形式言語理論の謎を解く

多重文脈自由言語に対するポンプの補題

金沢　誠（情報学プリンシプル研究系）

何が分かる？

文法の数学的モデルの中で一番基本的なものが「文脈自由文法」ですが、その自然な一般化に「多重文脈自由文法」があります。多重文脈自由文法に関する未解決問題とその部分的解答を解説します。

どんな研究？

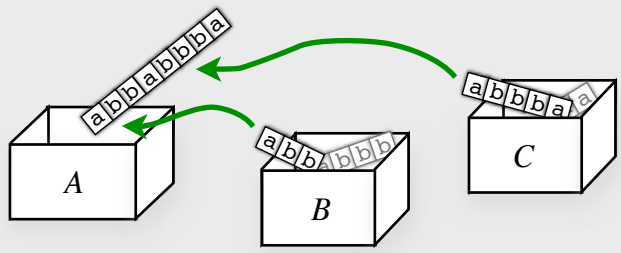
形式言語理論という理論計算機科学の分野に属する数学的な研究です。

文脈自由文法という言葉をご存知でしょうか。文脈自由文法は、人間の言語の文法の数学的モデルとして言語学者ノーム・チョムスキーが1950年代に導入したものですが、プログラミング言語の文法の記述にも標準的に使われています。文法の数学的モデルは**形式文法**と呼ばれ、さまざまな形式文法の数学的性質を研究する分野が**形式言語理論**です。 形式文法は、文字列の集合（**形式言語**）を定義する手段であり、いろいろな種類がありますが、その中でもっとも基本的なものが文脈自由文法です。

文脈自由文法は、直観的には、文字列をいくつでも格納できる箱と、箱の中に文字列を追加するための規則をそれぞれ有限個備えたものです。例えば、

$$A(xy) \leftarrow B(x), C(y)$$

という規則は、「箱Bの中から文字列xを取り出し、箱Cの中から文字列yを取り出して、xとyをつなげてできる文字列xyを箱Aに追加することができる」と解釈されます。



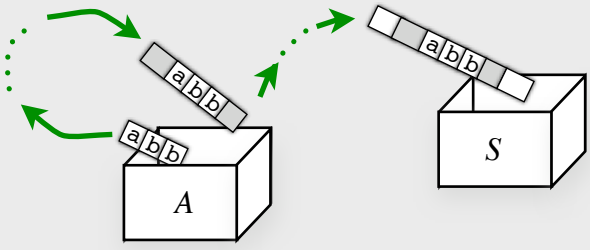
同様に、$A(\mathtt{aab}) \leftarrow$という規則は、「無条件で$\mathtt{aab}$ という文字列を箱Aに追加してよい」と解釈されます。

すべての箱が空の状態から始めて、与えられた文法Gの規則を適当な順番で適用して文字列xをSという特別な箱に入れることができれば、xは文法Gの記述する言語$L(G)$に属することになります。例えば、括弧 [,] の列で、開く括弧と閉じる括弧が正しく対応するように並んでいるものの集合（**ダイク言語**）を記述する文法は次のようになります。

$$S([\,]) \leftarrow$$
$$S(xy) \leftarrow T(x), S(y)$$
$$T([\,x\,]) \leftarrow S(x)$$

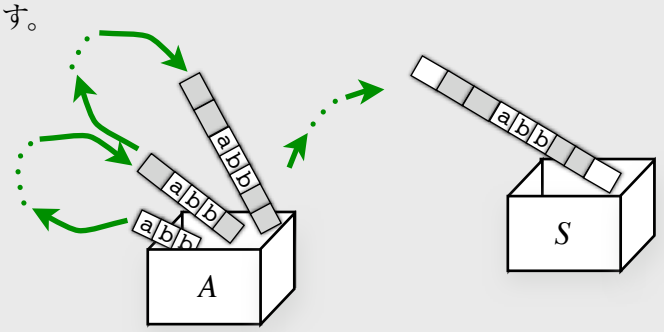
文脈自由文法はその50年の歴史の中で非常によく研究され、様々な性質が明らかにされて来ました。文脈自由文法に関するもっとも基本的な定理のひとつが1960年代に証明された**ポンプの補題**です。

Gを文脈自由文法、文字列xを$L(G)$の要素とし、xがGの規則を適用して作られる過程を考えます。xが長い文字列のとき、必ずこの過程は、ある文字列vが箱Aから取り出され、その後vを含む文字列がAに入れられるという過程を含んでいるはずです。



この途中の過程を省略したり繰り返したりすることによって$L(G)$に属する別の文字列を作ることができ

ます。 ポンプの補題はこの性質をとらえたものです。



文字列xに対して、$x = x_1 x_2 \ldots x_k$を満たす$(x_1, x_2, \ldots, x_k)$をxの**k分割**と呼びます。言語Lに属する文字列xの$2k+1$分割$(x_1, x_2, \ldots, x_{2k+1})$に対して、 偶数番目の要素$x_2, x_4, \ldots, x_{2k}$を$i$回繰り返して得られる文字列の集合

$$\{x_1 x_2^i x_3 x_4^i x_5 \ldots x_{2k-1} x_{2k}^i x_{2k+1} \mid i \geq 0\}$$

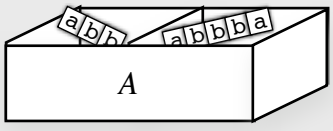
がLの無限部分集合となるように $(x_1, x_2, \ldots, x_{2k+1})$ を選べるとき、xはLに対して**k-ポンプ可能**であると言います。そして、Lの要素が、有限個の例外を除いてすべて、Lに対してk-ポンプ可能であるとき、L自体を**k-ポンプ可能**と言います。文脈自由文法に対するポンプの補題は、次のように述べることができます。

> すべての文脈自由文法Gについて、$L(G)$は2-ポンプ可能である。

ポンプの補題を使って、次のような言語が文脈自由文法によっては記述できないことを証明することができます。

$$\text{COUNT-3} = \{\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n \mid n \geq 1\}$$
$$\text{COPY-2} = \{xx \mid x \in \{\mathtt{a},\mathtt{b}\}^*\}$$

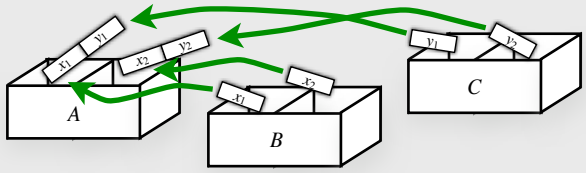
文脈自由文法の自然な拡張のひとつに、1991年に関らによって導入された**多重文脈自由文法**があります。多重文脈自由文法と文脈自由文法の違いは、多重文脈自由文法では、箱がいくつかの部屋に仕切られていることです。



まったく仕切りのない箱も許されます。特別な箱Sには仕切りがありません。箱に文字列を追加するときは、すべての部屋にひとつずつ追加します。箱から文字列を取り出すときは、すべての部屋からひとつずつ取り出し、また、取り出される文字列はその箱に同時に入れられたものでなければなりません。たとえば、

$$A(x_1y_1, x_2y_2) \leftarrow B(x_1, x_2), C(y_1, y_2)$$

という規則は、「箱Bの２つの部屋からx_1とx_2を、箱Cの２つの部屋からy_1とy_2を取り出し、箱Aの２つの部屋にx_1y_1とx_2y_2を追加してよい」という意味です。



多重文脈自由文法を使えば、文脈自由文法では記述できないCOUNT-3やCOPY-2のような言語も記述することができます。COUNT-3の文法は次のようになります。

$$S(x_1x_2) \leftarrow A(x_1, x_2)$$
$$A(\mathtt{a}x_1\mathtt{b}, \mathtt{c}x_2) \leftarrow A(x_1, x_2)$$
$$A(\mathtt{ab}, \mathtt{c}) \leftarrow$$

多重文脈自由文法は、文脈自由文法の記述力を拡張すると同時に望ましい性質を引き継いでおり、人間の言語の記述の他、 RNAの二次構造の予測など、生物配列の解析にも応用されています。

箱の中の部屋の数が最大mであるような文法を**m-多重文脈自由文法**と言います。上の文法は2-多重文脈自由文法です。（1-多重文脈自由文法は文脈自由文法に他なりません。）m-多重文脈自由文法については、文脈自由文法の場合との類推から、次の形のポンプの補題が成り立つと予想されています。

> すべてのm-文脈自由文法Gについて、$L(G)$は$2m$-ポンプ可能である。

しかし、この予想は未だに証明も否定もされておらず、20年ものあいだ未解決の問題となっています。これは数学的に非常に興味深い問題です。関らの1991年の論文では、これを非常に弱めた形の定理が成り立つことが示されています。

私の2009年の論文で、2-多重文脈自由文法に対してポンプの補題が成り立つことが初めて証明されました。 また、$m \geq 3$の場合については、文法が**非交差性**という自然な制約を満たす場合にポンプの補題が成り立つことが示されました。この制約を満たす多重文脈自由文法は、よく知られている他のいくつかの文法形式と等価になることがわかっています。