類推に基づく構文解析の変定数の影響評価

イヴ・ルパージュ & エディ・タイユフェール & 白井　諭
{yves.lepage, eddy.taillefer, satoshi.shirai}@atr.co.jp
エイ・ティ・アール - 音声言語コミュニケーション研究所

1　はじめに

類推に基づく構文解析が、比例類推関係に基づいて形式化され (Lepage 01)、妥当性の観点から評価されている (Lepage 99)。しかし、いくつかの問題が残されている。その一つは、検索空間が大きすぎる危険があり、この検索空間を減少させる方法があれば、本手法が速くなることが期待される。本論文では、検索空間を減少するため、2つの変定数を使用して、結果への影響を測定する。

2　類推に基づく構文解析

2.1　本手法の原則

ここで類推とは、3つのものを与えて残る1つを予測する方法をいう。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。例えば、次の類推関係は

昨日、彼が柿がを食べた	:	昨日、柿が彼に食べられた	=	官吏が打ち合わせ時間を決めた	:	打ち合わせ時間が官吏に決められた

類推方程式では次のように記述される

昨日、彼が柿を食べた	:	昨日、柿が彼に食べられた	=	官吏が打ち合わせ時間を決めた	: x	⇒	x = 打ち合わせ時間が官吏に決められた

2.2　類推に基づく直接的構文解析の原則

類推に基づく構文解析の基本原則は次の通りである。下位のレベルでの類推関係が満たされれば、（例えば、品詞のレベル）、

副詞、名詞が名詞を動詞	:	副詞、名詞が名詞に動詞	=	名詞が名詞を動詞	:	名詞が名詞に動詞

上位のレベルでも類推関係が満たされる（例えば、構文木のレベル）と考えられる（図 1、(Itkonen 94)）。

2.3　類推に基づく直接的構文解析の実現

本手法を実際に動作させるには、ツリーバンクが必要である。新しい文の構造を計算する時、次のようにする。先ず、その新しい文に対応する品詞列（D）がツリーバンクで見つかれば、ツリーバンクにある対応する構文木を単純に出力する。見つからなければ、それぞれの2つの品詞列の組み合わせ（A,B）を取って、$B : A = D : x$ の方程式を解決する。そういう組み合わせは言語学の観点から、変換を表すので、モデルと呼ぶ。また、この方程式には、$\mathbf{x} = C$ という正解があれば、C がツリーバンクにあるかどうかを確かめた上、A、B と C に対応する構文木で（$\widehat{A}$、$\widehat{B}$ と $\widehat{C}$）成立された方程式を解く。得られた解 $\widehat{D}$ を、新しい文 D に対応する構文木とする。

$$\begin{array}{ccccccc} \widehat{A} & : & \widehat{B} & = & \widehat{C} & : \mathbf{x} \Rightarrow & \mathbf{x} = \widehat{D} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ A & : & B & = & C & : \mathbf{x} \Rightarrow & \mathbf{x} = D \end{array}$$

従って、本手法で、N の大きさでのツリーバンクがあれば、モデルの数は $N \times (N-1)$ になる。我々の実験では、N が 5,000 であるから、モデル集合の大きさは 25×10^6 個になる。しかし、数十万規模の N にも適用できるようにするには、モデル空間を減少しなければならない。その目的で、本論文では、2つの可能性を検査した。

3　実験

3.1　実験手順

本実験では、2つの手法が判断される。

- 長さで制限されたモデルでの手法;
- 類似で制約されたモデルでの手法。

長さで制限されたモデル　類推関係の公理に基づいて、次の定理が証明できる。

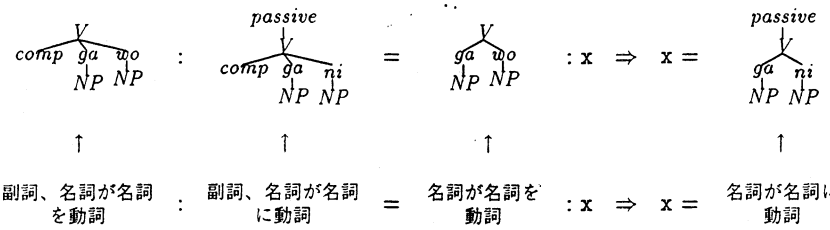


図 1: 類推に基づく構文解析の原則。

$$A : B = C : D \Rightarrow |A| + |D| = |B| + |C|$$

これに従って、類推に基づく構文解析を行なうとき、$|A| > |B|$ の制限があれば、以上に記述された手法で、探される C は D より短くて、検索の時間が減ると予想される。

類似で制約されたモデル 直感的には、モデルの中にある文が似ているなら、類推で生成されたものも似ていると考えられる。それに対して、ある程度以上類似するモデルだけを使おうという考えがある。制約として、類似度（σ）で定義する：$k \leq 2 \times \sigma(A,B)/|A| + |B|, 0 < k < 1$。

両方の場合では、計算時間が減ると考えられるが、モデルの数が減ると、結果へも影響があると予想される。従って、制約の変定数に対して、結果の妥当性を観察しなければならない。

3.2 評価

本実験では、5,000 文のツリーバンクを使用して、異なる 1,553 文の構文解析を行なった。この 1,553 文に対応している構文木が予め生成されているので、実験で得られた構文木とその構文木を比較した。また、以前の 2 つの手法の差異を明らかにするため、制限や制約を設けない根本手法の結果も測定した。

3.2.1 測定

結果の品質を計るため、一文あたりいくつの構文木が得られたかを数え、そのうち、いくつが正解であるかによって評価する。

さらに、より詳しい評価も行なう。それぞれの得られた構文木と正解の編集距離も計算する (Selkow 77)。その距離の大きさ、構文解析の品質も特徴付けられる。ある得られた構文木の品質は、正解距離に反比例する。距離 0 は、構文木が正解であることを示す。

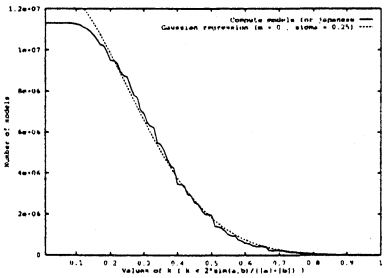
得られた構文木の品質分布をグラフで表す。そのグラフでは、横座標は正規化距離を表す：正解からの距離と正解の長さの比。縦座標である正規化距離の価に対して得られた構文木の数を示す。

3.3 結果

3.3.1 モデル集合の大きさ

長さ制限の影響 制限の使用では、モデルの集合のサイズは半分になる。それだけで、大幅に高速化される。以下に、このモデル集合のサイズの大幅な削減が結果の品質に余り影響を及ぼさないことを示す。さらに、C を検索するとき、検索の時間も減少できる：$|A| + |D| - |B|$ 以上の長さを持っている文からの検索が充分であるからである。

類似制約の影響 先ず、k の変定数で定義された類似制約の影響を調査する。以下のグラフでは、$2 \times \sigma(A,B)/|A| + |B| \geq k$ という類似制約での集合のサイズの変換を示す。ここは、百の k の価を計った（$0 < k < 1$）。また、ガウス回帰グラフは点線で表した。



最大のモデル集合のサイズは 1,100 万個であって、k は 0.75 の価で、14 万個になる。グラフは大体にガウスのグラフ形である。k の制約の適用では、サイズでの分布がランダムであって、データの分布が変わらないことを意味する。

3.4 結果の品質

長さ制限の影響 制約の影響を調査するため、色々な k の値に、2 つのグラフを計算した。図 2では制約なし、図 3では制約を適用した。この両者に差異はほとんど認められない。詳しいグラフの解析では、次のようなことがわかった。

- 両方の場合では、正解の数は正しくない結果の数を越えると見られる。
- 制約の適用が有利である。
  - 結果の全体の形を変えないで、
  - 正しくない結果の数が少しでも減少して、
  - 妥当性が（正解の数と全ての結果の数の比）微妙に上がる。

類似制約の影響　以下に類似制約（k の変定数）の影響を観察した。

色々な k の値に対する構文木の数を図 4に示す。予感されるように、k が上がると（モデルの中にある２つの文の類似度が上がると）、正しくない結果の数が減る。しかし、正解の数も微妙に減ると見られる。次のように説明できる。

- 一般的に、変定数が k が 0 に近くなると、モデルの中にある２つの文の間の変換が大きくなると考えられる。逆にいうと、k が 1 に近づくと、モデルの中にある２つの文の間の些細な言語的変換しか許されていないと考えられる。
- 従って、変定数 k が 0 に近くなると、変換が言語学的に無意味である可能性が大きくなって、正しくない結果が当然に増えると考えられる。

それに対して、手法の妥当性を検査した。次の量を計った。

1. 正しく解析された文の数の割合：それは、普通の検索実験では再現率に相当すると考えられる。
2. 一文あたりに正解の数と得られた構文木の数の比：それは、普通の検索実験では適合率に相当すると考えられる。
3. 全ての得られた正解の数と全て得られた構文木の数の比：それは、手法の妥当性を表す量と考えられる。

図 5では、最左のグラフは構文解析された文の割合も正解が得られた文の割合も正解が得られた文の割合も現われている。変定数 k が上がると、その２つの割合は収斂する。まとめると、次のように言うことができる。

- 変定数 k が上がると
  - 得られた構文木が正しくなって、
  - 全体の妥当性と一文あたりの適合率が上がる。
  - しかし、再現率が下がる。

4　おわりに

類推に基づく構文解析で使われるモデルの数を減少させるため、２つの変定数（長さの制限と類似の制約）の影響を計った。結果として、次の主な２つの知見が得られた。

- 長さの制限の適用で正解の数が変わらないが、正しくない構文木の数が微妙に減少する。従って、長さの制限の適用は有利である。その制限では検索空間が半分になって、結果が余り変わらずに、計算時間が削減できる。
- また、類似制約の適用で妥当性が少し向上するが、、解析できる文の割合が少し減少する。その類似制約を適用すると、ガウス形の検索空間の減少が観察される。

まとめると、提案された類推に基づく構文解析手法を使用するとき、長さ制限は有効であり、妥当性に応じて類似制約の調節が可能である。

謝辞

本研究は通信・放送機構の研究委託により実施したものである。

参考文献

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

Yves LEPAGE
Formalisation de l'analogie entre chaînes de symboles
Actes de CAp-2001, plateforme AFIA-2001, Grenoble, juin 2001, pp. 117–131.

Yves LEPAGE
Open Set Experiments with Direct Analysis by Analogy
Proceedings of NLPRS-99, Beijing, November 1999, p. 363–368.

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

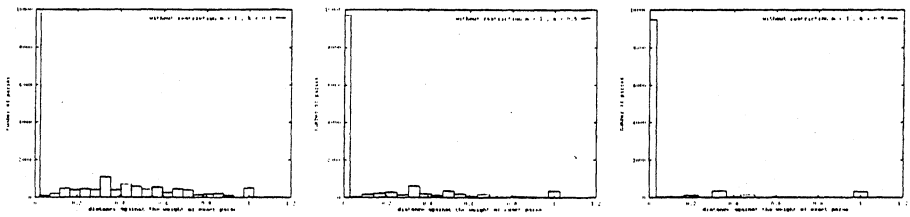


図 2: 長さ制限なし、類似制約なしの場合（ベースライン）。横座標：正解からの距離 / 正解の長さ、縦座標：得られた構文木の数。$k = 0.1, 0.5, 0.9$

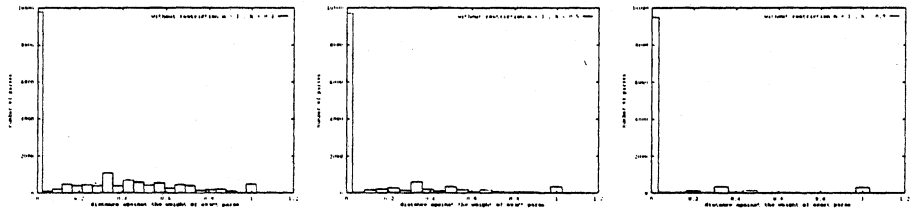


図 3: 長さ制限ありの場合。横座標：正解からの距離 / 正解の長さ、縦座標：得られた構文木の数。$k = 0.1, 0.5, 0.9$

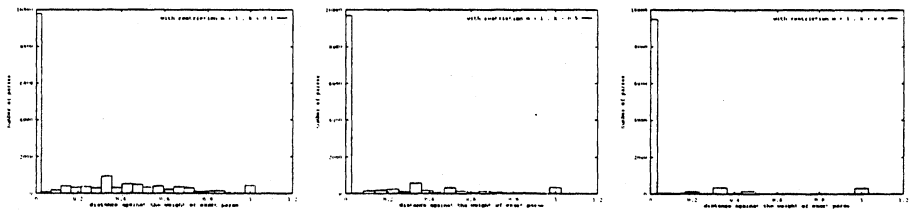


図 4: 類似制約ありの場合。横座標：正解からの距離 / 正解の長さ、縦座標：得られた構文木の数。$k = 0.1, 0.5, 0.9$

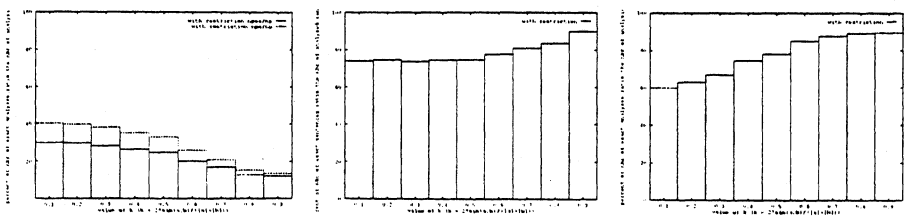


図 5: 本手法の妥当性。横座標：$k = 0.1, 0.3, 0.5, 0.7, 0.9$、縦座標：構文木の割合