類似検索機能を備えたツリーバンク構築エディタ*

4B－2

安藤 真一　Yves Lepage †
ATR音声翻訳通信研究所‡
e-mail: {ando,lepage}@itl.atr.co.jp

1　はじめに

ツリーバンクは、コーパスに構文解析木などの言語構造を付加した言語データ群であり、自然言語処理システムの評価や言語知識の抽出、統計情報の収集などに有用である[1]。しかし、ツリーバンクの構築は人手によって行われるため、多大なコストを要する。特に、言語的情報を表現する木構造の入力や編集、そして類似した文や句に異なった言語構造が付加されることを避ける無矛盾性のチェックに相当な時間と労力が費されている。そこで我々は木構造を直接入力、編集できるツリーエディタを作成し、類似した文や木構造の検索を可能とする類似検索機能を追加することにより、ツリーバンク構築エディタを試作した。さらにこのエディタは類似性に基づく解析機構を有しており、新たに入力された文に対する言語構造候補を提示することができる。

本稿では、試作したエディタのツリーエディット機能、類似検索機能、類似性に基づく解析機構について報告する。

2　ツリーエディット機能

一般に構文解析木などの言語構造は木構造によって記述できる。しかし、通常用いられるテキストエディタでは、木構造は特殊な表現形式を介して編集する必要がある。このためエディタ上でのデータの可視性が悪く、また構造を表す記号など余分な編集操作が必要となる。一方、木構造を木の形式で表示するツールもあるが、編集にはノード毎に編集用のダイヤログボックスを開く必要があるなど、間接的な編集操作しかできかった。これに対し、本エディタは木構造を木の形式のまま編集するツリーエディタを備えている。

図1に本エディタの画面例を示す。

この画面において下部はテキストエディタ、上部はツリーエディタである。ツリーエディタではツリー形式で表示された構造を直接入力、編集することができる。特にここでは、通常の単語/行に対する編集操作によってノード/サブツリーを編集することができる。例えば、操作列「NP <return> det <space> AP <return> adj <up> <space> N」で図1の木構造を入力することができる。このように木構造を木の形のまま編集できるため、編集効率の向上が期待できる。

また、画面中の上下それぞれに入力されたツリー構造の各ノードと文の各部分の間に対応関係を持たせることができる。これにより、サブツリーとそれに対応する文の部分を、同時に削除したり移動することができる。

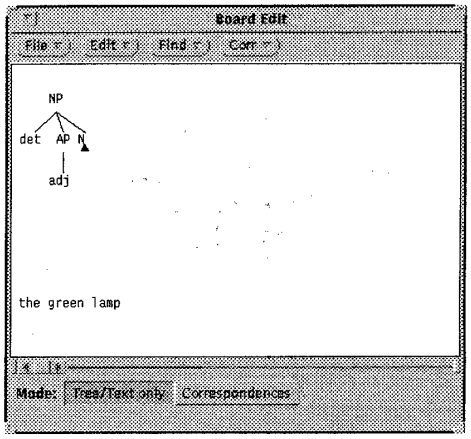

図1: エディタの画面例

3　類似検索機能

類似した文がある場合、各々の文が持つ言語構造は似ていると考えられる。このとき、もし入力文に類似した文がツリーバンク内にあれば、その文に付加された言語構造を再利用し修正することで、入力文に対する言語構造を低コストで得ることができる。またツリーバンク内のデータを再利用することで、データの信頼性(無矛盾性)も向上すると考えられる。そこで本エディタは類似した文を検索する機能を装備する。これにより、検索キー(文や木構造)との類似度が、閾値として与えられた値以下になるようなデータをツリーバンクから得ることができる。

今回試作した類似検索では類似度として編集距離を用いた。編集距離とは、2つのデータを同じものとするために必要な最小編集コストである[2, 3]。特にここでは、編集操作として削除、挿入、置換の3つを考え、その編集操作数によって類似度を定義した。例えば「The green lamp turns off.」と「The lamp turns on.」を考えると、第1文の「green」を削除し、「off」を「on」へ置換することで両者が同じ文になるため、この2文間の類似度は2となる。

同様の類似検索を実行するためのアルゴリズムは既にいくつか提案されている[4]が、我々はACアルゴリズムの拡張によって類似検索機能を実現した[5]。この手法では、まず与えられた閾値以下の類似度で検索キーとマッチするパターンを生成し、得られたパターンをACアルゴリズムによって検索する。特に検索パターンの生成では、他のパターンを部分的に含むパターンの生成を抑制することで、検索精度を落とすことなく高速化を実

*An editor exploiting similarities to build consistent treebank

†Shinichi ANDO　Yves Lepage

‡ATR Interpreting Telecommunications Research Labs.

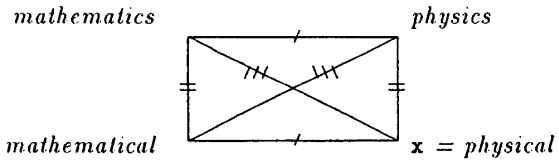

図 2: 類似性の関係を示すモデル

現している。

4 類似性に基づく解析機構

近年盛んに用例や類似性を用いた翻訳手法が研究されている [6] が、本エディタでも入力された文に対する言語構造を、類似性を用いてツリーバンク内のデータから計算する機構を備えている。以下では、この類似性に基づいた解析手法について記す。

Saussure は種々の言語について、ある語の語形変化を他の語の語形変化との類似関係によって説明している [7]。例えば、以下のように 3 つの単語の類似性から「physical」が導出できるというものである。

English: $mathematics : mathematical = physics : \mathbf{x}$

$\mathbf{x} = physical$

上記の等式において、各語間の類似性の間には特定の関係が見出せる。例えば、「mathematics」と「physics」は字面の一部が似ており、意味的にも近い。これに対し、「mathematical」と「physical」の間にも全く同じ類似性を見出すことができる。このような類似性の関係は、他の単語対同士に関しても見ることができる。そこで、図 2に示すような関係モデルを導入する。ここで図中の各語を結ぶ線は各語の間の類似性の大きさを表し、同じ記号の線はその大きさが同じであることを表す。このモデルでは、上記の 3 つの単語の類似性の関係と同質の関係を持った単語として「physical」が導出できる。上記の例では単語を例に取ったが、文においても同様の議論が成り立つと考えられる。

本手法では上記モデルに基づいて、入力文に対する言語構造を計算する。ここでは類似度として上述の編集距離を用いた。具体的には、まず入力文 (図 3中の a)) から最も近い 2 つ文 (図 3中の b), c)) をツリーバンクから検索する。そして入力文と得られた 2 文に上述のモデルを適用し、4 つ目の文 (図 3の d)) を得る。さらに、ツリーバンクから得られた 3 つの文に付加されている言語構造 (図 4中の b), c), d)) に上述のモデルを適用して、入力文に対する言語構造の候補 (図 4中の a)) を生成する。この解析では複数の候補が得られるが、もし入力文に対する言語構造がツリーバンクに存在すれば、距離空間の性質から正解を含むと考えられる。

5 おわりに

本稿では、質の高いツリーバンクを効率良く作成するために、ツリー構造の直接編集機能と類似検索機能を備えたツールについて報告した。また類似性の間の関係モデルを導入し、これを利用した解析手法を提案した。今後、理論的な検証や実データを用いた実験によって、この解析手法の正当性を検証、評価する予定である。

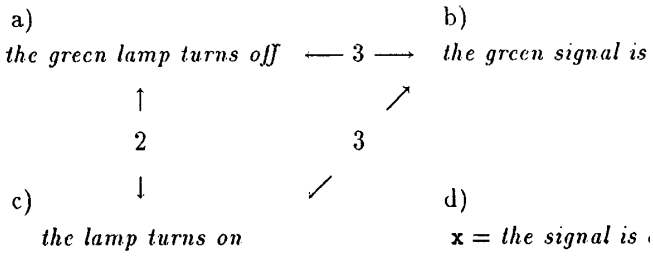

図 3: 文の解析

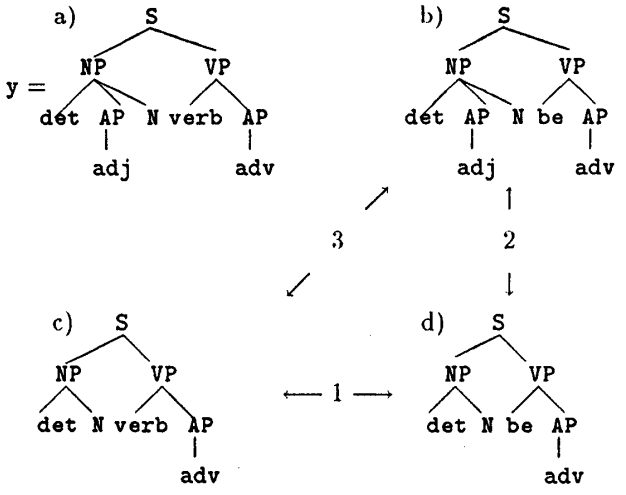

図 4: 言語構造の生成

参考文献

[1] 宇津呂他, コーパスを用いた言語知識の獲得, 人工知能学会誌, Vol.10, No.2, pp.33-40, 1995.

[2] R.A. Wagner 他, The String-to-String Correction Problem, *Journal for the ACM*, Vol.21, No.1, pp. 168-173, 1974.

[3] S.M. Selkow, The Tree-to-Tree Editing Problem, *Information Processing Letters*, Vol.6, No.6, pp.184-186, 1977.

[4] G.M. Landau 他, Fast String Matching with k Differences, *Journal of Computer and System Sciences*, Vol.37, pp.63-78, 1988.

[5] Y.Lepage 他, A first quantitative analysis of approximate pattern-matching, 人工知能学会研究会資料 *(SIG-J-9501)*, pp.17-23, 1995.

[6] Nagao Makoto, A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in *Artificial Intelligence and Human Intelligence*, Elithorn A. and Banerji R. eds., Elsevier Science Publishers, 1984.

[7] Ferdinand de Saussure, *Cours de linguistique générale*, publié par Charles Bally et Albert Sechehaye, Payot, Lausanne et Paris, 1916.