用例との多元的類似度計算に基づく文脈依存の構文解析法

1G－4

側嶋康博　Mark Seligman
ATR音声翻訳通信研究所
{sobasima, seligman}.itl.atr.co.jp

1. はじめに

本稿では、構文、意味、隣接する語句の文脈を含めた、用例との多元的類似度計算により、ボトムアップに構文解析を行う手法を提案する。本手法は、翻訳および文の解析のために用いられ成果を収めている用例との意味的近さ（または類似度）に基づく手法[1],[2],[3]を拡張することにより、頑強でかつ精度の高い構文解析を効率的に行う。

本手法の特徴は、第一に、多元的類似度を導入することにより頑強な解析が可能となることである。すなわち、意味的に類似する用例がない場合、構文的類似度に基づく選択ができ、構文･意味的に類似する複数パターンに対しては左右の語句と用例との類似度計算により、文脈に依存した選択ができる。また、ボトムアップな解析の各段階でトップダウン制約（隣接語句制約）がかかるため、全解探索やバックトラックなしに効率的な構文解析が可能となる。

本手法の有効性を確認するため、日英両言語の構文解析実験を行い、良好な結果を得た。

2. 発話の構造表現

本研究は、話し言葉の発話を対象としている。また、バイリンガル･コーパス[4]の日英表現の対応の便宜を考え、応答語句（「はい」「yes/okay」など）および表現末尾の働きかけ表現（「～さん」「right?」など）までをまとめて構文解析の単位（メッセージ[5]）としている。

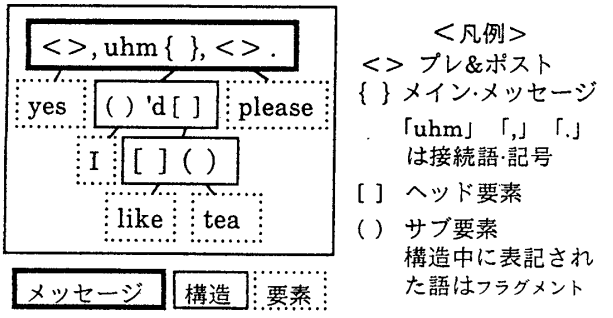

図1「Yes, uhm I'd like tea, please.」の構造表現例

A context-dependent parsing method for example-based analysis integrating multiple knowledge sources
Yasuhiro Sobashima and Mark Seligman
ATR Interpreting Telecommunications Research Laboratories

メッセージは、プレ、メイン、ポストの3通りのメッセージ要素と接続語･記号を持つ。おのおの、語（ここでは「要素」と呼ぶ）、または複数の語からなる「構造」で構成される。

構造は要素の並びと各要素の「役割」で規定する。役割にはヘッド、サブ、フラグメントの3通りがある(図1参照)。

3. 構文解析の概要

3.1 構文用例データの構成

類似度計算に基づく構文解析を行うため、図2に示す構文用例データを用意する。

図3には、構文用例データの例を示す。

構文用例データ：$"N \mid R \mid C_l \mid e_1 \mid e_2 \mid \ldots \mid C_r"$

N	構造を作る要素の数
R	構造全体の属性（機能･意味）セット
e_i	i番目の要素の属性（機能･意味･形態）セット
C_l	左に隣接する要素の属性セット（左文脈要素）
C_r	右に隣接する要素の属性セット（右文脈要素）

図2　文脈付き構文用例データ

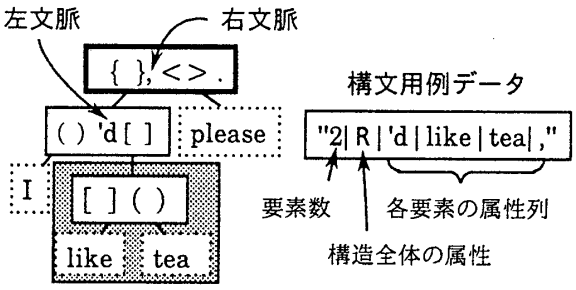

図3「like tea → like + tea」の構文用例データ

3.2 ボトムアップ構文解析法

形態素解析された要素列に対して、要素の組合せを変えながら構文用例データを適用し、用例との類似度（4参照）が最も高い組合せを採用して新たなノードを作る。この操作をメッセージ･ノードができるまで繰り返す（図4）。このように、全解探索、バックトラックせず、ボトムアップに構文解析を行う。

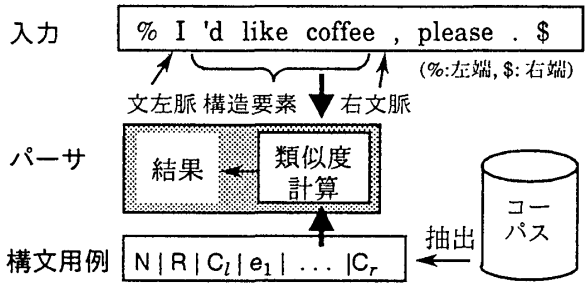

図4　文脈付き用例を用いた構文解析

4. 類似度計算

4.1 要素間の類似度

2つの要素 $A(A_f,A_s,A_g)$, $B(B_f,B_s,B_g)$ (添字 f, s, g は要素の機能、意味、形態の各属性を表す) 間の類似度 E-Sim を、要素間の機能、意味、形態の各類似度 (Sim-f, Sim-s, Sim-g) の関数であると定義する。また、各類似度は 0から 1の範囲で正規化する。

E-Sim(A,B)
$= f_1$ (Sim-f(A,B), Sim-s(A,B), Sim-g(A,B)) 　(式1)

4.2 構造間の類似度

同数の要素を持つ2つの構造 A, B について、構造間の類似度 S-Sim を、対応する要素間の類似度を用いて以下のように定義する。

$$\text{S-Sim}(A, B) = \alpha \Sigma_i \text{E-Sim}(A_i, B_i) * W_i \quad \text{(式2)}$$

ここで、A_i, B_i は A, B の i 番目の要素であり、W_i は重み、α は正規化係数 ($=1/\Sigma_i W_i$) である。

4.3 隣接する語句の類似度 (文脈的類似度)

2つの表現 A, B がそれぞれ左右に語句を a_1, a_2, b_1, b_2 のように伴っている場合を考える (下図)。

..., a_1, A, a_2, ... 　　..., b_1, B, b_2, ...

A, B 間の文脈的類似度 C-Sim を、左右の語句それぞれの要素の類似度を用いて式3で定義する。

$$\text{C-Sim}(A,B) = \text{E-Sim}(a_1,b_1)*\text{E-Sim}(a_2,b_2) \quad \text{(式3)}$$

4.4 統合した類似度

構文解析においては、統合した類似度 Sim (式4) を用いてノード作成の可能性を判定する。すなわち、ある構造を作成するかどうかを、その左右に隣接する語句の文脈とともに判断する。

$\text{Sim}(A,B) = f_2$ (S-Sim(A,B), C-Sim(A,B)) 　(式4)

5. 実験

ATRで収集している旅行会話コーパス[5] から構文データ (日英各言語 400メッセージ) を抽出し、表1の設定で、学習済みデータを使った構文解析実験を行った。

表1　実験における設定

属性	機能:3階層 72(日)/95(英) 分類および活用 意味:3階層 304分類 (日英共通) 形態:(一致または不一致の2通り)
類似度	Sim＝S-Sim*(1＋C-Sim)/2 E-Sim ＝ Sim-f*(1＋Sim-s) (1＋Sim-g)/4 Sim-f, Sim-s: 一致桁数/ 全桁数, Sim-g:0, 1
重み	head:10, sub:1, pre:1, mian:5, post:1

表2に、ランダム抽出した試験文2セット (各20メッセージ) に対する文脈類似度計算付きと無しの場合の構文解析実験の結果 (構造正答率) を示す。

今回の実験では最高得点の解を択一的に選択したため、文脈計算無し実験で約2割の誤りがある。それらは通常、全解探索かバックトラックをかけて回避するが、文脈付き実験では、この方法ですべて正しい構造を作った。ただし、依然、意味属性は0.6% (日), 0.2% (英) について選択の誤りがあり、その他の情報を用いて曖昧性を解消する必要がある。

表2　実験結果

	文脈計算付き実験	文脈計算無し実験
構造正答率	100%(日) 100% (英)	81.3% (日) 80% (英)

6. まとめと課題

隣接語句の文脈を含む用例との多元的な類似度計算を用いて、高精度で効率的なボトムアップ構文解析が可能であることが示された。

しかし、用例を増やしながら様々な入力に対する適用性を調べる必要がある。また、モデルを詳細化すること、計算式および重みについて妥当性を検討することが今後の課題である。

参考文献

[1] M. Nagao, "A Framework of a Mechanical Translation between Japanese and Engilsh by Analogy Principle," in Artifical and Human Intelligence, eds. A. Elithorn and R. Banerji, North-Holland, pp. 173-180 (1984).

[2] 隅田 ほか·"英語前置詞句係り先の用例主導あいまい性解消" 信学論 Vol. J77-D-II No.3 pp.557-565 (1994)

[3] 古瀬 ほか "経験的知識を活用する変換主導機械翻訳" 情処学論 Vol. 35 No. 3, pp. 414-425 (1994)

[4] O. Furuse, *et al*, "Bilingual corpus for speech translation, " Proc. of AAAI -94 Workshop on "Integration of Natural Language and Speech Processing," pp. 84-89 (1994)

[5] 側嶋 "バイリンガル·コーパスを用いた対話文翻訳のための局所文脈解析" 第47回情処学会全国大会 Vol. 3 pp. 3-205-206 (1993)