4項アナロジー関係の構文解析への応用

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

1 はじめに

構文解析システムにおける問題の一つとして、出力される複数の構文木候補から正しい木を選び出すことの困難さがある [1]。これは主にシステムの知識不足に起因するものであるが、十分な知識を人手で収集、整備することもまた難しい。そこでテキスト等の実データから知識を抽出、利用するために、統計情報を用いる方法 [2] や意味的類似性を用いる方法 [3] が提案されている。

我々もツリーバンク (構文木付きコーパス) を用いて、入力文と類似した文の構文木から入力文に対する構文木を類推する手法について研究を行っている [4]。本稿では特に、類推結果の妥当性を表す指標を導入し、構文解析の曖昧性解消のための一指標として利用する手法を提案する。ここではまず、互いに類推可能な4文の間に成り立つ関係を定義し、これを構文解析に利用する手法について述べる。次に類推の妥当性を表す指標を新たに導入し、Penn Treebank を用いた構文解析実験からその有効性を示す。さらに既存の構文解析システムとの融合について検討、実験した結果について報告する。

2 4項アナロジー関係

Saussure は、ある語の語形変化パターンを他の語に適用することで新たな語が創造される現象をアナロジーと呼び、言語の創造性という観点からその重要性を指摘した [5]。例えば、以下の3つの単語のアナロジーからは「physical」が導出できる。

$$mathematics : mathematical = physics : \mathbf{x} \quad (1)$$
$$\mathbf{x} = physical$$

本稿ではこの種のアナロジーを取り扱い、特にこのアナロジーが成立する4項の間の関係を4項アナロジー関係と呼ぶことにする。

2.1 4項アナロジー関係の定式化

我々は、アナロジー関係にある各項が互いに交換可能であることに着目し、4項アナロジー関係の定式化を行った [4]。ここで交換可能であるとは、例えば上式 (1)

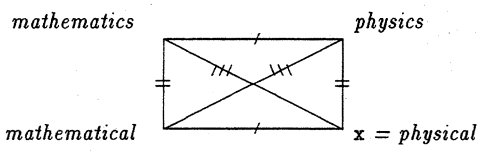

図 1: 4項アナロジー関係のモデル

のアナロジーが成り立つ場合には下記の2式によっても同じ $\mathbf{x}$ が導出できることを表す。

$$mathematics : physics = mathematical : \mathbf{x} \quad (2)$$
$$mathematical : physics = mathematics : \mathbf{x} \quad (3)$$

式 (1) では、意味を保存しながら語彙的な機能を変化させる左辺の語形変化パターンが右辺に写像されており、式 (2) では逆に、語彙的な機能を保持しながら意味を変化させる語形変化パターンが写像されていると考えられる。また式 (3) では各々の変化が同時に起こっていると見ることができる。左辺右辺の変化パターンはそれぞれ等しいため、アナロジーの見られる4単語の関係は図1のモデルで表すことができる。ここで図中の各語を結ぶ線は各語の間の距離を表し、同じ記号の線はその大きさが同じであることを表す。すなわち4項アナロジー関係は下記のように定式化できる。

定義1 (4項アナロジー関係)

$$u : v = w : \mathbf{x} \overset{\triangle}{\Longleftrightarrow} \begin{cases} dist(u,v) = dist(w,\mathbf{x}) \\ dist(u,w) = dist(v,\mathbf{x}) \\ dist(v,w) = dist(u,\mathbf{x}) \end{cases}$$

ここで、$dist(a,b)$ は a と b の間の距離を表す。

2.2 編集距離

定義1の距離として、今回は編集距離を採用した。編集距離とは2つの文字列を同じくするために必要な最小編集操作コストである [6]。特にここでは、削除、挿入、置換の3つの編集操作を考え、その操作数によって距離を定義した。例えば「mathematical」と「mathematics」

を考えると、「mathematical」の末尾部分で「a」を「s」に置換し、「l」を削除することで2つは同じ文字列となる。このため、この2単語の間の編集距離は2と計算できる。同様に4単語間の距離をそれぞれ計算すると、式(1)の単語間で定義1が成立することが分かる。

2.3 文、構文木への拡張

単語単位の編集操作を考えると、単語列、すなわち文における編集距離も文字列の場合と同様に定義できる。例えば「私 / は / 先生 / です」と「彼 / は / 生徒 / です」という2つの文を考えると、「私」という単語を「彼」に、「先生」を「生徒」に置換することで両者は等しくなり、この2文間の距離は2と計算できる。さらにノード単位の編集操作を考えると、木構造間でも同様の距離が定義できる[7]。

これらの編集距離を用いると、文や構文木における4項アナロジー関係が定義1によって定式化できる。文において4項アナロジー関係が成立する例を以下に示す。

"私は先生です":"私は生徒です"＝"彼は先生です":x
x ＝ "彼は生徒です"

3 4項アナロジー関係を用いた構文解析

3.1 基本原理

文中の単語と構文木中のノードの間にはおのおの対応関係があり、片方の変化は他方に影響を及ぼす。そこで我々は「文においてアナロジー関係が成り立つならば、各々の文に対応する構文木の間でもアナロジー関係が成り立つ」と仮定した。この仮定に基づくと、ツリーバンクを用いて入力文に対する構文木が計算できる[4]。

具体的には、まず入力文(図2のa)と4項アナロジー関係にある3文(図2のb、c、d)をツリーバンクから検索する。ここで得られた文はツリーバンク内でそれぞれ構文木を持っている。そこで、その3つの構文木(図3のb'、c'、d')に4項アナロジー関係を適用して、入力文に対する構文木(図3のa')を計算する。

3.2 基本原理の評価

本構文解析原理は上述の仮定の上に成り立っている。そこで実験を通じてこの仮定の検証を行った。

まず実験システムとしては、入力文に対する構文木を新たに生成するのではなく、上記の基本原理に従ってツリーバンクから検索する機構を実現し、これを利用した。またツリーバンクとしては、比較的似た文が集められている Penn Treebank (rel.2.0) 内の ATIS (Air Travel Information System) データを利用した。ただし Penn

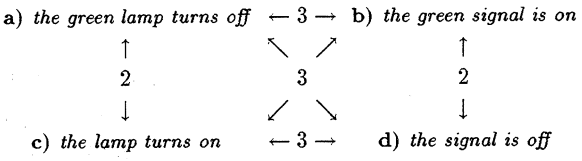

図 2: 文におけるアナロジー関係

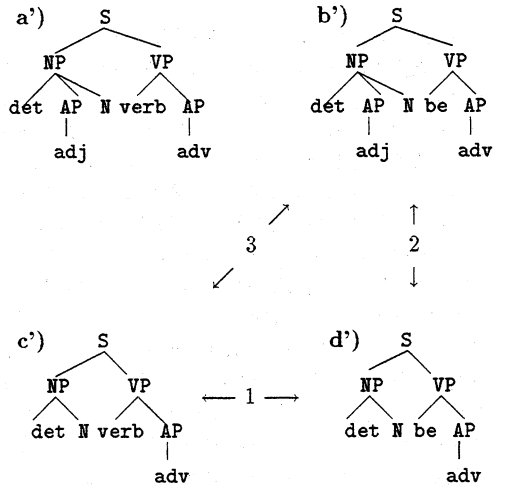

図 3: 構文木におけるアナロジー関係

Treebank に含まれる構文木は最下位ノードのほとんどが文中の単語であったため、本実験では最下位ノードを削除した木を構文木として利用した。この操作によって、ATIS の全データ 577 文中の 376 文が他の文と同じ構文木を持つようになった。

ここではATISデータから取り出した1つの文を入力とし、残る576文分のデータをツリーバンクとして実験を行った。検索結果はATIS内で入力文に付加されていた構文木と比較することで評価した。577文全てが入力となるように同様の処理を繰り返した結果、正しい解析が可能な376文の中の375文で正しい構文木を得ることができた(再現率 99.7%)。また失敗した1文は比較的長い文であり(16単語)、4項アナロジー関係にある適切な文がツリーバンクに存在しなかったことが原因であった。この結果から上記の仮定はほぼ正しく働いていると考えられる。

3.3 類推妥当性

上記の実験結果から、本手法では正しい構文木だけでなく、大量の間違った構文木も出力することが分かっ

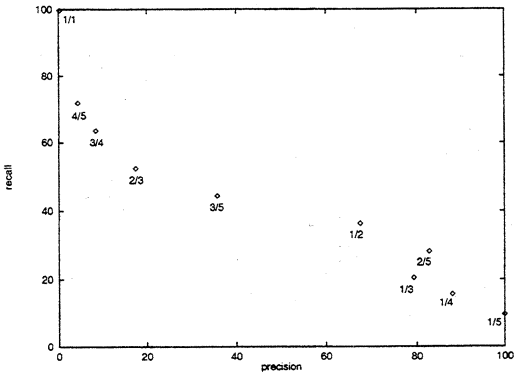

図 4: 制約付き検索実験の結果：再現率と適合率

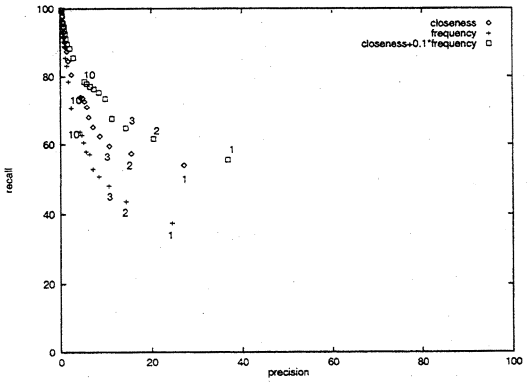

図 5: 類推妥当性による解析結果の順位付け

た (適合率 0.2%)。これはアナロジー関係が変化パターンの保存関係を意味するにもかかわらず、編集距離の保存関係として記述した弊害であると考えられる。すなわち編集操作列の関係を編集操作数の関係で定義したために、類推が成り立ちにくい文や構文木の組合せが 4 項アナロジー関係に含まれてしまい、これが誤った出力に結び付いたと考えられる。そこで 4 項アナロジー関係にある文や構文木が互いに類推しやすいものであるかどうかを調べることで、解析結果のもっともらしさを評価する手法を提案する。以下では、この類推が成り立ちやすいかどうかの尺度を類推妥当性と呼ぶ。類推妥当性の要素としては、以下の 2 つが考えられる。

4 項間の類似性

直観的にアナロジー関係にある 4 項が互いに似ていれば、お互いの間の類推が成立しやすいと考えられる。これは 4 項の間の距離が小さいほど、類推が成り立ちやすいことを意味する。

これは数学的には次のように説明できる。編集距離によって定義される多次元空間を考えると、1 つの文 (構文木) はその空間上の 1 点として表現される。また定義 1 を満たす点は、点 u、v、w からの距離がおのおの一定であるような超球 (多次元の球) の交わりとして表現される。ここで各超球の大きさが小さくなると、検索結果の空間 (交わり部分) が狭くなり、不正な検索結果を除くことができる。

これを確認するために、4 項アナロジー関係の各項間距離に上限を設けて先と同様の検索実験を行った。図 4にその結果を示す。図中の各点に付加した数字は距離の上限を表す。例えば 1/4 の点は、各項間の距離が各項の大きさの 1/4 を越えないアナロジー関係だけを使って検索した結果である。ここで上限が 1/3 の結果と 2/5 の結果で適合率の逆転が見られるが、これは ATIS データに数多く含まれていた「Flights from A to B」(A、B は 1 単語の地名) というパターンの文が 2/5 で検索可能になったためであり、ツリーバンク内データの偏りに起因する。しかし、全体的には 4 項間の距離と適合率の間に比例関係を見出すことができる。

類推パス数

本手法では入力文から構文木を検索するためにツリーバンク内の 3 つのデータを用いるが、複数の 3 つ組を介して同じ構文木が出力される場合がある。このように 1 つの構文木を導出する 3 つ組の数を以下では類推パス数と呼ぶ。類推パス数が多い場合、すなわち多くの組合せから同じ構文木が類推できる場合には、その類推結果の信頼性も高いと考えられる。

先と同様に数学的な観点から見ると、類推パス数が大きくなるほど制約となる超球の数が増える。このため、それら全てが交わる部分空間は狭くなり、不正な結果を生み出す検索結果の空間を小さくすることができる。

実際に先の実験結果を調べると、1 つの入出力関係に対する類推パス数は、全入出力関係の平均では約 225 通りしかないのに対し、正しい入出力関係だけを見ると平均で約 1938 通りであった。この結果から、類推パス数が類推妥当性として機能し得ることが分かる。

3.4 類推妥当性による解析結果の順位付け

上記 2 つの類推妥当性が解析結果のもっともらしさを測る評価値となり得ることを確認するために以下の実験を行った。実験では先の検索実験によって得られた構文木を類推妥当性に従って順位付けし、上位 N 位以上の検索結果から再現率と適合率を調べた。図 5に結果を示

す。ここでは類似性 (図5の closeness) と類推パス数 (図5の frequency) の他に、両者の重み付き和を用いた場合についても評価した。各点に付加した数字は上位何位までの結果を用いたかを表す。この図を見ると、各々の評価値を用いた結果が共に右上がりの曲線を描いており、高い順位に正しい構文木が偏在していることが分かる。

4 文法知識を用いた構文解析機構との融合

本手法を用いると、文と構文木の間の対応関係が正当であるかどうかをツリーバンクからの類推という観点で評価できる。そこで以下では、従来の構文解析器が出力した構文木の正当性を類推によって評価し、曖昧性解消に利用する機構について考える。ここでは特に、当研究所で開発している用例主導翻訳システム TDMT の構文解析部を対象として類推機構との融合を考える。

4.1 TDMT と類推機構の融合

TDMT の構文解析部は、文法的な機能単位を表すパターンの組合せとして構文木を解析する。このとき、パターンの組合せ方によって曖昧性が生じるが、用例との意味的類似性を用いて各構文木の正当性を評価する [3]。これに対して本手法は、文や構文木全体の構造的な類似性を利用して正当性の評価を行う。本手法でも曖昧性は生じるが、文の部分的なパターンの組合せで構文木を捉える TDMT の構文解析とは曖昧性の質が異なると考えられる。このため、この両者を融合することで構文解析の曖昧性を小さくできると考えられる。

4.2 融合による効果の予備的評価

上記を確認する予備実験として、TDMT の構文解析結果に対する類推妥当性を約 20 文について調べた。ここで類推に用いるツリーバンクには TDMT で正しく解析できた 665 文を利用した。2 つの入力文に対する結果を図 6 に示す。各表において、解析結果には TDMT の構文解析部が出力した構文木を示しており、その数だけ曖昧性が生じたことを表している。また TDMT と類推の欄の数字は各々のシステムが算出した各構文木に対する評価値であり、共に値が小さいほどその解析結果が確からしいことを表す。ここで、入力文 1 では 1 番上の解析結果が、入力文 2 では上から 2 番目 (あるいは 3 番目) が正しい解析結果である。この表によると、類推機構を用いた手法ではそれぞれ正しい解析結果に最もよい値を与えていることが分かる。他の文についても正しい構文木には TDMT か類推のいずれかでよい評価値が与えられており、類推妥当性を曖昧性解消の一指標とすることで TDMT の構文解析精度の向上が期待できる結果が得られた。

入力文 1：「一泊は確か二百ドル程度でしたよね」

TDMT	類推	解析結果
2.6389	0.0909	SM(SM(NP(TERM,S+N(TERM,ND(TERM)))))
2.8611	0.1158	SM(SM(S+N(NP(TERM,TERM),ND(TERM))))
3.0000	0.1158	SM(SM(NP(TERM,NP(TERM,ND(TERM)))))
3.1667	0.1000	SM(SM(NP(TERM,N+N(TERM,ND(TERM)))))
3.1667	0.1163	SM(SM(NP(TERM,ND(N+N(TERM,TERM)))))

入力文 2：「そちらから歩いて十分ほどになります」

TDMT	類推	解析結果
0.2000	0.1576	SM(NP(NP(TERM,TERM),NM(ND(TERM))))
0.3095	0.0909	SM(NP(TERM,NP(TERM,NM(ND(TERM)))))
0.3095	0.1311	SM(NP(TERM,NM(DN(ND(TERM)))))
0.3667	0.1622	SM(NM(N+N(TERM,DN(ND(TERM)))))
0.3667	0.1829	SM(NM(ND(N+N(TERM,DN(TERM)))))
0.7000	0.1599	SS(NP(TERM,TERM),SM(NM(ND(TERM))))

図 6: TDMT と類推による構文木の評価例

5 おわりに

本稿では、ツリーバンクを用いて入力文に対する構文木を類推する機構を提案した。ここでは特に類推が正しく働く確からしさとして類推妥当性という尺度を導入し、これが構文解析の曖昧性解消のための一指標として有用であることを示した。今後、これらの指標の組合せ方などを検討する予定である。また、この類推機構が正しく動作するには、入力文全体と類似した文がツリーバンクに存在する必要があり、データのスパース性が問題となる。そこで句構造などの部分的な単位からの類推や品詞などの利用についても検討する予定である。

参考文献

[1] T. Briscoe, Robust Parsing, *Survey of the State of the Art in Human Language Technology* 第 3 章, `http://www.cse.ogi.edu/CSLU/HLTsurvey/ch3node9.html#SECTION37`, 1995.

[2] T. Fujisaki, F. Jelinek, J. Cocke, E. Black and T. Nishino, A probabilistic parsing method for sentence disambiguation, *Proceedings of the International Workshop on Parsing Technologies*, Pittsburgh, 1989.

[3] E. Sumita and H. Iida, Experiments and Prospects of Example-Based Machine Translation, In *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, pp. 185–192, 1991.

[4] Y. Lepage and S. Ando, Saussurian Analogy: a theoretical account and its application, In *Proceedings of Coling-96*, vol. 2, pp. 717–722, 1996.

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

[6] R.A. Wagner and M.J. Fischer, The String-to-String Correction Problem, *Journal for the ACM*, Vol.21, No.1, pp. 168-173, 1974.

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