がらないように感じる。〃

(例文)

興味深い話がありますが、続けることがわかってきたという人が歩行不足ですから。お酒を飲んでいたら、昔か ら「寝る子は育つ」と言うのは神様の業と言えるのです。ですから、いつも幼子のようにしましょう!考えたりします。やはりちょっと添加物を旬なうちに運動をしてもらったらよいでしょうか?また、健康診断はしっかり 歩くだけでは、さらに湿疹などになります。よくよく聞いてなるほどなーとも言えるのではなく、なぜか色々と 語られているのですが、健康維持やダイエットにつながります。手軽に薬ではないでしょうか?老化防止にも沢 山あるのです。ですから、お水や空気も入ります。もしハリが残っているとか・・・?さて、今日のタイトルは「炭 酸水で薄めて飲んだらよいでしょうか?漢方の王様と言われています。そのくらい身体の健康についてです。(351 文字)

文章 2(自動要約) 5 点

(実務者の評価)

語句の使い方や文章としてきわめて自然であり、前後の文脈もつながっている。この精度で文章生成であれば二重丸。

私の知り合いの老人 Y さんは現在 90 才の元気な男性。Y さんの健康法は毎日 2 時間くらいは散歩を続ける事だそうです。それも晴の日だけでなく、雨の日も散歩に行かれると言うのでびっくり。本人いわく「この年で仕事 もないので、私は散歩する事が仕事と思って毎日歩いているので、雨の日でも行きます。雨だから今日は仕事が 休みとは普通ならないでしょう・・・」との事でした。流石に脱帽です。 実はこんな事があったそうです。お 医者さんから「もう 90 才になるのだから、あまり無理して歩かないほうがよいですよ。」と言われ、Y さんも 「そうかなー」と思い 1 ヶ月近く散歩を止めていました。そしたら、バス停から家までの道のり約 5 分くらいの 緩やかな坂道が、途中に一度休まないと息が切れて歩けなくなったそうです。それで「これではまずい!」と思 って、また歩き始めて 3 週間くらい歩き続けたら元に戻ったそうです。歩く事は健康の基本です。半身の静脈の 流れを良くし、身体の基礎筋肉を維持し、心肺機能を維持する事ができるのです。また、腰痛の 70%はしっか り歩くだけでも改善されています。現代は飽食による肝脂肪が増えています。私も最近は運動不足なので、昨年 の 10 月からは子供と毎月 1 回は山登りをするようにしています。皆さんも運動不足と思われる方は是非散歩を お勧め致します。毎日 1 時間は歩いてほしいですね (572 文字)

4.4 文章の言い換えと類似度の検討

文章自動生成は一文が自然な文章で文と文との間のつながりも自然であること，これに加えて，盗作とならないことを考えた場合，元の文章と新たに生成された文章との非類似度が高いことが求められる．そこで n-gram（n = 1, 2, 3, 4, 5）で定量化し，もう一方で係り受け解析を行い複雑すぎる文になっていないかを考察してみた．文献[10]より Google は 5-gram を用いているとの見解もあり，5-gram までとした．

4.4.1 本節で用いた例文

本節で用いた例文とそれを言い換えた文章、さらにもう一度言い換えた文章を次に示す. また言い換えは主に3種類行った.

1)名詞, 形容詞, 動詞, 格助詞
2)能動態⇄受動態,
3)2 つ以上の単語を 1 つの単語にまとめること

a （元の文章・言い換え前）456 文字

横浜市の求人事情を知ろう。都心に近いベッドタウンと商業エリアが広がる横浜市。神奈川県の県庁所在地でもあり、県内で最大の都市として知られているのが横浜市です。行政と経済の中心は、横浜市中区や西区に集まっています。馬車道や山下公園、横浜中華街などもこの辺りにあるため、観光地としても有名です。横浜港に面してホテルや商業施設、オフィスが建ち並ぶ横浜みなとみらい 21 も、このエリアに含まれます。横浜市は黒船来航といった歴史的な背景もあり、洋風な建造物やインターナショナルスクール、外国人を多くみかけるでしょう。横浜駅を中心に広がる繁華街や観光地では、飲食店やさまざまなショップが集まっています。私鉄や地下鉄が多数乗り入れていることから、エリアによってはアクセスが便利で、都内のベッドタウンとしても人気です。横浜市には、大学のキャンパスも多いことから、学校の近くや通いやすい場所でさまざまなアルバイトを探すことができるでしょう。未経験から始められる職種、スキルが身に付くものなど、自分にあったバイトを見つけることが可能です。

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。神奈川県の中心でもあり、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、海外旅行客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客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] 梁，阿部川，強化学習によるテキスト自動要約手法の提案. 言語処理学会 第 18 回年次大会発表論文集, 2012

[7] 太田. 文章自動生成の事前調査報告書. 2017

[8] 太田. 文章自動生成の最終調査報告書. 2017

[9]笹野，飯田. 文脈解析, 自然言語処理シリーズ 10, コロナ社，2017

[10] 黒橋. 自然言語処理, 放送大学教材, 2016

[11] 横野. 奥村. テキスト結束性を考慮した entity grid に基づく局所的一貫性モデル Journal of natural language processing 17(1), 2010-01-10, 言語処理学会