

逆構文解析を用いた意味からの文章生成

柳沢 孝[†] 三浦 孝夫[†]

[†] 法政大学 情報電気電子工学科 〒184-8584 東京都小金井市梶野町 3-7-2

E-mail: [†]takashi.yanagisawa.st@eng.hosei.ac.jp, ^{††}miurat@k.hosei.ac.jp

あらまし 本研究では、与えられた意味を表現する文章生成を論じる。通常、文章は各部分に対応する解釈の構造化によって表現されるが、当該する意味を表現する「文章」が妥当かどうかの判断は行わない。本稿では、四則式を表現する文章の生成を例にとって、文章生成条件を設定し、その条件と文章との関連性を検討する。

キーワード 逆構文解析, 文章生成

Generating Sentences by Reverse-Parsing

Takashi YANAGISAWA[†] and Takao MIURA[†]

[†] Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184-8584 Japan

E-mail: [†]takashi.yanagisawa.st@eng.hosei.ac.jp, ^{††}miurat@k.hosei.ac.jp

Abstract In this investigation we discuss how to generate sentence to represent given meaning. Generally, sentences are represented by structuring of interpretation corresponding to each part. However it is hard to grasp what sentences are going to say. In this work we propose a method which generate sentences for mathematical formula which has only four arithmetic operations then we put conditions. We show relationship to the condition and sentences.

Key words Reverse Parsing, Generating Sentences

1. 前 書 き

我々が日常生活で触れる大量の情報は、車内放送や電光掲示板のように、大半が順次に流れ去る単語の並び（文字列ストリーム）であることが多い。多くの場合、「2 分の 3 です」のように単純な文章で表される。高度な情報を伝えるためには、精密で複雑な文章が必要となる。例えば、「 $\frac{3}{2}$ 」を伝えるには、「5 分の 3 を考え、それを 2 で割る」のように多くの文章を用いばよい。だが、一般的には受け手に多くの記憶量を課すため、通常は文章構造を複雑にすることになる。即ち、ある文法を構築しその意図を表現するための解析法を探ることになる。「5 分の 3 を 2 で割る」は $\frac{3}{2}$ とも理解できるから、左連想的文法を構築すればよい。では、後者の式はどのように表現されれば良いのだろうか。このように、意図を表現するのに都合の良い文法・文章はどうあるべきか、を論じることはこれまであまり無い。

例えば、コンパイラでは、文字列から意味を抽出するため、文法構造の解析手法を用いる。文法構造の各部をある意図と対応させ、文章全体は各部の構成で表現できるという仮定の下で、コードを出力する。しかし、意図（生成されたコード）を得るための「文章」を得るための研究はこれまであまり行われていない。どのような条件のもとで $\frac{3}{2}$ も $\frac{3}{5}$ も生成できる文法とはどうあるべきであろうか。

本研究では逆構文解析を用いて、与えられた意味を表現する

文章（文字列、トークン列）の生成のあり方を論じる。文章の生成を行うときこれまで提案されている方針として「最小記述原理 (MDL)」、「最尤推定 (MLE)」、「認知心理学」などが想定できるであろう。例えば「2 分の 1」という文字列は「1 を 2 で割る」という意味を「短く」表記する。

本研究では四則式の表現を行う文章の生成を例にとって、逆構文解析の意義と方向性を提案する。また、方針の評価として「文章は適切であるか」、「曖昧性が生じてはいないか」、「演算子量や入れ子のレベルなどにより、元の意図を正しく理解できるか」などの充足性を検証する。

与えられた意味をもつ文章は一般に意味が存在するが、本研究では、「意味制約」と「意味分割手法」を導入する。意味制約とは、単純さを保つため意味記述に課する条件であり、「複雑さ」を定義する。一方、意味分割とは、複雑さを取り除くため、意味記述を、これと等価で複雑でないものに分割すること、及びそれらの関係を生成することをいう。

本研究では、四則式の記述を例にとって扱う。意味制約として四則式内の演算子量、あるいは四則式の複雑さを四則式の深さで定義する。ここでは、意味分割は（木構造として扱われる）数式を四則木とし、これを部分木にわけけることをいう。また、分割した部分を変数記号で関連つける手法を用いる。

本報告では LaTeX による四則式表現の文章生成用文章の生成、変数記号の導入、記号を含む式の変数化などの文章生成条

件を設定し、実験システムを試作して有効性を検証する。

2章では問題点と解法について、3章で手法、4章で評価実験、実験方法、評価方法、実験結果、および考察、5章で結びとする。

2. 逆構文解析

本研究では、逆構文解析による文章生成の効果調べるため四則木を意味とする文章文字列を生成する。すでに述べたように、生成された文章が正しい(文法に従う)ことや曖昧性がない(一意な解釈を有する)といった必要条件だけではない。文章木の深さや演算子量の数、文章を設定し、これに応じて文章の分割や構造の導入を行い、元の意図を理解できるかを検証する必要がある。

ここで四則木に対し深さで分割するための閾値が意味制約条件として与えられた場合、木の根から閾値の深さの演算子のノードを意味分割の対象とする。同様に、四則木に対し演算子の量で分割するための閾値が条件として与えられた場合、木の葉レベルから演算子の量が閾値となるまでの部分木を分割の対象とする。四則木が分割されたとき、もとの関連を保つために変数記号を導入する。

たとえば $\frac{3}{2}$ を、 $\frac{3}{5}$ を表す記号 A を用いて表すとき、 $\frac{A}{2}$ と表現する。

意味分割は意味制約が満たされないときに生じる。本稿では意味制約を、演算子数、木の深さなどの閾値として設定する。意味分割によって、与えられた四則式はいくつかの小さな四則式(部分木)の集合で表される。このときどの部分木も意味の制約を満たし、部分木間は境界ノードが分割場所で関連付けられる。

例えば、意味制約を「演算子の数が2」とする。このとき図1aは意味制約を満たさない。

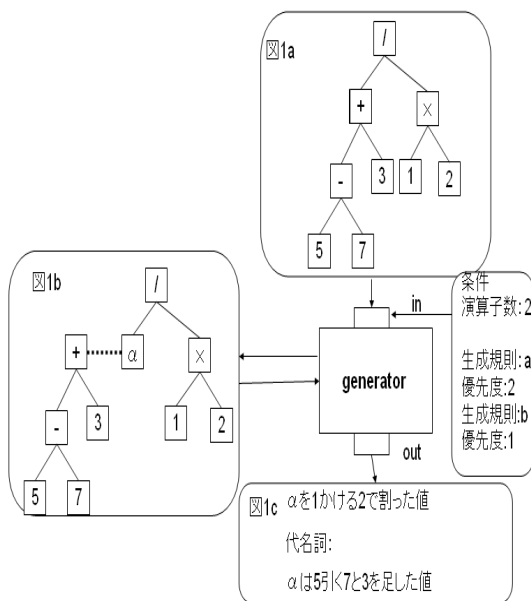


図1 四則木の文章生成

意味分割は制約を満たさないときに必要となる。図1aの葉レベルの演算子から順に以下の動作を行う。この結果を図1bに示す。

演算子“-”から意味制約を満たす限り、上位に部分木を走査する。この拡張を続け、ついに意味制約を満たさなくなったとき当該境界ノードで分割する。このとき、分割した場所 α が部分木との関連を記述する。

文章生成規則は、(部分)木から文章を導出する規則、および部分木の関連を表現する原則から構成される。部分木は、あらかじめ定まったいくつかの生成規則(と優先度)によって文章へ変換される。図1bでは部分木“(5,7)”は「5 引く 7」、あるいは「5 から 7 を引く」として表記される。境界ノードでは、これを関連付けるため変数記号を代名詞のようにして用いる。例えば、「 α は「5 引く 7」」のように意味を置き換える。このことにより、部分木が意味生成を満たしたまま文章を生成することが可能となる。

文章生成を図1bに示す。図1bにおいて、「+」を根とする文章木を A_1 、「/」を根とする部分木を A_2 とする。 A_1 に関して文章を生成するには、優先度の高い規則aを用いる。 A_1 は後置順に処理され、次に規則bが用いられ「5 引く 7 を 3 で割った値」の文章が生成される。同様に A_2 についても「 α を「1 かける 2」で割った値」と表す。この両者の関連は α が分割点を表すことから、

- 1) α を1かける2で割った値
- 2) α は5引く7と3を足した値

がこの順に生成され、意味分割された四則式(部分木)が関連付けられ。

3. 意味分割と文章生成

3.1 四則木の分割

意味制約として深さ、演算子数の2通りを考える。深さは四則式の複雑さを表し、演算子数は大きさを表すことによる。

意味制約の深さは、閾値 d を与えるとき意味制約となる。このとき意味分割手順が必要となる。四則木 T が入力されると、部分木への分割を行う。つまり深さ d を超える四則木が与えられたとき、 T を分割する。さらに、再帰的に分割された部分木にも充足性を調べる。この結果すべての部分木が意味制約を満たすまで分割を繰り返す。

意味制約を深さ $d > 0$ とし、生成される部分木数を n とする。

入力: 四則木 T

出力: 部分木 $\{S_0, S_1, S_2, \dots, S_n\}$

手順:

T の深さ D に対して $D > d$ とする。このとき根から見て、深さ d の部分木を T から分割し、分割ノードを変数で置き換える。同時に当該部分木をその変数記号に関連付ける。残った T についてこの手順を繰り返し、分割された部分木および(分割された後の) T は深さに関する意味制約を満たす。

四則式の大きさに関する意味制約として、演算子数 $k > 0$ に閾値を与える．四則木 T が入力されると、先と同様に部分木への分割生成を行う．すなわち、演算子数が k を超えるとき意味制約を満たさないとし、 T を意味分割する．この手順も分割された部分木に対して再帰的に適用する．

生成される部分木の個数を n とする．

入力: 四則木 T

出力: 部分木 $\{S_0, S_1, S_2, \dots, S_n\}$

手順:

T の演算子数 K に対して $K > k$ とする．このとき葉から見て、演算子数 k となる部分木を T から分割し、分割ノードを変数で置き換える．同時に当該部分木をその変数記号に関連付ける．

残った T についてこの手順を繰り返し、分割された部分木および (分割された後の) T は演算子数に関する意味制約を満たす．

3.2 文章生成規則

一つの演算子に対して文章は複数類を用意し、演算子としての表現が含まれる言葉 (表 1 内の下線部) の位置で 3 つに分け、表 1 で示す．

規則				
数式	規則 a		規則 b	規則 c
$\alpha + \beta$	α	たす β	α と β を 足した値	β を α に 足した値
$\alpha - \beta$	α	引く β	α から β を 引いた値	β を α から 引いた値
$\alpha \times \beta$	α	かける β	α と β を かけた値	β を α に かけた値
$\frac{\alpha}{\beta}$	β	分の α	α を β で 割った値	β で α を 割った値

表 1 文 章 表

規則 a, b, c を組み合わせて文章生成規則とする．

例えば、式 $\frac{3}{2}$ に対し規則 a, 規則 c という順に適用すれば、

”2 で 5 分の 3 を割った値”

のように、下線部の文章に対応する順と適用させる．またこれに対応する生成規則を適用する．

4. 実 験

4.1 実験方法

本実験では、演算子数 (2, 4, 6, 8, 10, 15) を含んだ 6 種類の数式を用意する．

- (a) $\frac{1 \times 2}{8}$
- (b) $(\frac{1}{1 \times 2 + 3}) - 4$
- (c) $(\frac{3 \times 4 + 5}{4}) + (\frac{3}{1 \times 2})$
- (d) $\frac{(3+4) \times 9}{(\frac{5}{1-(7+4)}) - (2 \times 9)}$
- (e) $((\frac{(2-4) \times 3}{1 \times 2}) + 7) - (\frac{2+4}{5 \times 6})$
- (f) $\frac{(2+(4 \times 5) - 9) \times (7 \times 8)}{((8 \times 3) - (1+2)) \times (\frac{5}{4})}$

図 2, 3 にそれぞれの LaTeX としての数式及び四則木を示す．

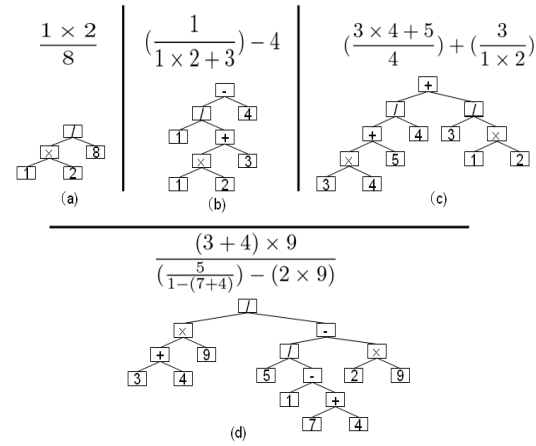


図 2 文章木 (1)

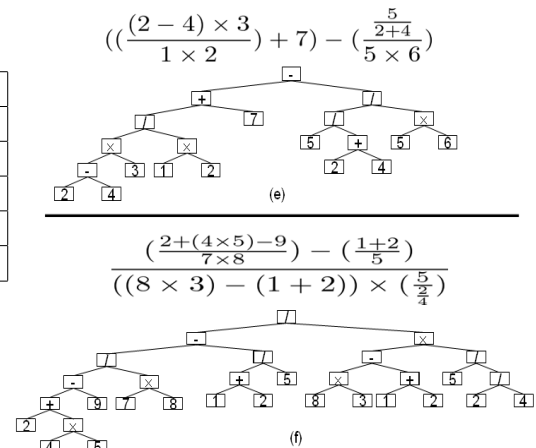


図 3 文章木 (2)

この式に対して、さまざまな意味制約および文章生成規則を組み合わせ、生成される文章を評価する．本実験では、意味制約を、深さおよび演算子数に設定する．このとき上記 6 種類の四則式に以下の項目の実験をする．

実験ア 文章生成規則考慮なし

実験イ 文章生成規則考慮

実験ウ 深さによる意味分割+文章生成規則考慮なし

実験エ 演算子数による意味分割+文章生成規則考慮なし

実験オ 深さによる意味分割+文章生成規則考慮

実験カ 演算子数による意味分割+文章生成規則考慮

この際、実験ウ から実験カ については 閾値を変えて数回行い、検証の際に実際に LaTeX としての式と比較する．

4.2 評価方法

文章評価にできるだけ客観性を与えるため、生成された文章について文字数を比較する方法と、理解しやすさを比較する方法をとる。後者の実験のために被験者を用意し、アンケートの集計を行う。ここで対象とした被験者は、50代:男性:会社員、40代:女性:主婦、30代:男性:会社員、20代:男性:学生、20代:男性:学生の5人である。

4.3 結果(1): 記述長

表2は数値が文字数、括弧内の数字は文の数を表す。この数値は実験(2)で高評価の文章を利用している。

	四則木					
実験	(a)	(b)	(c)	(d)	(e)	(f)
ア	8(1)	14(1)	21(1)	27(1)	34(1)	50(1)
イ	11(1)	27(1)	35(1)	51(1)	62(1)	86(1)
ウ	13(2)	19(2)	26(2)	42(4)	49(4)	65(4)
エ	13(2)	19(2)	31(3)	37(3)	44(3)	65(4)
オ	16(2)	28(2)	40(2)	62(4)	82(4)	110(4)
カ	20(2)	28(2)	43(3)	61(3)	74(3)	117(5)

表2 生成された文章の記述長

4.4 実験結果(2)

評価は、各式に対して実験アを基準とし、”5(大変わかりやすい)”、”4(わかりやすい)”、”3(どちらともいえない)”、”2(わかりにくい)”、”1(大変わかりにくい)”のいずれかを選択する。実験ウからカにおいて数回の実験のうちもっとも評価の高いものを実験結果とする。実験結果を示す。

	評価					
実験	5	4	3	2	1	平均
ア	5	0	0	0	0	5.0
イ	4	1	0	0	0	4.8
ウ	2	3	0	0	0	4.4
エ	2	3	0	0	0	4.4
オ	2	3	0	0	0	4.4
カ	2	3	0	0	0	4.4

表3 演算子数2 式:(a)

意味分割を行う実験ウからカは、平均が4.4、意味分割を行わない実験1,2の平均4.9で後者の方が高評価である。なお全体の平均は4.7である。

	評価					
実験	5	4	3	2	1	平均
ア	0	0	1	3	1	2.0
イ	0	1	2	1	1	2.6
ウ	5	0	0	0	0	5.0
エ	5	0	0	0	0	5.0
オ	4	1	0	0	0	4.8
カ	5	0	0	0	0	5.0

表4 演算子数4 式:(b)

表4において、意味分割を行う実験ウからカは平均が5.0と

意味分割を行わない実験ア、イの平均2.3で差が意味分割を行う実験に偏っている。なお全体の平均は4.1である。

	評価					
実験	5	4	3	2	1	平均
ア	0	0	0	2	3	1.4
イ	0	0	1	2	2	1.8
ウ	4	0	0	0	1	4.2
エ	4	0	1	0	0	4.6
オ	2	3	0	0	0	4.4
カ	1	3	1	0	0	4.0

表5 演算子数6 式:(c)

表5において、意味分割を行う実験ウからカは平均が4.3、意味分割を行わない実験ア、イの平均1.6と表4と同様で差が著しい。意味分割を行う実験に偏っている。なお全体の平均は3.4である。

	評価					
実験	5	4	3	2	1	平均
ア	0	0	0	0	5	1.0
イ	0	0	0	0	5	1.0
ウ	1	1	1	1	1	3.0
エ	0	1	3	0	1	2.8
オ	0	3	0	0	1	3.0
カ	1	2	1	1	0	3.6

表6 演算子数8 式:(d)

表6において、意味分割を行う実験ウからカの平均が3.1、意味分割を行わない実験ア、イの平均1.0で差は出たものの表45と比べると差は少なくなったが意味分割を行う実験に偏っている。なお全体の平均は2.4である。

	評価					
実験	5	4	3	2	1	平均
ア	0	0	0	0	5	1.0
イ	0	0	0	0	5	1.0
ウ	0	1	2	0	2	2.4
エ	0	1	2	0	2	2.4
オ	1	2	0	0	2	3.0
カ	0	1	3	0	1	2.8

表7 演算子数10 式:(e)

表7において、意味分割を行う実験ウからカの平均が2.7、意味分割を行わない実験ア、イの平均1.0で差が生じ代名詞を利用したものに偏っている。また、ここで、文章生成規則考慮を利用した実験イ、オ、カの平均は2.3、文章生成規則考慮をしていない実験ア、ウ、エの平均は1.9である。これは実験全体の平均2.1であるにもかかわらず文章生成規則考慮に偏った結果である。

表8において、意味分割を行う実験ウからカの平均が1.5、意味分割を行わない実験ア、イの平均1.0と差がなくなってきたが意味分割を行う実験に偏っている。また、文章生成規則考

	評価					
実験	5	4	3	2	1	平均
ア	0	0	0	0	5	1.0
イ	0	0	0	0	5	1.0
ウ	0	0	0	1	4	1.2
エ	0	0	1	0	3	1.2
オ	0	0	0	2	3	1.4
カ	0	0	2	1	2	2.0

表 8 演算子数 15 式:(f)

慮を利用した実験イ,オ,カの平均は 1.5,文章生成規則考慮をしていない実験ア,ウ,エの平均は 1.1 と文章生成規則考慮を利用した実験に偏っている.なお,全体の平均は 1.3 である.

4.5 考 察

本研究での実験結果について考察する.記述長で判断すれば,どの数式に関しても実験アは実験カのおよそ半分であり,コンパクトな表現を得ているといつてよい.しかし,容易に想像できるように,これが被験者の理解と必ずしも対応していない.

全体の実験結果の特徴から文章生成規則考慮の実験イ,オ,カと,文章生成規則考慮なしの実験のア,ウ,エを比較する.

文章生成規則考慮全体の平均が 3.1,文章生成規則なしが全体の平均は 2.9 となり,これは文章生成規則を考慮すると理解度の向上が期待できることを示す.特に,表の特徴から式(e),(f)の文章生成規則の適用結果は評価が高い.

例えば,実験カは式(f)について文章生成規則考慮を扱う.

\$3 分の\$2 から 5 で 2 を 1 に足した値を割った値を引いた値
代名詞
\$0 は 7 と 8 をかけた値
\$1 は 5 を 2 を 4 で割った値で割った値
\$2 は\$0 分の 2 と 5 を 4 にかけて値に足した値から 9 を引いた値
\$3 は 1 たす 2 を 3 を 8 にかけて値から引いた値と\$1 をかけた値

同様に実験エは式(f)について文章生成規則を考慮しない場合を扱う.

\$3 分の\$2 引く 5 分の 1 たす 2
代名詞\$0 は 7 かける 8
\$1 は 4 分の 2 分の 5
\$2 は\$0 分の 2 たす 4 かける 5 引く 9
\$3 は 8 かける 3 引く 1 たす 2 かける\$1

2 つの文章はもともと同一だが,文章生成規則の適用,意味分割され,容易な理解を得た.この考察から,演算子が増えるほど文章生成規則の適用が高い理解を得ることを示している.

同様に,意味分割を行う実験ウからカでの全体の平均が 3.5 であるのに対し,利用しない実験ア,イの全体の平均が 2.0 であることから,意味分割は一般に理解度を向上させる働きを有することがわかる.

次の表 9 に,数式ごとに最も評価の高い実験結果の(実験アとの)記述長比を示している.数式(a)を除くとほぼ倍前後で長さを要しており,理解可能な意味分割が最小記述長に対して

2.0 ないし 2.4 程度の効果を生むことがわかる.

数式	実験	記述長比
(a)	ア	1.0
(b)	ウ	1.4
	エ	1.4
	カ	2.0
(c)	エ	1.5
(d)	カ	2.2
(e)	オ	2.4
(f)	カ	2.3

表 9 記述長と理解度

5. 結 論

本研究では,逆構文解析による文章生成の効果を調べるため数式木を意味とする文章生成手法を提案した.木の大きさや演算子量の数,木の複雑さに制約を加え,閾値をこえた場合に数式木を分割分割する.関連付けを与えるため変数記号を導入し,元の関連を保存するものとした.この方法を,生成された文章の大きさ,およびその理解の深さを検証し,対応関係を論じた.

今後は四則式以外への適用を論じる予定である.

文 献

- [1] 中田 育男: “コンパイラの構成と最適化”, 朝倉書店, 1999.
- [2] 森敏昭, 井上毅, 松井孝雄: “グラフィック認知心理学”, サイエンス社, 1995