言語学的類推による生成文における非文法性の分析

イヴ・ルパージュ、白井 諭
{lepage, shirai}@itl.atr.co.jp
ATR 音声翻訳通信研究所

はじめに

人間の言語活動において，類推による新たな言語表現の生成は一般的に行なわれていると考えられるが，言語処理への具体的な応用例は少ない。また，単純な類推により文を生成しただけでは，生成結果の多くが非文法的となることから，あらかじめ一定の制約を設けた上で類推を行なう場合が多い。

これに対して，本論文では，非文法性の分析を狙いとして，記号レベルの類推処理により文生成を行なった結果を報告する。具体的には，既存のテキストコーパスを対象として，この中に現れる任意の文に対し文字レベルで言語学的類推を適用し，新たに生成された文の非文法性の原因を分析した。

1 言語学的類推

1.1 文の間の類推

類推関係は，4つのものの比例関係に基づいた概念である。Hermann Paulは新しい文の生成にも類推関係が適用されると考えた (Paul 20. 109ページ)。さらにBloomfieldは文のパターンはほぼ類推により関係づけられることを示した (Bloomfield 33. 275ページ)。以下に日本語の例を挙げる。

場所が彼に決められた。 : 場所を彼が決めた。 = 柿が彼に食べられた。 : x ⇒

x = 柿を彼が食べた。

1.2 類推を制約する

記号レベルで類推を行なうと非文法的な文や非意味的な文が生成されることがある。例えば：

場所が彼に決められた。 : 場所を彼が決めた。 = 先生が学院に来られた。 : x ⇒

x = *先生を学院が来た。

その理由で，Chomskyは類推が文法性の判断基準にならないとした (Chomsky 86. 12ページ)。

しかし，Itkonenは言語的情報を構文木に与えることにより，生成される文の構文の正しさが類推関係で制御され，非文法的な文の生成を防げることを示した (Itkonen & Haukioja 97)。

1.3 本研究の目的

本論文では，制約を設けることにより類推による非文法的な文の生成を防止するのではなく，逆に，記号レベルで一様に類推による文の生成を行なってから，非文法性の原因を分析を試みる。

2 実験の条件

2.1 データセット

ATR-NECのツリーバンク (Lepage & al. 98) から，「持」という文字が含まれる153文を対象として取り出した。以下，「基礎データ」と呼ぶ。この基礎データの諸元を表1，図1に示す。

表 1: データの諸元

サイズ（文字の数）			文の数
最小	平均 ± 標準偏差	最大	
8	19.5 ± 6.2	37.5	153

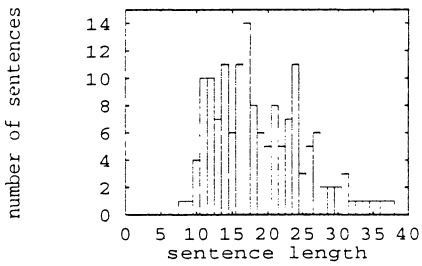

図 1: 文の長さの分布

表 2: 生成文の頻度分布

	新しい文	基礎データと同じ文		合計
		一回出現	二回出現	
文数	1095	151	2	1248
頻度	1-8	すべて 142	すべて 280	約 25.000

表 3: 結果の分類（異なる文）

正しい文		非文		合計
自然	文脈依存	非文法的	非意味的	
453	2	769	24	1248
36.3%	0.2%	61.6%	1.9%	100%

2.2 類推解決

類推方程式を次のようにして解く (Lepage & Iida 98)。はじめに、3 つの記号列の間（持っています：持っていません ＝ お持ちします：x ）から最長共通部分系列を見いだし，余り部分を組み合わせることにより、解答となる部分列を生成する。

x = ε − ε + お
　持 − 持 + 持 ·
　って い − って い + ち し ·
　ま − ま + ま ·
　す + せ ん − す
= お · 持 · ち し · ま · せ ん

記号レベルの類推では一般に 2 つの場合がある。1 つは，3 文の組み合わせからは類推文が生成されない場合である。

お持ちしませんか。 : 手持ちのドルがありません。 = カードは持っていません。 :「正解無し」

もう 1 つは，3 文の組み合わせから類推文が生成される場合である。

お持ちしませんか。 : いいえ、お持ちします。 = 持っていませんか。 : x ⇒

x = いいえ、持っています。

また，類推解決アルゴリズムは非決定的であるが、本実験では複数の類推文が生成される場合、1 番目のみを対象とすることにする。

こっちこい : そっちいって = ここへこい : x ⇒

x = 1/ そいへこって
　 2/ そこへいって

3 実験結果の概要

3.1 生成された類推文の数

基礎データにおいて，単純に 3 文を組み合わせると $153^3/8 \simeq 3.6\times10^6/8 = 450{,}000$ となる。これに対し、類推解決を適用すると，成功する場合が約 25.000 で，成功率は約 1/18 である。

成功した場合の生成文を異なりで見ると 1248 文である。基礎データに含まれる文と同じ文が 153 文生成された。表 2に概要を示す。

基礎データの文数（153 文）と比べると、生成された文の数は約 8 倍になる。

3.2 文法的の種類

表 3に大雑把の分類を示す。3 分の 1 以上の場合には、文法的にも意味的にも正しい文が得られた。

4 正しい文

4.1 含蓄性

文法的，意味的に正しい 452 文のうち，基礎データに含まれる文と同じ文を除くと、299 文が真に新しい文である。は改めて生成された。すなわち，基礎データの文数（153 文）に比べると約 2 倍の新しい文が生成されたことになる。類推による生成能力を特徴づける量として「含蓄性」という概念が提案されている (Lepage & Ando 96)。これは、類推で生成された文法的に正しい文ともとの文の比で表せる。本実験の場合、含蓄性は 195% となる。

以下に新しい文の例を挙げる。

あとでテーブルにそこにあるもの何でも持ってきて下さい。

大きな荷物は船内にはそこにあるもの何でも持ち込めません。
小皿を二，三枚そこにあるもの何でも持ってきて下さい。
確認書を持っていないのですが。
手持ちのドルがあります。

ここは，はっきりと見えるが，類推で文の部分の変換が起こる。

例文から，類推により文の部分置換の様子が観察される。例えば，「そこにあるもの何でも」という部分列が，「あとでテーブルに持ってきて下さい。」や「大きな荷物は船内には持ち込めません。」という基礎データの文に挿入されることにより新しい文が生成されている。。

以上の例では１つの部分列が挿入されているが，一般的には複数の部分列が挿入されることもある。

4.2 利用できない部分置換

ある対立関係が基礎データ中に１組しか見いだせない場合，その対立関係を利用した新しい文の生成を行なうことはできない。

例えば，本実験の基礎データでは，以下の２文の間の対立関係（「は別の」と「違う」）はこの１組み以外には発見できなかった。

このワインとは別のものを持ってきてください。
このワインと違うものを持ってきてください。

5 換入により発生した誤り

対立関係の利用が正しくなかったのは，不適切な文脈において利用されたからであると考えられる。文字列が単語として不適切や，文法的には正しくても意味をなさない場合があるが，記号列に対して類推を行なった結果として生じた問題であり，基本的に同じ原因である。

5.1 語尾表現「す」「せん」の換入

言語における活用の規則性は類推の関係により説明できる。例えば，文献(Lepage 99)ではフランス語の動詞活用への適用が述べられている。本実験では，日本文の語尾表現に現われる「ます」と「ません」を「す」と「せん」の対立として生成した文が多く見られた。

生ものや植物の種などは持っておられませんか。
（基）
生ものや植物の種などは持っておられますか。
キーホルダーを持っています。（基）
キーホルダーを持っていません。

また，この対立関係を活用が異なる「です」に適用したため，次のような非文法的な文が生成されている。

缶ジュースなら持ち込んでも宜しいですか。
（基）
＊缶ジュースなら持ち込んでも宜しいでせんか。

5.2 挿入の位置

前節せは，部分列が連続的に挿入された場合の例を挙げたが，複数の個所で挿入されたものは．置換，挿入．削除が生じる場合がある。

「小皿を二，三枚持ってきて下さい。」という基礎データの文から，「小皿」と「を二，三枚持ってきて下さい。」が，無意味に挿入され，さらに，別の基礎データの文から「日」が文頭に挿入された例を示す。

＊日小皿のマンガを二，三枚おみやげに持ってきていますが問題はありませんか。
＊日小皿の雑誌を二，三枚たくさん持って行きたいのですが，税関で問題はないですか。
＊日小皿の植木を二，三枚アメリカに持って行きたいのですが。

これらの例では「日」を削除すれば文法的には正しくなるが，意味的には不都合が生じている。

5.3 挿入部分

無意味な対立関係を利用したことによる誤りもある。以下の例では，挿入や置換がどの文との関係で生じたかの判断が難しく，原因が解明されていない。

＊グラスは[手荷物]おいくつ持[って]いしましょう，アイスはボックス一つで宜しい[ま]すか。

基礎データのうち最も似ている文は「グラスはおいくつお持ちしましょう，アイスはボックス一つで宜しいですか。」である。「持ち」と「持って」の対立，「ます」と「です」の対立は基礎データからある程度は予想される。しかし，「手荷物」の誤挿入は予想外であった。これは，アルゴリズムの適用の問題である可能性が考えられるので，次節で議論する。

5.4 文字ずれ

さき述べたとおり，アルゴリズムの一番目の正解しか本実験で見いださないので，文字のずれが見える。以下の例では，3つの基礎データ文から．次の文が得られたが，

持っていません。 : 石鹸を持ってきてください。 = お持ちしません。 : x ⇒
x = ＊石鹸をお持てきちくださし。

ほかの類推文としては，「*石鹸をお持ちきてくださし。」がある。文法的に正しくないが，上の例と比べると考えられる。これらの差を編集距離を使って説明できないか検討したい。例えば，入力文と最初の例の編集距離は 11 でありが，後の例は 10 である。

6 自然さの観点からの誤り

非意味的というのは，文法的に正しい文に対する誤りである。

6.1 不自然さ

部分挿入では，文法的には問題のない位置であっても，挿入する部分列が整合するか否かによって，正しい場合と無意味な場合がある。

しかし，判断が難しい場合が少なくない。例えば，次の生成文は特殊な文脈では生成しそうである。

> ? このバッグの大きさなら機内にそこにあるもの何でも持ち込むことが出来ますか。
> ? すぐお持ちしません。

次の文は文法的には可能だか，意味的には相当無理があると考えられる。

> ?? ホテルの看板をそこにあるもの何でも持ってるそうなので見つけて下さい。

この例において，「ホテルの看板を持ってる」場面は考えにくいので不自然である。

このような例は言語学ではよく知られているが，一般に正確に判断することは難しい。

6.2 不整合な文

次の例では，「問題は」の後にポーズを入れ，「禁止されています」を疑問と解釈すれば話し言葉としては言えるそうである。

> 日本のマンガをおみやげに持ってきていますが問題は禁止されています。

また，次の例では前半と後半が論理的に不整合であるが，文法的な問題はない。

> ? すみません，ウォンは持ち合わせていますので日本円で支払ってもいいですか。

まとめ

本論においては，類推解決のアルゴリズムを利用し，153 文の基礎データのうち，任意の 3 文の組み合わせに対し，類推生成を試した。その結果，文法的にも意味的にも正しい文を 299 文新たに生成することができた。

生成された文のうち誤りを含む 807 文を対象として，例を挙げながら誤りの分類を試みた。文字レベルで部分列の対立関係に着目するだけでは，挿入の位置，文脈との不整合が誤りの大部分を占める。今後は，文法や意味による制約のし方について検討を進める予定である。

参考文献

Leonard Bloomfield
Language
Holt, New York, 1933.

Noam Chomsky
Knowledge of Language
Praeger, New York, 1986.

Esa Itkonen & Jussi Haukioja
A rehabilitation of analogy in syntax (and elsewhere)
in András Kertész (ed.) *Metalinguistik im Wandel: die kognitive Wende in Wissenschaftstheorie und Linguistik* Frankfurt a/M. Peter Lang, 1997, pp. 131-177.

Yves Lepage & Ando Shin-Ichi
Un éditeur pour la construction de banques d'arbres
Actes de TALN-96, Marseille, mai 1996. pp. 104-111.

Yves Lepage & 飯田仁
言語に依存しない早期終了型類推解決手法
言語処理学会第 4 回年次大会，九州大学，1998 年 3 月，pp. 266-269.

Yves Lepage, Ando Shin-Ichi, Akamine Susumu, Iida Hitoshi
An annotated corpus in Japanese using Tesnière's structural syntax
ACL-COLING Workshop on Processing of Dependency-Based Grammars, Montréal, August 1998, pp. 109–115.

Yves Lepage
Analogy + Tables = Conjugation
Proceedings of NLDB'99, G.Friedl, H.C. Mayr (eds.), Klagenfurt, June 1999, pp.197–201.

Hermann Paul
Prinzipien der Sprachgeschichte
Niemayer, Tübingen. 1920.