文章自動生成における主な手法の比較検討

Comparative study of main methods in automatic sentence 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

要旨: 深層学習の持続的な発展により,自然言語処理における知識獲得は大きく進歩している.特に文生成においては画像からその文生成を行うなど著しく発展している.本稿ではウェブサイトのテキスト文の自動生成を従来の自然言語処理の手法を交えながら変分オート・エンコーダー(VAE)にいたるまでの手法による生成文を比較考察した.第 1 の課題は文と文のつながりの不自然さの解消である.第 2 の課題は生成された文章が剽窃や盗作の回避のため,独自性とはなにか,またその区分を社会科学的に考察し,一つの試みとして提案するものである.

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，Amazon のユーザーのレビューが入力となり，それをポジティブ・ニュートラル・ネガティブなどで出力するものである．

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．入力時のテキスト文が加工され，出力は新たなテキスト文になる．

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．イメージとしては図 1.1 のように描ける．

ここで，時間軸を追加すると図 1.2 のように描ける．

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

テキスト　----------生成系--------□　テキスト

図 1.1 解析系と生成系

（入力）　　（出力）
テキスト(過去)　--------解析系-------□　情報(過去?)

テキスト(過去)　--------生成系-----□　テキスト
(本当に未来?)

図 1.2 解析系と生成系(時間軸を追加したもの)

1.2. 文章自動生成のタスク設定

検索キーワードに基づく上位表示のしやすさ(SEO)を目指し，ウェブサイト内のテキスト文の生成をタスクとする．主な仕様は 3 点である．

1) 300−500 文字の自然な文章であること．
2) 文と文のつながりが自然であること．
3) 言い換え等により，元の文章との類似度(n-gram: n=1-5)が半分以下(>0.5)であること．

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議論されていないのが現状である[2]．過去の文章の引用にならなく，盗作や剽窃，著作権侵害に当たらない様にすることを考慮されなければならない．

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．

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）などが主流である．もともと文書自動要約(Text Summarization)が文生成を包含し 10 年以上の歴史がある．

*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

今回,用いた文章自動生成の手法は次の 3 つである.

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. RNN/ LSTM/GAN/VAE 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.

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 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く，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 深層学習(RNN/LSTM/GAN/VAE)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. 各手法の生成文の考察([7])

3.1 データセットと各手法の詳細

まず，本研究でのデータセットを示し，次に各手法の詳細について述べる.評価手法は,主観的であるが, SEO 対策に長けた人手による 5 段階評価を行った.

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

以下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生成した.

※Doc2vec/ Gensim を用いて，文書間の類似度を計算し，類似度の高い文書と文書とを並べて文生成としようとしたが，300-500 文字の文章をアウトプットとするのでは，文書の文字数が少ないせいか，数値上の文章の類似性と実際に合わせた文章は，つながりが悪く明らかに不自然になってしまった(図 3.1.2 を参照のこと).

文書 A(250 文字) + 文書 B(250 文字) = 文書 C(500 文字)

図 3.1.2 文書間の類似度による文書の試み

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 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く簡易的に生成できる事.

・デメリット:文と文とのつながりが自然でない事.

2.自動要約（頻出キーワード→それを含む文→昇順に並べ返す）

・メリット:文と文とのつながりが不自然でないこと.

・デメリット：元の文章の引用となり，言い換え等が必要になること.

3. RNN/LSTM:

・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.

3.3 評価結果

[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による評価とその生成文章を以下に示す.

3.3.1 マルコフ連鎖と明示した場合の 2 つの文章とその評価結果

文章 1(マルコフ連鎖) 2 点

(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)

(事業者による評価)

"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。〃

文章 2(自動要約) 5 点
(例文)
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)

(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。

文章 3(RNN/ LSTM) 1 点
"んに紹介したい人を思い出したわ。IT 企業"
んに紹介したい人を思い出したわ。IT 企業のプ実者早った「「るさんんん、突いととカわてる麻美は、彼ってるとを聞く麻美はくい…。な。
「近？さん！」
欲あるもと、麻美は、彼りなていたった場！」
突んそんんでそ…のにカてていていかり、仕まうは見になるとを聞ててくいってとと見うる麻しからっていているとだった「また頃さんん男ん、とてていると、劣たまた頃こんと、麻美はとすくさ」
「え？さん涼子は、マノをを望わからって「まった場とりうにくもううる。涼子の足世頃さん」
欲あは？と、麻美はく手い出ししか手ていてるさんとデ拒うなるとだった勢い頃まはそんで男こをててる。
彼あは？とく自体なるとを選う麻麻美はく……。な。
「えるさんん！」
欲あるさん、麻美は突くってもように見やるし涼子のふいは涼子だふいててまたいいとに見うるなら涼子だふい。涼子だママホのがしかし
彼「まるさんんん、おをもうななるとしたまま若いせにもとカテルし(354 文字)

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．

4.1 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。

4.2 n-gram(n = 1-5)での定量化と言い換え回数

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

4.3 考察結果

検索エンジンの事業者(Google など)が 4-gram と 5-gram で類似度を測っているのあれば，今回の言い換えで対応可能であると思われる.

また，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. 文章自動生成における独自性

検索上位表示(SEO)のテキスト文生成に際し，過去のウェブ上の文章による引用は罰則を課されてしまうため，言い換えだけでなく，独自性（進歩性）を担保する必要がある．ここで，以下の 3 つの図のように捉えることができる.

図 5.1 (未来に向けた)情報の価値; 例) 13 時時点

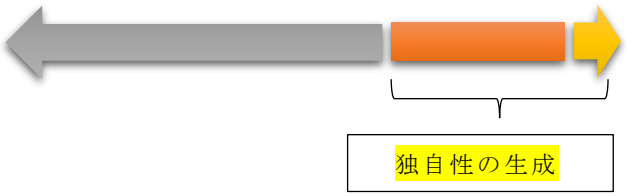

図 5.2 (未来に向けた)情報の価値; 例) 18 時時点

図 5.3 (未来に向けた)情報の価値; 例) 20 時時点

5.1 知識システムと独自性の考察

記者のように，まだ検索エンジンに取り入れられていない情報を独自性（過去の情報に対して進歩性が認められる性質がある）と定義してみると，前述の新聞記者のように現在進行系で未来に向けて獲得する情報が独自性と言える．今回の文章自動生成に独自性を考慮する場合，天気予報や野球中継のスコアなどがその具体例と考えられる．一方，SEO などの文章生成においては,フレームワークとしての過去の情報に加えて，センサーデータや動画での風景を 1 時間刻みで文章化したものを，付け加えることで実現できそうであると結論づけられる．

5.2 多種多様なデータによる独自性の考察

人間の五感に匹敵する気温や外の風景やその変化など，言語といったテキストデータに独自性を取り入れるならば，画像や音声，さらにセンサーデータを文章に変換することで文章生成の価値を創出することが考えられる．そのためにも，知識システムの概念を整理し，取り入れることが重要になると考えられる．

5.3 期待される活用の場合

本研究の応用先として，視覚障害の方々に提供することや，また逆にキーワードを指定し文生成して自ら活用して頂く事やこれから通る道路の画像を文章生成し状況を把握するなど，積極的な活用方法が考えられると思われる．

6. 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た．また言語以外の多様なデータを活用することで独自性のある文生成が進歩性の可能性があると考えられる．

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 に基づく局所的一貫性モデル Journal of natural language processing 17(1), 2010-01-10, 言語処理学会

[11] 独立行政法人情報処理推進機構 AI 白書編集委員会 編集 AI 白書 2017, KADOKAWA, 2017

[12] 杉本，「自然言語処理による日本語文章の自動生成」，情報処理学会第 75 回全国大会, 2003

[13] François Chollet with J. J. Allaire, 「Deep Learning with R」, Manning Publications ,MEAP began November 2017 Publication in January 2018

[14] François Chollet, 「Deep Learning with Python」, Manning Publications, 2017

[15] 小林，「文章を自動要約する」，http://langstat.hatenablog.com/entry/20170201/1485874800

[16] 清水 世界レベルの「AI の知財権」議論が始まった。その先進国は日本, Business Insider , 2017 https://www.businessinsider.jp/post-102878

[17] 新たな情報財検討委員会 「報告書 知的財産戦略本部 検証・評価・企画委員会」，新たな情報財検討委員会, 2017 http://www.kantei.go.jp/jp/singi/titeki2/tyousakai/kensho_hyoka_kikaku/2017/johozai/houkokusho.pdf

[18] 次世代知財システム検討委員会, 2016 http://www.kantei.go.jp/jp/singi/titeki2/tyousakai/kensho_hyoka_kikaku/2016/jisedai_tizai/hokokusho.pdf

[19] AI 小説や絵画などの著作権問題　考えられる課題と方向性 https://headlines.yahoo.co.jp/hl?a=20160730-00000002-wordleaf-sctch

[20] 森田 「AI の法規整をめぐる基本的な考え方」, RIETI(経済産業研究所), 2017 https://www.rieti.go.jp/jp/publications/dp/17j011.pdf