4S−8

自然な日本語生成のための指針

熊野 明　吉村裕美子　野上 宏康
(株)東芝 総合研究所

1. はじめに

筆者らは，英日機械翻訳システムを開発してきた．機械翻訳において訳文の品質を左右するのは，辞書・文法の質と量である．トランスファ方式を採用している当社のシステムでは，文法知識は解析，変換，生成の3つに分けられる．ここで生成処理とは，概念依存構造を示す木構造データから最終結果である日本語文字列を出力する過程と位置付ける．

今回自然な日本語を生成する目的で，日本語生成文法を改良した．本論文では特に，句読点の挿入基準を中心に報告する．

2. 自然な日本語生成の要因

日本語の訳文における自然さ，理解度(わかりやすさ)の要因はいくつか考えられる．

(1) 語彙選択
(2) 語順
(3) 句読点

(1)は翻訳する文章の分野(技術文，報道文，会話文など)によって，同じ意味を表す場合でも異なった語彙を用いることがある．(2)は原文である英語の情報を反映することによって，訳文としての明確さ，理解度を増すことがある．例えば，目的を表す不定詞句を訳出する際に，英文における位置が異なればその違いを日本語に反映させることによって理解度を増すことができる．

1. The buffer size is used to determine the number of records.

バッファサイズはレコードの数を決定するために使用される．

2. To read input records, the statements must be specified.

入力レコードを読むために，その文を指定しなければならない．

また，(1), (2)とも文脈を考慮してよりよい生成結果を実現する可能性がある．

これに対して，(3)の句読点の問題は日本語独自の課題であり，(1), (2)に比べ1文内だけの情報でほぼ解決されるものである．そこで英日機械翻訳における日本語生成のキー技術のひとつとして句読点挿入に着目することは，訳文品質の向上のために重要なことである．

3. 日本語における句読点の生成

3.1 句読点の重要性

欧州語と違って単語が分かち書きされない日本語において，読点の果たす役割は文の理解度の点から大きいものである．

読点は語句や節の切れ目に挿入することは自明であるが，その数が多すぎても少なすぎても理解度を損う．例えば次の英文に対する3通りの日本語訳を考える．

3. This is most often used to stop processes started by the user.

(a) これはそのユーザによって起動されたプロセスを止めるために最も頻繁に用いられる．

(b) これは，そのユーザによって，起動されたプロセスを，止めるために，最も頻繁に，用いられる．

(c) これはそのユーザによって起動されたプロセスを止めるために，最も頻繁に用いられる．

3つの訳文は語順はまったく同じであるが，(a)では読点がまったくないために，(b)では逆に読点が多すぎて理解度が低くなってしまう．これに対して(c)はかなりよい読点の使用を示している．

英文3の変形例についても，同様に適切な読点を考える．

4. This is most often used.

これは最も頻繁に用いられる．

5. The commands in your directory are most often used to stop processes started by the user.

あなたのホームディレクトリ中のコマンドは，そのユーザによって起動されたプロセスを止めるために，最も頻繁に用いられる．

これらのことから，読点の挿入可能位置は構文的に決っているが，実際に読点を挿入すべきか否かは文全体の長さによって左右されることがわかる．

3.2 句読点生成の問題点

これまでも日本語生成文法で，読点の挿入を制御してきた．その基準は複雑なものであるが，主なものは次のとおりである．

(1) 従属節の直後
(2) 複雑な修飾句・節を伴う格要素の直後
(3) 構成単語数が一定数を超える格要素の直後
(4) 並列要素が3つ以上の並列句

Guidelines on the Generation of Natural Japanese Sentences
Akira KUMANO, Yumiko YOSHIMURA and Hiroyasu NOGAMI
Toshiba Corporation

(1)は構文的に明確な基準であり，経験的にも適切な読点生成を実現している．(2)でいう複雑な修飾句・節とは，埋め込み文や並列句を含むものである．(3)は1つの格要素に相当する木構造の重みを量ること[1]で判断している．(4)は英語のカンマの使用に倣ったものである．

ところがこれらの生成基準は，読点生成に直接関連する句や節の局所的な情報を用いるものである．そのため訳文全体としては，バランスを欠いた読点挿入になる可能性を残していた．

3.3 句読点の使用状況

適切な句読点の挿入を実現するために，実際の英文，訳文における句読点の使われ方を調査した．科学技術分野のマニュアルを中心に英文(原文)と日本文(訳文)を整理している文献[2]から，対訳関係の明確な514対を任意に抽出して，英文中のカンマとピリオドの数，日本文の句読点の数を調査した．その結果を表1に示す．

表1．英日対訳文中の区切り記号の数

	英単語数	カンマ	ピリオド	読点	句点
総数	8,095	284	516	874	524
1文当り	15.7	0.55	1.00	1.70	1.02
1語当り	1.000	0.0351	0.0637	0.1080	0.0647

このデータでわかるように，同じ内容を表す英文と訳文とではカンマと読点の使われ方が数の上で大きく異なる．カンマが1文当り0.55個で少ないのに対し，読点は1.70個でかなり多い．両者の比は 0.55 : 1.70 = 1.00 : 3.08であり，日本語の読点が約3倍多く使われている．本来日本語文の単語数に対する読点の数を算出すべきであるが，複合語の単語認定が多くの可能性をもっているため，代りに英語の単語数に対する値を調べた．その結果は 0.108，すなわち英単語約10語につき読点が1つ含まれることがわかった．

以上のことから，英文の訳文では英文中の単語数の約1割にあたる数の読点を挿入すると，バランスのとれた日本文に近付くと予想できる．この統計では文中の語数を英文でカウントしたが，この数は訳文における日本語の単語数とかなり相関しているはずである．ここでの1文中の単語数と読点の数の比を，読点生成比 α と呼ぶ．上の統計結果から $\alpha = 0.1$ と仮定する．

日本語の文章を書く際に，読点のバランスを考慮しないで書き始め，1文の長さがある程度長くなるとどこか適当な場所を捜して読点を挿入することはよくあることである．上の統計結果は，そのような我々の日本文生成過程を模擬していると考えられる．

3.4 新しい読点挿入基準

現実の句読点の使用状況をシミュレートするために，読点挿入ルールに新たな基準を追加した．その処理は次の通りである．

(1) あらかじめ日本語生成文法で，読点の入り得る位置に「読点可能位置」を設定する

(2) 各読点可能位置には，局所的な性質をもとに優先度(必要度)を割り当てておく

その上で1文ごとの翻訳処理において，次の手順で実行する．

(3) 英文の構成単語数 n をカウントし，訳文中の読点数の期待値 $x = \alpha n$ (整数値) を計算する

(4) 1文全体の生成が終了した時点で，優先度がある閾値より高い読点可能位置には，無条件に読点を挿入する

(5) (4)で挿入した読点の総数が(3)で求めた期待値 x に満たない場合は，残った読点可能位置に優先度の高い順に，x個に達するまで読点を挿入する

こうすることによって，訳文の長さに応じて適切な数の読点を挿入し，より自然な日本語訳文を生成することができる．

3.5 処理例

3.4のアルゴリズムを利用して読点挿入を行った例を次に示す．下線のない読点は(4)で無条件に挿入されたもの，下線付きの読点は(5)の条件で追加挿入されたもの，()は読点可能位置であるが実際に挿入の行われなかった位置を示している．

6. Other signals have the values that the C-Shell inherited from its environment.

[構成単語数 n=12，読点数期待値 x=0.1×12=1]

他の信号には、Cシェルが()その環境から継承した値がある．

7. You need set path once in the .cshrc or .login file.

[構成単語数 n=11，読点数期待値 x=0.1×11=1]

.cshrcあるいは.loginファイルの中で、パスを一度設定する必要がある．

8. These often occur because a shell script or a .cshrc file does something like that command to ensure that the named directories are in the pathname list.

[構成単語数 n=27，読点数期待値 x=0.1×27=3]

シェルスクリプトあるいは.cshrcファイルが、指定されたディレクトリが()パスネームリストの中にあることを保証するために、そのコマンドに類似した何かを行うので，これらは頻繁に生じる．

4. おわりに

日本語訳文の中で読点可能位置に優先度を付与し，1文全体のバランスを考慮することにより適切な数の読点を挿入することができ，結果として自然な日本語文を生成することができた．

読点生成比 α の値は常に一定と考えるより，文の性質によって変動があると予想される．実際に用いる際には，ある程度の幅を考慮して読点の期待値を求めるべきであろう．現在の α は，処理の簡単のために英文中の単語数との比を用いているが，日本語の構成単語数との比，あるいは日本語の表層文字数との比のほうが人間の文生成過程に近いシミュレーションが実現できるかも知れない．

また，読点可能位置の優先度の値も，まだ調整の余地がある．

今後はこれらの値を少し変えて，比較実験を行う．

参考文献

[1] 吉村他:「自然な文章生成のための規範」，情報処理学会自然言語処理研究会資料，NL74-3 (1989).

[2] 富井編:「科学技術和英大辞典」，オーム社(1988).