日本語の文章を機械的に作るには

How to Make Japanese Text Algorithmically

名古屋大学大学院工学研究科教授
佐藤　理史

京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005 年より現職。現在、言語処理学会会長。

1 コンピュータにブルーバックスを書かせる

昨年、ある会議の分科会に呼ばれて話をした。その分科会の名称は「人工知能は科学技術論文を理解できるのか？」。依頼のメールには、「人工知能が科学論文を読んで理解し、ブルーバックスを 1 冊書けてしまうような理解度をもち、ストーリーテラーになれないかということを議論したい」と書かれていたので、それに対する考えを準備して、当日の分科会に臨んだ。

そこで驚いたのは、出席者の多くが、「近い将来、それが実現できる」と思っていたという事実である。主催者が掲げた根拠は３つ。

(1) 科学技術論文の要約サービスが存在――PaperDigest（https://www.paper-digest.com）

(2) 科学技術論文執筆支援サービスの存在――SciNote（https://scinote.net）

(3) 小説を作るプログラムの存在

最後の根拠は、我々の研究プロジェクト「きまぐれ人工知能　作家ですのよ」による「日経星新一賞」への応募であった。

よくよく話を聞いてみると、この分科会の首謀者のひとりが思い描いていた青写真は、次のようなものである。ある研究者が書いた一連の研究論文を AI に入力する。そこから、その研究者の研究のエッセンスを抜き出し、それをブルーバックスのような一般の方にもわかるような形で一冊の本にまとめる。

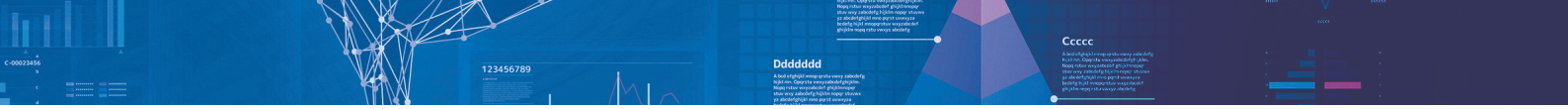
「AI と言った途端に漂う万能感」。これは、最近お会いした新聞記者の方が使ったフレーズであるが、多くの出席者が実現できると思っていた理由は、ここにあるように思う。出席者のほとんどは色々な分野の研究者だったのであるが、AI という言葉は、SF や映画のイメージが強すぎるのであろう。最近の報道も、このイメージの定着に拍車をかけているように思う。

では冷静に考えて、先ほどの青写真のどこにネックがあるかといえば、多くの言語処理研究者は「一冊の本にまとめる（文章を生成する）部分」と答えると思われるが、私は逆である。「研究のエッセンスを抜き出す（論文を理解する）」という解析側の実現の方が遠いと考える。もし、論文を理解した結果を適切な形で出力する解析システムがあれば、おそらく生成システムは作れるのではないかと思う。

2 日本語の文章を機械的に作るには

人間がどのように文章を紡いでいるのかは、自省してもよくわからない。その理由は、入力をうまく規定できないからである。今、私は、この文章を書いているが、それに何らかの明確な入力があるわけではない。こういう文章を書こうという方針はあるが、その内容はどこからともなく湧いてくるだけである。

あまり意識されないことではあるが、文章には何らかの目的がある。多くの場合、それは情報を伝えることであるが、読み手の心を動かすことが目的の文章もある。前者の典型例は新聞記事、後者は小説である。文章生成システムは、目的に沿った文章を作らなければならないが、それはシステムの中から湧いて出るわけではない。



つまり、何らかの形式で外から入力しなければならない。

それを自然言語（文章）で入力するのであれば――それが完成された文章なら、文章生成システムはいらない。それを何らかのデータの形で入力するのであれば――これが、現在の研究のひとつの方向性である。データは、結局、数字を含む記号列であるから、これは一種の翻訳（メディア変換）である。これら以外に、何らかの装置を使って脳の状態を読み取るといった方法も考えられるが、実用的に使えるレベルに達するのは、まだまだ先であろう。となると、文章生成システムの立ち位置は、はっきりしてくる。文章の断片を入力して完成した文章を出力するか、あるいは、テキスト以外のメディアで記述されたデータをテキスト化するか、である。現在の研究の主流は後者であるが、私が取り組んでいるは前者である。

さて、何らかの入力の形式が定められていると仮定して、文章生成を機械的に行うために必要なことを突き詰めていくと、私の答は次のようになる。

① 語から文を組み立てる

② 文から文章を組み立てる

3 文を組み立てる

日本語の文は、述語をひとつだけ持つ単文と、２つ以上の単文から構成される複文に分けられる。

つぎのような例文を考えよう。

(1) 本発明は上記問題点に鑑みてなされたものであり、単純かつ文法的に正しい述部の言い換えを行うことができる述部正規化装置、方法、及びプログラムを提供することを目的とする。（出典：特許第 5585961 の詳細説明）

この文には、次のような６つの単文が含まれている。ここでは、ゴシック体を述語とみなす。

(2) 上記問題点に**鑑みる**

(3) ――**なされたものである**

(4) 単純かつ文法的に**正しい**

(5) ――述部の言い換えを**行うことができる**

(6) ――述部正規化装置、方法、及びプログラムを**提供する**

(7) 本発明は――ことを**目的とする**

複文（1）を組み立てるためには、まず、これらの単文を組み立てなければならない。しかし、ここまで分解すれば、それらは、次のような述語構造テンプレートに語を当てはめれば組み立てられることがわかる。

(8) ～に**述語**

(9) ～は**述語**

(10)～を**述語**

次に、複文の組み立てを考えよう。(4) から (6) の単文は、それぞれ次の単文のヲ格要素を連体修飾していると考えることができる。日本語文法では、(4) と (5) は連体節、(6) は「こと」を含めて補足節とみなすのが一般的であるが、形式的にはいずれも連体修飾である。一方、(2) は、副詞節（テ節）として (3) に結合される。これは連用修飾の形式である。最後に、(3) は、並列節として主節 (7) に結合される。これは列挙の形式であるが、連用修飾の一種とも考えることができる。このように考えると、複文は、連体修飾、連用修飾の形式によって、単文から構成されることとなる。

このように文の構造を分解して捉えれば、比較的長いこの複文でも、機械的に合成することが可能となる。ただ、機械的に合成できるとしても、何らかの意味で省力化が達成できなければ、生成システムを作る意味がない。そこで、文の目的を「発明の目的を伝えること」と固定し、次のようにテンプレート化する。

(11)本発明は［**どのように**］なされたものであり、［**何**］を目的とする

こうすれば、次のようなフィラーを用意してテンプレート (11) を埋めれば、複文 (1) を生成することができる。

(12)**どのように** ＝ 上記問題点を鑑みる

(13)**何 ＝［どんな］**ことができる［**何 2**］を提供すること

(14)**どんな** ＝ 単純かつ文法的に正しい述部の言い換えを行う

(15)**何 2** ＝ 述部正規化装置、方法、及びプログラム

ここで重要なことは、テンプレートを埋めるフィラーもテンプレートを使って合成できるようにすることである。上記の例では、文 (11) の［何］を、(13) から (15) を使って合成できることである。さらに、(12) の末尾を「鑑みて」ような適切な形式に変換することも必要となる。

4 文章を組み立てる

所望の文が組み立てられるとして、それらを使って文章を組み立てることを考えよう。

文にはその構造を規定する制約があり、それが文法としてかなりよく研究されているが、残念ながら、文章の構造に対する制約は、十分には明らかにされていない。しかしながら、文章の種類を固定すれば、多くの場合、文章の典型的な流れは定まるので、これをモデル化すればよい。

我々は現在、通販商品のテレビ CM のシナリオの自動生成に取り組んでいるが、そのような CM の典型的な構造は、次のようになる。

① 視聴者の注意を引く
② 商品を説明し、購入を促す
③ 商品の購入方法を説明する

つまり、このような内容を伝えるパート（段落）をこの順に並べればよいということである。それぞれのパートも、同じように典型的な構造を考えることができる。たとえば、パート③の典型的な構造は次のようになる。

(a) 商品名を再度提示する
(b) オファー（価格や特典）を伝える
(c) 電話番号を提示する
(d) 電話を要請する

そして、それぞれに対して文を作れば、次のような文章を作ることができる。

(16)A 社の「おいしい X」。1 週間分の無料サンプルを、抽選で 1 万名様にお届けします。お申し込みは、0120-XXX-XXX。今すぐお電話を。

結局、段落を作るために決めなければならないことは、どんな目的・内容の文をどのような順に並べるかである。そして、どんな段落をどのような順で並べるかを決めれば、文章が作れる。より長い文章を作るためには、このような組み立てを多段階で行えばよい。

つまり、典型的な構造が存在し、それをモデル化できるのであれば、文章を合成することは可能である。科学啓蒙を目的とした書籍にも、ある種の典型的構造が観察される。原理的にはブルーバックスも生成可能であろうと考える根拠はここにある。

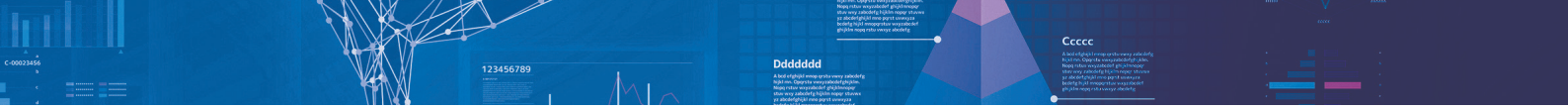
5 定型性が実用性を決める

解析システムと比べて、生成システムの研究は活発とはいえない。多くの応用では、文章生成システムを作るよりも「人間が文章を書いてしまった方が速い」ので、システムに対するニーズが低いのであろう。

現在のコンピュータは、同じようなタスクを何度も繰り返して実行する場合に威力を発揮する。多くの異なるタスクを実行しなければならない用途には向いていない。これを文章生成システムに当てはめると、同じような文章を何度も繰り返し生成する場合にのみ、システムを作るメリットが発生する。

しかしながら、そのような状況は、それほど多くはない。文章を書くことを生業にしている人に限られる。ところが、彼ら・彼女は、文章を書くことが苦にならないので、文章生成システムを必要としない。文章生成システムを欲するのは、彼ら・彼女らに文章作成を発注する側であろう。

どのような実現法をとるにしても、実用的な文章生成システムを作れるか否かは、生成すべき文章の内容と形式の定型性に決定的に左右される。内容の定型性が高ければ、システムの入力形式の設計が容易になり、形式の定型性が高ければ、システムは比較的単純な機構で実現できる。一方、定型性が低ければ、実用的なシステムを作ることは諦めたほうがよい。これが、現在の技術の到達レベルであり、多くの人々が期待する「思ったことを文章化してくれるシステム」というのは、まだまだ幻想に過ぎない。


123456789
Dddddd
Ccccc