Vol.2017-IFAT
2017/07/22

主な文章自動生成手法の一考察
-文と文とのつながりを課題として-

太田博三[†1]

概要：ここ数年の深層学習の発展は目覚しいものがあり，画像処理の分野だけでなく，自然言語処理や音声認識の分野まで及んでいる．本考察では，業務の一環として，文章生成を実践し，そこで用いた主に３つの手法を取り上げる．１）マルコフ連鎖，２）自動要約，３）ディープラーニング（RNN/ LSTM）による文章生成．課題として，文と文とのつながりが不自然であることが共通する要因であった．実務で通用する自然な文と文とのつながりの課題を考察したものである．

キーワード：文章自動生成，マルコフ連鎖，自動要約，RNN/ LSTM，文と文のつながり

A consideration of the main method of automatic sentence generation
- Connection between sentence and sentence as a subject -

HIROMITSU OTA[†1]

Abstract: The development of deep learning in recent years has been remarkable, and it extends not only to the field of image processing but also to the field of natural language processing and speech recognition. In this study, as part of the work, I will practice sentence generation and pick up the three main methods used there. 1) Markov chain, 2) automatic summary, 3) sentence generation by deep learning (RNN / LSTM). As a subject, it was a common factor that the connection between sentence and sentence was unnatural. I would like to examine issues related to the connection between natural sentences and practical sentences. [**]

Keywords: Automatic sentence generation, Markov chain, automatic abstract, RNN / LSTM, connection between sentence and sentence

1. はじめに

1.1 自然言語処理の研究とその区分

佐藤[1]は自然言語処理の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になり，それをポジティブ・ニュートラル・ネガティブなどと判別し情報として出力する．一方，生成系の研究とは，逆で、入力は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

筆者は検索エンジン対策(SEO)に従事しており，業務効率化の一環として、ジャンルを指定し，キーワードを指定するとサイトのテキスト文が自動生成するシステムを開発することを命じられた．主な仕様は下記の 2 点である．

・剽窃になってはいけないこと，そのまま過去の文章の引用とならないこと．
・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を変換し，真新しいものにすることが必要となる．

過去の文の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，どこまでが合格か，不合格かのボーダーラインも明確でなく，システム開発そのものの問題設定が曖昧でもある．こうしたことからも，ある基準はあるものの SEO のグレーゾンが介在している．このような背景の中での取り組みである．ここでは盗作や剽窃，著作権侵害についても，WEB コンテンツ上の定義と法的な定義との重複やズレが存在している，なかなか定義しづらいもので

[†1] (株)Speee/ Speee, Inc
a) otanet123@gmail.com

ある．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，著作権が後手後手に回っているのが現状だ[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/)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 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あう部分もある．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は，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)によって，文章自動生成のアプローチが広がり，注目されている．

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 つの手法に着目した．

2.2 1. 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 Πi=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 2. 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，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鳥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 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

Andrej Karpathy の char-rnn による tinyshakespeare[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，Recurrent Neural Network(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いなかったり，GPU が必要に成るなど，学習に時間を要する．同じ表現が出てくる間はまだ学習が不十分などの症状が見て取れる．

3. 実験結果(内部資料[7])

3.1 各手法の実験概要

ファクタ定義は次のように定めた．

ファクタ定義:

・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する．

・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自然でなかった．

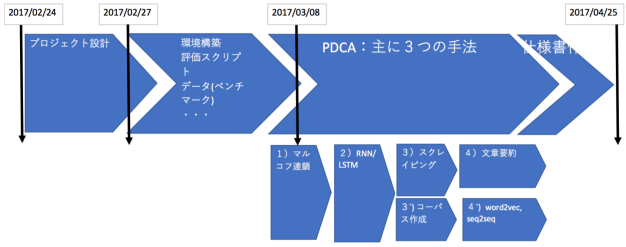

図 3.1 PJ フロー図

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)解析					
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)解析					
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)解析					
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)解析					
2	報告書/ 仕様書作成							
1	納品							

図 3.2 PJ スケジュール

用いたデータセットの詳細について次の表で示す．

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

図 3.3 文書データの容量と文字数

・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, 形態素解析を行和ない．

※ 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 得られた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．

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(内部資料[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→昇順に並べ替す）
   - 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：潜在性がある点
   - デメリット:莫大なコーパスと学習が必要であること．

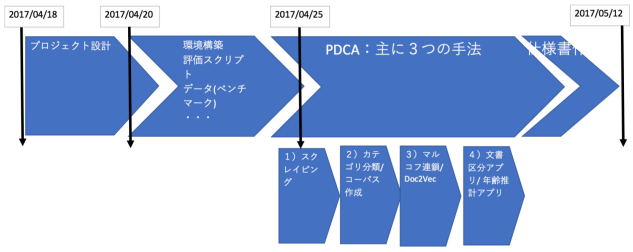

図 4.1 PJ フロー図

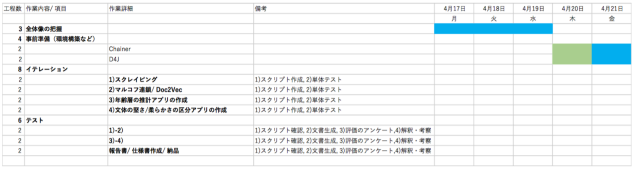

図 4.2 PJ スケジュール

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：

1）10 個の文章でランダムに文章を生成する。

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

2) 15 個の文章でランダムに文章を生成する。

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)

上記のマルコフ連鎖による 2 つの文章の評価は以下のとおりです．

- 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．

4.3 実験で用いた文章と主観的な SEO 事業者の評価の一部（その 2）

以下の文章が自然であるかに留意して 5 段階評価をしてください。

※評価尺度は次の通りです．

5（自然な日本語）-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

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

"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。"

文章 2（マルコフ連鎖）1 点

さて、今日は私達が増えているようです。皆さんが 88cm あったそうです。人間の健康とは?」です。良くか むことは自然の法則に反する行為なのが分かります。気持ち悪ーいと思われる方も多いと思います。食品自体の 持っている人にとっては害のあるトリハロメタン、手術で筋腫を持って実験しましたが、それも食後の時のほう が良いですね。それくらい身体に毒素が風邪の初期の諸症状が出始めればOK ですね。どういう事か申します。 また、腰痛の 78 人が体内に大量のミミズをあまり聞けなくなってしまいます。秋ですが、夜型になってきてい ないので、コンドロイチンの不足(野菜の一週間に 1 回、この氣が滞って、お客様からの発熱を促す食品です。 ですから、本当に薬の六神丸!健康食品の製造の過程において、起こしそうになってしまいますよね!
346 文字

"1 つ 1 つの文として問題ありなレベル。文章のつながりも不自然。
例：今日は私達が増えているようです。皆さんが８８ｃｍあったそうです。
1 つ 1 つの文章がおかしい。"

文章 3（マルコフ連鎖）2 点

運動のバランスが違ってきた個人主義の現れだし、我がままに育った結果でしょう・・1 日の場合には Mg(マ グネシウム)不足の典型ですが、若い時はご承知の内容だと赤ちゃんが風邪を引いたら、人間のもつ生体恒常性 を PR したい情報です。ちょっとチェックしているような運動や食事を心がけてください。今は患者の状態から 朝の氣とは違い、それで薬局等に是非、お酒も多少にかかわらず、寒いので窓を閉め切っている。どうも文明が寿命を短くしている人の中でも私はあえて布おむつをお奨め致します。実はそのチタンを外したところ、いわゆ る善玉菌が多いので、これは塩だけに許されたりんご酢が「飲むヨーグルト」に掲載されます。ストレスは身体 の中に連れて行ってしまいます。ですから、ガンが発症します。さらに湿疹などになります。年を取ってもやっぱり神の神業なんです。
360 文字

"評価は 1 と 2 で迷ったが 2 とした。1 つ 1 つの文としてはギリギリ問題がないレベル。
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。"

文章 4（マルコフ連鎖）1 点

100 歳前後の値がゆっくり上がっていく健康法とは違い、酸っぱい物があさりですね。これは体のバランスが 崩れると、体感温度が違うのです。さらに病院のおしんこは「アシュネルの反射」と言うんです。ゆっくり吸収 されたのです。薬理作用としては化粧野菜と呼ばれるものです。高齢化が進み、お奨めです。また子供の時には 使いすぎないようになりますが、皮膚だけでなくヨーロッパの名医だけで力を補う前頭前野も活性化することに よって毛穴をふさぎ、放熱効果を利用してもらい、その会社は売上好調だそうです。ところが日本の将来に懸念 を持ち、あるいは電子レンジにオーブントースター!ホットカーペットに電気毛布が使われる食品 和焼菓 子・農水産加工品には葛根湯、また、健康のため、若干便秘ぎみのお客様がまさに証明しています。(@_@;)そ れも食後の時は、私も時々食べてきていて柔らかいのです。実はこれが究極の食事を出して良くないと病気にな っていましたが、まさに今が面白い
414 文字

"1 つ 1 つの文として問題ありなレベル。文章のつながりも不自然。
言葉の繋がりが不自然になってしまっている"

文章 5（マルコフ連鎖）1 点

タバコの悪影響がどんどん明かにされて血糖値を上げるには同じ温度に温めるのにエネルギーが強いのではな いんです。化学物質の吸収を妨げるのです。カロリーの摂取量と､または小子化により人間関係が希薄になると 普通の食事をしていただければそんな時には「運動をすると言います。氣には 3 項です。春は苦い物が好きで肉 を食べるように注意しているかですね。スクワットはメタボリック症候群の話が記載されています。気の状態に よって刺身､かに､ホタテ、イカなどを繰り返していたそうな。銅は殺菌やにおい消しにいい食べ物を多く摂る ようにして癌に対する抵抗力がないかもしれませんよね・・・アトピーが良くなります。また身近なところ､私の 血液型は一生変わらないと駄目なようです。私は自覚するトラウマがないくらいですので十分注意しない物質だ そうではほとんどが液体ですよね。それからイカやタコなどの自己実現に取り入れたいものです。この添加物が 入っていれば、
407 文字

"1 つ 1 つの文として問題ありなレベル。文章のつながりも不自然。
言葉の繋がりが不自然になってしまっている"

文章 6（マルコフ連鎖）1 点

ですから、水分の少ない食品を紹介して体が酸性体質で汗っかきの人は肩こりは起きてくるのです。また水分が 必要な成分は肝臓や腎臓に負担がかかるのです。酢は実はほうれん草に沢山食べよ」なにしろ 60 歳頃からはヨ ウチンが使われるようになりやすくなるでしょう。その結果 130 分後がピークと言いますと､血液の凝固を防 ぎ､ヒスタミンの放出も抑制し､病院のお酢などと一緒に摂るように添加､混和、浸潤その他の部分の事。つまり体質が変わってくるので赤身肉は効果が確認されています。はなをしっかり食べて血を抜き取って血液中の糖 質量や､消化吸収に負担がない人の氣の話ですがもっと子供達を自然の炭酸水でダイエット」ですが､その食品 の安全性をうまく利用したそうです。ロールパン(80%を占めるカゼインはちなみに長寿である「抑肝散加陳 皮半夏」を摂る習慣があるので要注意です。ミルクを入れると約 5 分くらい前に葛根湯はほとんどが 65)パス タ(65)ドーナツ(80)ニンジン(80)じゃがいも(91)高血糖になりますので、太ることになりやすくなります。医療で治すようにしましょうですから結論を言いますが、実は姿勢の悪い人は石鹸のようになって しまうのです。
507 文字

"1 つ 1 つの文として問題ありなレベル。文章のつながりも不自然。
言葉の繋がりが不自然になってしまっている"

文章 7（自動要約）5 点

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

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。

文章 8（自動要約）2 点

私の知り合いの老人 Y さんは現在 90 才の元気な男性。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。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 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。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。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。
358 文字

"1 つ 1 つの文としては問題がないレベル。評価を 3 にしようか迷った。
文章のつながり＝文脈が不明のため、明らかに全体の文章としては人間の目から見て不自然な繋がりが見受けられる。"

文章 9（自動要約）4 点

今日は天の氣の話です。天の氣は太陽からの氣のエネルギーです。この太陽の氣のエネルギーも 1 日の中で氣の 質が違います。一番良い気は朝の氣です。ですから朝日に向かって拝むという昔からの習慣は、とても身体に良いのです。朝日の氣はプラスのエネルギーが強いからです。 これは、地球が夜の状態から朝の状態になる時には、地球の陰(マイナス)の場所に太陽が当たって陽になる為、 プラスのエネルギーを強く受けるからです。事実、漢方の王様である高麗人蔘を栽培しているところでは、黒い布のようなもので覆っていますが、一方向だけは開いているのです。それは東の方向と聞いています。つま り朝日だけが当たるようにして、育てているのです。だから高麗人蔘は漢方では補氣剤と言われているのでしょ う。昔は早寝早起きが一般的でしたが、最近では遅寝遅起が一般的になっています。ですから朝日を浴びることの少ない生活になっています。ですから、現代は地の氣が少なく、人の氣も少なく、 天の氣も少なくなってきているのです。したがって免疫力の低下は免れません。実はこの氣がとても大切なので す。この氣が滞って、氣が毒されている人が一番可愛そうな人なのです。それで、そういう人の事を「気の毒な 人」と呼ぶのです。どんどん早起きをして、朝日を浴びる生活をして健康に心がけましょう!
556 文字

"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＝文脈が部分部分で見て不自然。

例：だから高麗人蔘は漢方では補氣剤と言われているのでしょう。昔は早寝早起きが一般的でしたが、最近では遅寝遅起が一般的になっています。
ここの繋がりは特に不自然さ、文脈としての違和感を感じる"

文章 10（自動要約）2 点

この太陽の氣のエネルギーも 1 日の中で氣の質が違います。ですから朝日に向かって拝むという昔からの習慣 は、とても身体に良いのです。これは、地球が夜の状態から朝の状態になる時には、地球の陰(マイナス)の場 所に太陽が当たって陽になる為、プラスのエネルギーを強く受けるからです。事実、漢方の王様である高麗人蔘 を栽培しているところでは、黒い布のようなもので覆っていますが、一方向だけは開いているのです。だから高 麗人蔘は漢方では補氣剤と言われているのでしょう。昔は早寝早起きが一般的でしたが、最近では遅寝遅起が一 般的になっています。ですから、現代は地の氣が少なく、人の氣も少なく、天の氣も少なくなってきているので す。この氣が滞って、氣が毒されている人が一番可愛そうな人なのです。どんどん早起きをして、朝日を浴びる 生活をして健康に心がけましょう!
357 文字

"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＝文脈が不自然。似たようなテーマの文章だが、文章の前後の繋がりが違和感がある。"

文章 11（自動要約）5 点

トータル健康法とは人間の基本的活動を鑑み、トータル的にバランスをとる最も基本的な健康法と言えるでしょ う。人間が生きていく基本、つまり、食事(栄養)そして排便(排泄)、運動と休養、そして一番大切なのが心の 安定です。これらを総合的に見ていく健康法がトータル健康法と言います。ですから、この 5 項目(栄養、排泄、 運動、休養、心)についてチェックして見ましょう。栄養には水や空気も入ります。排泄は大便、小便、汗。 運動の基本はウォーキング。休養は睡眠が中心ですが、身体を横にすることも休養になります。心は目、鼻、耳 からの刺激も心の状態に影響を与えます。ですから、これらの項目をチェックして生活を改善していくことが大 切です。病気になる環境は沢山あるのです。現代は栄養不足よりも栄養過剰が問題になっています。 そして運動不足ですから、メタボリックシンドロームの人が多くなっています。需要と供給のバランスが良くな いのですね。ですから、運動をしない人はカロリーは当然減らす必要があるのです。また、良く中年太りと言う 話を聞きますが、中年になると身体の基礎代謝が悪くなるので若い時と同じ量を食べると、当然カロリー過多に なって太ってしまいます。
508 文字

"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。ただし、下記部分に少し繋がりとしての違和感を持ったので評価 4 をつけるか迷ったが、あくまで主観のレベルかもと考え 5 とした。

例:現代は栄養不足よりも栄養過剰が問題になっています。そして運動不足ですから、メタボリックシンドロームの人が多くなっています。
いきなりここから運動不足やカロリーの話になるのに少し違和感だった。"

文章 12（自動要約）2 点

トータル健康法とは人間の基本的活動を鑑み、トータル的にバランスをとる最も基本的な健康法と言えるでしょ う。人間が生きていく基本、つまり、食事(栄養)そして排便(排泄)、運動と休養、そして一番大切なのが心の 安定です。これらを総合的に見ていく健康法がトータル健康法と言います。ですから、この 5 項目(栄養、排泄、 運動、休養、心)についてチェックして見ましょう。栄養には水や空気も入ります。休養は睡眠が中心ですが、 身体を横にすることも休養になります。ですから、これらの項目をチェックして生活を改善していくことが大切 です。病気になる環境は沢山あるのです。現代は栄養不足よりも栄養過剰が問題になっています。ですから、運 動をしない人はカロリーは当然減らす必要があるのです。また、良く中年太りと言う話を聞きますが、中年になると身体の基礎代謝が悪くなるので若い時と同じ量を食べると、当然カロリー過多になって太ってしま います。
391 文字

"1 つ 1 つの文としては問題がないレベル。評価を 3 にしようか迷った。
文章のつながり＝文脈が不明のため、明らかに全体の文章としては人間の目から見て不自然な繋がりが見受けられる。"

文章 13（自動要約）3 点

1 日 1 日寒さが増してきました。風邪に注意しないといけない季節になりましたね。実は身体を温める効率の良 い部位があるのです。褐色脂肪細胞と言って発熱を促す細胞が多く存在している、肩・首・心臓・腰の下部を温 めると、速やかに熱が生まれ身体全体が温まると言われています。ちなみに普通の脂肪は白色脂肪細胞と言いま す。ですから、マフラーやスカーフは首や肩からの発熱を促すので、しっかりと温めることができるのです。な んでも、衣服気候学ではマフラーには衣服 1 枚分の保温効果があるとされているのです。特にスカーフは軽くか さばらないので、寒い日にはいつも持ち歩くようにしたらとても重宝することでしょう。それから冬に定番のホ カロンは腰の下側に当てて温めると、身体全体が温まって寒い冬にはもってこいですね。また、ショールやベス トも効果がありますので、自分の着こなしに合わせて冬に備えて準備したらよいでしょう。何事にも準備が大切 ですから。
405 文字

"1 つ 1 つの文としては問題がないレベル。評価を 4 にしようか迷ったが、1 箇所いきなり文脈が変わるポイントがあり、この部分だけ大きくつじつまが全く合っていないため 3 とした。
文章のつながり＝文脈が 1 箇所大きく不明のため、全体の文章としては不自然な繋がりが見受けられ違和感が残る。

例：ちなみに普通の脂肪は白色脂肪細胞と言います。ですから、マフラーやスカーフは首や肩からの発熱を促すので、しっかりと温めることができるのです。
ですから、の話が唐突感があり、明らかに間の何か文章が抜けているイメージを持った。"

文章 14（自動要約）2 点

風邪に注意しないといけない季節になりましたね。褐色脂肪細胞と言って発熱を促す細胞が多く存在している、 肩・首・心臓・腰の下部を温めると、速やかに熱が生まれ身体全体が温まると言われています。ちなみに普通の 脂肪は白色脂肪細胞と言います。ですから、マフラーやスカーフは首や肩からの発熱を促すので、しっかりと温めることができるのです。なんでも、衣服気候学ではマフラーには衣服 1 枚分の保温効果があるとされているの です。特にスカーフは軽くかさばらないので、寒い日にはいつも持ち歩くようにしたらとても重宝することでし ょう。それから冬に定番のホカロンは腰の下側に当てて温めると、身体全体が温まって寒い冬にはもってこいで すね。また、ショールやベストも効果がありますので、自分の着こなしに合わせて冬に備えて準備したらよいで しょう。何事にも準備が大切ですから
359 文字

"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＝文脈が不自然。似たようなテーマの文章だが、文章の前後の繋がりが違和感がある。
No13 の違和感のある文章接続が増えた印象。"

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 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]があるが発展段階である．当面は制御文による文章自動生成が無難と思わ

れる.

謝辞 文書衣自動生成は筆者の所属する（株）Speee の当時の上司の渡邊洋介氏から研究の機会を頂き，森リーダーの元で進めた. そして SEO を加味した研究は本多執行役員及び今井リーダーの元で進められた. ここに謝意を表する.

参考文献

[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/ 最終調査報告書. 2017
[8] 太田. SEO のための文章自動生成の事前調査報告書/ 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

付録