自然言語処理と文章自動生成

-文と文とのつながりを課題として-

2017年7月1日（土）16:00-16:50

株式会社Speee デジタルコンサルティング事業本部

太田 博三

db analytics showcase

Sapporo 2017

1. 背景

- 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，文章生成の主な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.

1.1 自然言語処理の研究とその区分

解析系の研究とは，Amazonのレビューなどのポジ・ネガ判別

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.

（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①

システム開発を命じられた．主な仕様は下記の3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．

＜懸念点＞

- 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い．
- 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

1.2 文章自動生成の入力の問題設定とその難しさ②

- SEO自体がグレーゾン.

- 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.

- 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.

1.3 文章自動生成の注目度

- 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.

- Cf. 文書自動要約(Text Summarization)は 10年以上前から盛ん.

2. 本研究で用いた手法について

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つの手法に集約できると思われる．

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

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.

- Refere to this !

マルコフモデル ～概要から原理まで～ （前編）

http://postd.cc/from-what-is-a-markov-model-to-here-is-how-markov-models-work-1/

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
- 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.
- 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.

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.
- 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
- 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が必要になるなど，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. 社内での実験結果 （一部）

・得られた各手法と好ましいと思われる文字数

- 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
- 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. 実験で用いた文章と評価の一部（その1）

- 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。

酸ガと刻
やーこ一が、か鹸ケルよ
剤オる、すを加デラせ
定ズすうま体添ボトます。
安一造よい身無や肌れま
、タ製くて日。ルのしり
剤、クて届っ毎ねオ分もあ
腐ドっにな。いタ自かが
防。よ元にすさンにる加
をたに手りまだ口合す添
水し法お作いくイ具がの
粧まスたたててナうじど
化しセましれめ。い感な
まとロ。にさかすとた感
まりプすム奨確でっ激
のとドま一推を物すま刺
そっルきリが身禁ま締
ねずコに一導ク方中はいきす。
さかと常下選品旨れにり引
さつ料正いな粧、わ的あ
だ材が良う化て使実もと
くかの謝もよ礎っく現と
ててで、品代ての基補よはこ
立響化新、陳と次のにてれるこ
泡影加（と、は達分しこな
での添す性品自を材い、に
など無ま全製様で、油のさ基なが
なのり安様で、油のさ基なが
水鉛クかい仕と。品こ乾燥
粧亜ッわ早菌こ…。用起
化ニがも抗すで…。用起
ア
ね？

361文字

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/		

4.1 実験で用いた文章と評価の一部（その2）

- 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.

1. 元の文章のままではない点が評価できる.

2. 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.

3. 内包表記などで工夫できれば簡易的で良い.

4.2 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）

- 次のスライド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.

（自然な日本語）5-4-3-2-1（機械的な日本語）

次のスライド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。

- **文章1　__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

SEO事業者の評価の例

- **文章1（マルコフ連鎖）2点**

- "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。"

次のスライド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。

- **文章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。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

SEO事業者の評価の例

- **文章2（自動要約）5点**
- "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。"

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

- 今回の講演は尊敬する中村良幸さんにお声がけ頂き，実現したものである．中村さんにはPythonプログラミングのみならず，多くのパワーを頂いている．ここに謝意を表する．

ご清聴、どうもありがとうございました.