EXPERIMENTING WITH ANALOGY
言語学的な類推とその言語処理への適用性

Yves LEPAGE (ルパージュ)
lepage@itl.atr.co.jp
ATR 音声翻訳通信研究所

はじめに

一般には、言語学的類推は形態素への適用に制限されると考えられる。本論では、構文解析への適用性を示す。具体的には、言語学的類推関係による解析システムの構成と、787文のコーパスを用いた実験結果について述べる。

1 言語学的類推に基づく言語解析

1.1 言語学的類推関係とは

アリストテレスの類推では、A：B = C：Dの関係に基づき、3つ（AとBとC）を与えて残る1つ（D）を予測する。即ち、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 や Bloomfield は新しい文の生成にも言語学的類推関係が適用されると考えた：

$$\begin{matrix} the\ green \\ lamp\ turns\ on \end{matrix} : \begin{matrix} the\ lamp \\ turns\ on \end{matrix} = \begin{matrix} the\ green \\ signal\ is\ off \end{matrix} : \mathbf{x} \Rightarrow \mathbf{x} = \begin{matrix} the\ signal \\ is\ off \end{matrix}$$

チョムスキーは類推に対して否定的見解を示したが、(Itkonen & Haukioja 97) では生成された文の構文の正しさを類推関係でコントロールできることが示されている。

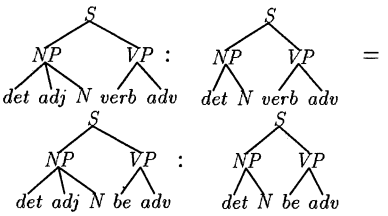
1.2 手法の基本

言語学的類推に基づく言語解析というのは、次の発想を根拠とする。品詞のレベルでの類推関係が満たさせれれば、構文木のレベルでも類推関係が満たされると考えられる (Itkonen 94)。

例えば、前期の例の場合は、品詞のレベルでの類推関係は次のようになる。

$$\begin{matrix} det\ adj\ N \\ V\ adv \end{matrix} : \begin{matrix} det\ N\ V \\ adv \end{matrix} = \begin{matrix} det\ adj\ N \\ be\ adv \end{matrix} : \begin{matrix} det\ N\ be \\ adv \end{matrix}$$

構文木のレベルでも類推関係を見いだすことができる。



1.3 手法の実現

(Lepage & Ando 96) で以下のような実現を提案した。まず第一に、ツリーバンクを使用する。ツリーバンクとは、文とそれに対応する構文木の集合を言う（図1）。

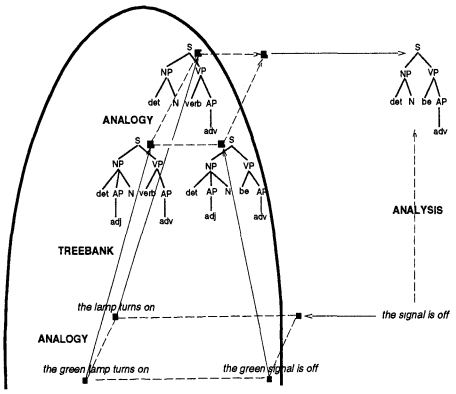


図 1: 言語学的類推に基づく言語解析の原理

第二に、新しい文（例えば、図 1の右にある文）を解析するため、ツリーバンクの3つの文を検索する。新しい文と3つの文の間に類推関係があれば、3つの文に対応する構文木に類推関係を適用することにより、新しい文の構文木を推測する。本当に正解の構文木かどうかの検証が今回の実験の目的である。

以上の考え方を前論文で提案したが、前論文では本論文とは異なる類推関係を利用した。その方法では、ある場合には類推方程式を解くことがで

	サイズ			異なり品詞や
	最小	平均 ± 標準偏差	最大	ノードの数
文（品詞列）	3	12 ± 5.3	57	18
構文木	3	22 ± 9.6	105	38

表 1: データの諸元

きないことが分かったが、それを解消するのは困難であると思われる。

1.4 類推解決

最近、(Lepage & Iida 98) と (Lepage 98) の論文で、記号列の間の類推方程式の解決ためのアルゴリズムを提案した。

例えば、次の類推関係を考える。

$$\begin{matrix} det\ adj\ N \\ V\ adv \end{matrix} : \begin{matrix} det\ N\ V \\ adv \end{matrix} = \begin{matrix} det\ adj\ N \\ be\ adv \end{matrix} : \mathbf{x}$$

類推方程式を次のようにして解く。はじめに、2組の記号列の対 (*det adj N V adv*, *det N V adv*) と (*det adj N V adv*, *det adj N be adv*) から最長共通部分系列を見いだし (Wagner & Fischer 74)、それらを組み合わせることにより、1つ目の記号列と正解の記号列を結合する [*det adj N*+*V*+*adv* / *det adj N*+*be*+*adv*]。

理論的には、このアルゴリズムによれば接頭辞や接尾辞や多数の接中辞の変換ができる。有限トランスデュサーの分野では、多数の接中辞の変換は最近になって得られた (Beesley 98)。我々のアルゴリズムは、記号列の長さの2条の時間以下で動くとともに、正解のない場合、早期終了の特徴も持っている。

2 実験

2.1 データ

提案手法の正しさと実行可能性を検証するための実験を行なった。使用したデータは、ATIS のタスクの 787 文（本実験では、品詞列を文と言う）とその対応する 787 個の構文木である。このデータの諸元（平均の長さや多様性）を表 1に示す。

2.2 評価

実験結果に対し、2種類の評価を行なう。第1に、本手法によって得られた構文木のうち、正しい結果が得られた割合を示す。第2に、本手法により生成構文木と正解の構文木を詳しく比較する。

一般的に、統計言語処理では、表面的な構造比較にとどまり、ノードのラベルは比べないことが多い。これに対して、本研究では、もっと詳しく本手法で得られた構文木と正解構文木の編集距離を計算する (Selkow 77)。従って、構造的な比較の上、ノードの違いも数える。

2.3 実験条件

2つの実験を行なった。両実験では、それぞれ、787 文に言語学的類推に基づく解析を適用した。

第1実験では、クローズド・セットで実験を行なった。その時解析されている文を除いてから、すべてのツリーバンクの中から類推関係文を検索する。従って、この実験の結果は同じ条件で行なわれた (Ando & al. 96) の実験結果と比べることができる。

第2実験では、類推関係文は、その時解析されている文のツリーバンク上の位置より前にある文の中からしか検索しない。そうすることにより、学習的な実験になる。ツリーバンクを読み進めることにより、システムが使用する知識が増えていく。

3 実験結果

3.1 正解

クローズド・セットの実験では、正しく解析された文の割合が高く、約 2 / 3 に達する。学習的な実験では、当然ではあるが減少し、約 1 / 5 になる（表 2と 3）。

(Ando & al. 96) の実験では、類似検索によりツリーバンクが検索されたので、構文木が得られなくてよい場合にも類似度の低い構文木が得られる場合があった。それに対して、本研究の結果では、構文木のレベルで類推方程式を正しく解くことができるため、構文木が得られるべき場合だけが解析できたのである。

2種類の実験において、構文木が生成された場合、1文あたりの構文木の生成数は正解の有無に関係なく、よく一致している。その比は、本手法の非決定性を表す。

3.2 妥当性

出力された構文木と正解構文木の間の距離が結果の妥当性を表す。各距離の価に構文木がいくつ出力されたかを、図 2と 4に示す。解析成功の場合、距離が小さくなると、構文木が増える。さらに、構文木の半数は、対応する正解構文木から3ノードよりも近くに存在する。

	構文木生成		構文木	全体	
	正解あり		生成なし		
文数	506 (64%)	678 (86%)	109 (14%)	787 (100%)	構文木合計
正解構文木数 (1文あたり)	14.3 (7 221/506)	10.7 (7 221/678)	-	9.2 (7 221/787)	7 221
生成構文木数 (1文あたり)	-	725.8 (492 111/678)	-	625.3 (492 111/787)	492 111

表 2: クローズド・セット実験：それぞれの文に出力された構文木の数。

	構文木生成		構文木	全体	
	正解あり		生成なし		
文数	165 (21%)	294 (37%)	493 (63%)	787 (100%)	構文木合計
正解構文木数 (1文あたり)	15.9 (2 621/165)	8.9 (2 621/294)	-	3.3 (2 621/787)	2 621
生成構文木数 (1文あたり)	-	432.0 (127 026/294)	-	161.4 (127 026/787)	127 026

表 3: 学習的な実験：それぞれの文に出力された構文木の数。

3.3 最尤解のみ

本章では、それぞれの文に本手法に出力された構文木の中から最尤解のみに着目する。最尤解と正解構文木の間の距離の分布を、図 3と 5に示す。

前と同じように、近い距離の方が木の数が増える。2種類の実験では、最尤解の95%が正解構文木から2ノードよりも近くに存在する。

まとめ

本論においては、新しい類推解決のアルゴリズムを利用し、言語学的類推に基づく言語解析手法を実行し、ATIS のコーパスの 787 文を使用する実験を行なった。クローズド・セットの実験では約3分の2、学習的な実験では約5分の1の文が正しく解析された。

正解から最も近い構文木だけを注視すると、望ましい結果が得られる。最尤解の95%が正解構文木の2ノード以内に存在する。どのようにして最尤解を選別するかが次の課題になる。文の距離や長さ、及び、木の距離や重みを区別できるかどうかについて、今後さらなる研究が必要である。

参考文献

Kenneth R. Beesley
Consonant Spreading in Arabic Stems
Proceedings of COLING-ACL'98, vol I, Montréal, August 1998, pp. 117–123.

安藤 真一, Yves Lepage & 飯田仁
4項アナロジー関係の構文解析への応用
言語処理学会第3回年次大会発表論文集、1997年3月27－28日、213－216ページ

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.

Esa Itkonen
Iconicity, analogy, and universal grammar
Journal of Pragmatics, 1994, vol. 22, pp. 37-53.

Yves Lepage & Ando Shin-Ichi
Saussurian analogy: a theoretical account and its application
Proceedings of COLING-96, Copenhagen, August 1996, vol .2, pp. 717–722.

Yves Lepage
Ambiguities in analysis by analogy
Proceedings of MIDDIM-96, post-COLING seminar on interactive desambiguation, Christian Boitet ed., August 1996, pp. 93–100.

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
Solving Analogies on Words: an Algorithm
Proceedings of COLING-ACL'98, vol I, Montréal, August 1998, pp. 728–735.

Stanley M. Selkow
The Tree-to-Tree Editing Problem
Information Processing Letters, Vol. 6, No. 6, December 1977, pp. 184-186.

Robert A. Wagner and Michael J. Fischer
The String-to-String Correction Problem
Journal for the Association of Computing Machinery, Vol. 21, No. 1, January 1974, pp. 168-173.

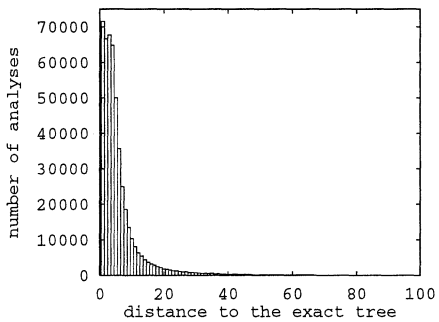


図 2: クローズド・セット実験：妥当性（101 と 481 の距離の間には、642 の構文木がある）。

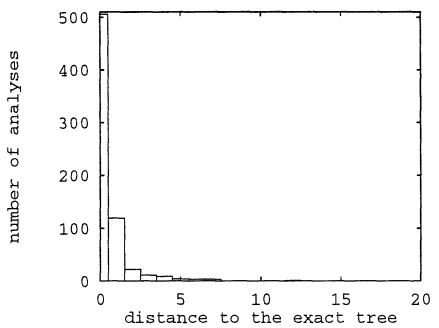


図 3: クローズド・セット実験：妥当性 (正解から最尤解までの距離分布)。

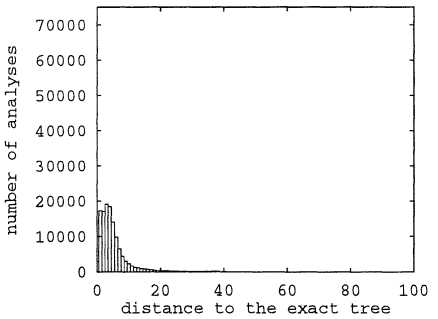


図 4: 学習的な実験：妥当性（101 と 480 の距離の間には、172 の構文木がある）。

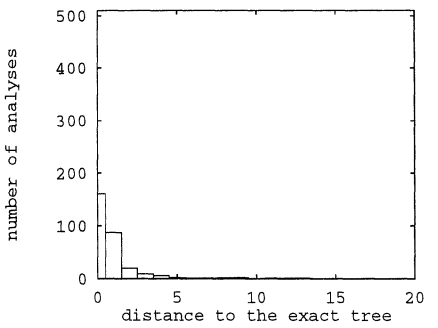


図 5: 学習的な実験：妥当性 (正解から最尤解までの距離分布)。