コンピュータは文章を書けるのか

Can Computer Write Text?

名古屋大学大学院工学研究科教授 **佐藤　理史**

PROFILE 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、言語処理学会編集委員長。

1 はじめに

現在の自然言語処理研究は、明らかに、テキストの自動解析・自動理解の方向を向いており、テキストの自動生成･自動産出に関する研究は、相対的に少ない。特に、日本語の処理に関しては、その傾向が顕著で、言語処理学会の年次大会では「生成」に関するセッションを１つも組めないのが普通である。

多くの人は、読むことより書くことに苦労を感じているように思えるのに、書くこと（あるいは、その作業の一部）をコンピュータに代行させようという動きが本格化しないのは､どうしてなのだろうか。ほとんどの人は､いまのワープロソフトに満足し、それ以上のサポートはいらないと思っているのだろうか。

読むことと書くことは、書き言葉を扱う能力の両輪である。その両方が実現できて、はじめて、コンピュータは書き言葉を扱う能力を持ったと言うことができよう。我々のグループは、ショートショートの自動創作を目指すグランドチャレンジ「きまぐれ人工知能プロジェクト 作家ですのよ」への参加をきっかけに、2013 年から本格的に文章の自動生成の実現に取り組んでいる。このプロジェクトにおいて我々が追求する問いは、「コンピュータは文章を書けるのか？」である。

2 文章を書くということ

文や文章の自動解析では、入力はテキストで、出力は解析結果である。一方、文や文章の自動生成では、出力はテキストであるが、入力は自明ではない。何を入力とすべきか、それを考えるところから研究はスタートし、それが定まれば、研究はほとんど終了する。

特に、小説のような文章を作ることを考えるとはっきりするのだが、概念的には、「語る内容を考える（作る）こと」と､「それを個別言語(たとえば日本語)で語ること」は別物である。前者は、言語に依存しないが、後者は特定の言語に依存する。たとえば、映画やドラマのノベライゼーションで行われることのほとんどは､後者である。しかしながら、普通に文章を書く場合、前者と後者は渾然一体とした作業となる。少なくとも、私の実感ではそうである。

文章を書くという作業のうち、前者のプロセスがどのようなものなのかは、よくわからない。当面、機械化のターゲットとなるのは、後者である。これを文章化と呼ぶことにしよう。

ちなみに、文章が書けずに困っている場合の多くは、実は前者の部分なのではないかと睨んでいる。日本語の文章が書けないのではなく、何をどういう順序で伝えるか、その内容を決めるのに困っているである。

3 機械的に文章を生成する方法

文章化の部分しか実現できなくとも、実は、多くの潜在的な応用がある。たとえば、

・箇条書きのメモから文章を作成する
・研究発表スライドから論文を作成する
・定型データを文章化する

これらは、広い意味でのメディア変換（他の媒体から

文章への変換）である。特許という文脈では、特許出願書類の生成などがこの範疇に含まれる。

文章を出力する基本的な方法は、雛形（テンプレート）を用いる方法である。定型性が高い文章であれば、この方法で十分なことが多い。雛形には、普通、いくつかの空欄（スロット）があり、これらを適切に埋めると、文章が完成する。

しかしながら、文章の長さが長くなると、この方法は破綻する。第一に、長い文章の雛形を作ることは不可能に近いし、もしそれができたとしても、完全に剽窃となる。機械的に生成したテキストが著作物とみなされるかどうかは将来の問題だが、一定以上の長さのテキストをそのまま流用した場合は、著作権に抵触する可能性が高い。私の調査によれば、小説で使い回される文字列の長さは、最長で 20 文字強であり、それを超えた同一文字列が異なる作品間で見受けられる場合は、剽窃である可能性が濃厚である。

ということは、小説に限らず、文章一般において、それを生成するためには、かなり短い部品から合成することが不可欠である。日本語の書籍における文長の最頻値は 23 文字、中央値は 33 文字であるから [1]、文より小さい単位からの合成が必要である。

4 文生成器と文章生成器

文章生成に対する我々の基本的な道具立ては、文生成器と文章生成器である。文生成器は、文の骨格から表層文を生成するプログラムで、その主要な機能は、機能語の選択と活用形の調整である [2,3]。文章生成器としては、文（または文章）の順序を規定する規則（文脈自由文法）によって駆動される生成器を作成中である。非終端記号・終端記号の属性拡張により、文間の整合性や連鎖をコントロールすることが可能である。

これらのツールは、機能的にはまだまだ不十分である。しかしながら、出力したい文章とそのバリエーションが明確になっていれば、それらを生成するシステムを作成できるレベルには達している。実際に、1200 字程度の文章を生成するデモが動いており、利用できる部品や規則を増強すれば、より長い文章の生成も可能である。

5 文章生成が流行らない理由

現在のレベルは、「コンピュータ『が』文章を生成する」と言えるレベルではない。「コンピュータ『で』文章を生成する」という表現が適切である。

個人的には、文章生成の研究はかなり面白く、当面続けようと考えているのだが、まだしばらくは流行らないだろうとも思う。それにはいくつか理由がある。

(1) 生成した文章を読むのは人間なので、高い品質が求められる。

(2) 短い文章であれば雛形で十分。品質も高い。

(3) 長い文章を日常的に書かなければならない職業人は、かなり限られる。

つまり、ニーズはあるようで、実はあまりないのかもしれないということである。英語では、実際に、企業の決算発表記事を自動的に作成するシステムが用いられ始めているとのことであるが、日本の通信社・新聞社での取り組みは聞こえてこない。小説の自動生成のような酔狂なプロジェクト以外では、日本語文章の自動生成は不要なのであろうか。

参考文献

[1] 刀山将大、佐藤理史、近藤秀、吉田達平．日本語の文の平均像を体現した文を探す（1）文の特徴量の抽出．第 13 回情報科学技術フォーラム（FIT2014）、E-006、第 2 分冊、pp217-218、2014.

[2] 佐藤理史、「文生成器を作る」とはどういうことか．言語処理学会第 21 回年次大会発表論文集、D7-4、pp.1080-1083、2015.

[3] 緒方健人、佐藤理史、松崎拓也．文節木の段階的実体化による日本語文生成器の作成．人工知能学会第 29 回全国大会、3M3-1、2015.