非線形言語モデルに基づく文型パターン型言語知識ベースの開発

Development of Language Knowledge Base in Sentence Patterns based on Non-Compositional Language Model

徳久雅人　　池原 悟　　村上仁一
Masato Tokuhisa　Satoru Ikehara　Jin'ichi Murakami
鳥取大学　工学部　知能情報工学科
Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University

1 はじめに

言語解析のためにパターンが広く使用されている. たとえば, 日英機械翻訳, 言い換え, 情報抽出, 感情推定などが挙げられる[1][2][3][4][5]. パターンについて, 単にルールや表現構造の記述形式の1つとして捉えられ, 言語解析上の位置付けがあまり議論されてこなかった. 本稿では, 非線形言語モデル[6]を軸として, 文型パターン型の言語知識ベースの性質を示し, 知識ベースの開発方法について分類する.

2 文型パターンの解析能力

2.1 言語表現の概念と意味処理

言語表現は, 単語, 句, 節, 文, 文章である. 単語の文字通りの解釈は, 人々の間で共通性がある. 言い換えると, 単語に対応する概念についての社会的な約束がある. 単語の集まりである文も同様に文字通りの解釈は共通性がある. ここで, 1つの単語の表す概念を「単一概念」と呼び, 複数の単語による表現の表す概念を「複合概念」という[6].

言語表現の生成において, 話し手は, 言わんとする対象の特徴にできるだけ一致する概念を表す言語表現を用いる. 逆に, 言語表現の理解においては, 聞き手は, 言語表現の表す概念を自己の内部に取り入れ, 話し手の言わんとする対象の特徴を内部に再構成する.

ここで, 言語の意味処理を実現する上で次の 2 点が区別されている[7]:

- **(A) 意味解析**: 言語表現に社会的約束として対応する概念を求めること
- **(B) 意味理解**: 聞き手として, 言語表現に対応する概念を他と関係付けて内部に持つこと

2.2 非線形言語モデル

「非線形言語モデル」は, 言語表現の構造と概念の関係に着目した意味解析のモデルである[6]. 表現構造とは, 言語表現の組み合わせ方である. 言語表現の組み合わせ方も, 解釈の仕方に人々の間の共通性がある. すなわち, 表現構造に対応する概念についての社会的な約束がある.

言語表現の表す複合概念は, 言語表現のどの一部でも変更すると, 変化してしまう. たとえば, 「母は父より若い」という言語表現の表す概念は, 「ママはパパより若い」の表す概念と異なる. 一方, 表現構造の表す概念は, 抽象的である. たとえば, これら 2 つの例文は, 「〜は〜より〜」という構造であり, 「 2 者の比較」という抽象的な概念を表していると言える.

言語表現のうち, 表現構造を形成するための部分表現を変更すると抽象的概念が変化してしまう. たとえば, 上記の 2 つの例文では, いずれも, 「2 者の比較」を表しているが, 「母も父より若い」と変更すると「他の者も比較されていたこと」が伺えるため, 抽象的概念が変化してしまう.

非線形言語モデルでは, 表現構造を形成するための部分表現を「非線形要素」と呼び, 抽象的概念が変化しない範囲で書き換え可能な部分表現を「線形要素」と呼ぶ. 線形要素は, 言語表現なので, 複合概念を表す限り, 再帰的に分解できる.

2.3 文型パターンの作成

線形要素と非線形要素の関係を記号処理するために記述したものが「文型パターン」である. 文型パターンは基本的に言語表現から線形要素と非線形要素を区別することで作られる. 非線形要素は, 言語表現のうち, 表現構造か抽象的概念かの少なくとも一つを固定することで, 言語表現から抽出でき, 一方で, 線形要素はその抽象的概念を変化させない程度に汎化可能な部分表現として抽出できる.

たとえば, 上記の 2 つの例に対して, 「 2 者の比較」という抽象的概念でとらえるとき, 「川は山より流れる」という表現には「〜は〜より〜」という字面が含まれているが, 「比較」は表していない. 線形要素として品詞に制限があることが分かる. さらに, 抽象的概念として主題が問われないので, 「〜より〜は〜」というように, 格要素の位置の異なる構造も同一視できる. したがって, パターンを作ると, 「{ *N*1 は, *N*2 より} *AJ*3 」となる. ここで, *N* と *AJ* はそれぞれ名詞と形容詞の変数, {〜, 〜} は順序任意記号である. これらの記述子の仕様は, 文献[2]で提案されている.

以上の通り, 言語表現から構造と抽象的概念に着目して文型パターンを作成することができる. しかし, 実際には抽象的概念を明示的に扱うコストが問題となり, 工夫を要する (第 3.2 節).

2.4 文型パターンの照合検索と応用

文型パターンは, 表現構造の種類数だけ作成される. 通常, 応用のために, 文型パターンには, 表現構造の表す概念に関連する情報を付随させるので, 文型パターンを収録した言語知識ベースは, 「文型パターン辞書」とも呼ばれる.

本辞書を用いて意味を解析する過程は次のようになる:

(1) 解析対象の言語表現に適合する文型パターンを, 文型パターン辞書から検索する.
(2) 適合した文型パターン (1個以上) の中から, 適切なものを選択する. その結果は, 意味解析結果に相当する.

概念そのものを計算機上で処理する方法は現在確立できていないので, 概念を経由して応用する際に必要な情報を, 概念の代わりに文型パターンにあらかじめ付随しておく.

3 知識ベースの開発

文型パターンの作成, および, 応用のための情報付与は, 抽象的概念が捉えられるかどうかに依存して, 実現方法が異なる. 具体的取り組みについて紹介する.

3.1 抽象的概念を特定しながら作成する方法

3.1.1 決定できる場合の作成事例

情緒生起の原因に基づく情緒推定という応用のために, 日本語語彙大系[8]の文型パターン辞書を拡張した[9]. この文型パターンからは, 用言を中心とした語義が抽象的概念として読み取れる. 情緒生起の原因についての概念分析をあらかじめ行っておき[10], その原因の特徴と語義の一致を人手で判断しながら, 情緒情報を対応付けた. 線形要素である格要素の変数は, 情緒主や情緒対象として関係付けた. その結果, 約 14,800 パターンに対して約 11,700 件の情緒情報が対応付けられた. なお, 「獲得」のように語義を捉えるまでは(A)意味解析であり, 獲得が起因して生じる情緒を推定することは(B)意味理解である.

文型パターン: *N*1 (3 主体) が *N*2 (533 具体物…) を *N*3 (*) から/より 入手する
用言意味属性: 所有的移動　情緒生起原因: 獲得
情緒名: 喜び　情緒主: *N*1　情緒対象: *N*2

情緒状態を明示する用言や情緒的反応を表す用言については, 文型パターンの表す用言の語義を直接的に分析することで可能である[11]. なお, これは(A)意味解析に位置付けられる.

3.1.2 曖昧な場合の作成事例

用言を中心とした場合は, 語義が明確なので1件ずつ手作業でパターンの拡張ができた. しかし, 文末表現に着目した情緒推定に応用する際, 文末表現の抽象的概念が定め難いためパターン化が容易ではない.

そこで, 話し手の情緒をタグで表したタグ付きコーパスを作成し, タグの共起関係から, 情緒を表す可能性のある文末表現を収集した[12]. 文末表現を非線形要素に, 文末表現に先行する表現を線形要素にしてパターン化した. さらにパターンには意図情報を追加した. こうして, 約 1,000 件を得た. 文末表現は元来話し手の主体性を表すので, 文末表現と共起する情緒や, 追加した意図情報が抽象的概念に相当する.

これにより, 話し手の意図と情緒が解析できるのだが, さらに, 抽象的概念を介して同一意図の別パターンを選出し, 同一意図で異なる情緒を表す文に書き換えることも可能である[13].

文型パターン: (*VP*1|*AJP*1)^shushi かも…。
情緒成分: 悲しみ(25%), 恐れ(58%), 嫌だ(8%), 驚き(8%)
意図:　〔伝達〕

3.2 抽象的概念を特定せずに作成する方法

日英機械翻訳への応用において, 日英の表現に対する抽象的概念が, 現実的には見出すことが難しい. たとえば, 日本語の重文･複文について意味的分類[14]は, 大量の文に対してその分類を割り当てることが難しい. 機械翻訳においては, 抽象的概念が得られることよりも, それを介した対訳側のパターンの得られることの方が応用に直結する.

そこで日英対訳文対において, アライメントのとれる部分表現のうち線形要素と言えるものを判別することでパターン化を行った. 線形要素の最大の大きさを, 単語, 句, 節と 3 レベルに設定することで, 約 12 万文対から 22.7 万件の日英パターン対を作成した[2]. 統計翻訳や用例翻訳ではアライメントのとれる部分, すなわち線形要素を扱うが, 本辞書の特徴は, アライメントのとれなかった部分 (非線形要素) が明示的な知識となる点である.

原文:明け方になってやっととろとろ眠った。
対訳: I dropped into a fitful doze at dawn.
WJ : /ytcfk*TIME*1(IM:16930,IM:16970)に!(なっ|成っ)て< /y*N*2(IM:11100)は>!やっと!とろとろ/f 眠った。
WE :<I | *N*2> dropped into a fitful doze at *N*1.
PJ :/ytcfk*TIME*1(IM:16930,IM:16970)に!(なっ|成っ)て< /y*N*2(IM:11100)は>!やっと!*VP*3(IY:1120,IY:3210).kako 。
PE :<I | *N*2> *VP*3^past at *N*1.

4 おわりに

非線形言語モデルに基づく文型パターン型言語知識ベースは, 原理上, 表現構造の表す抽象的概念を解析することができ, パターン毎に対応付けられた情報を介して応用に結びつけることができる. 言語知識ベースの開発には, 抽象的概念を明示的あるいは非明示的に意識した方法があることを示した.

参考文献

[1] 池原, 宮崎, 白井, 林:言語における話者の認識と多段翻訳方式, 情報処理学会論文誌, **28**(12), pp.1269-1279, 1987.
[2] 池原, 阿部, 徳久, 村上: 非線形な表現構造に着目した重文と複文の日英文型パターン化, 自然言語処理, **11**(3), pp.69-95, 2004.
[3] Takahashi,T., Iwakura,T., Iida,R., Fujita,A., Inui,K.: KURA: A Transfer-Based Lexico-Structural Paraphrasing Engine, NLPRS, pp.37-46, 2001.
[4] 川浪, 大熊, 増市, 杉原, 石崎: ウェブからの情報抽出システムの構築-定義型質問に対する情報検索に基づく回答の作成-, 言語処理学会第 12 回年次大会発表論文集, pp.797-780, 2006.
[5] 松本, 三品, 任, 黒岩: 感情生起事象文型パターンに基づいた会話文からの感情推定手法, 自然言語処理, **14**(3), pp.239-271, 2007.
[6] 池原: 言語で表現される概念と翻訳の原理, 電子情報通信学会技術研究報告, TL2003-25, pp.7-12, 2003.
[7] 池原: 自然言語処理の基本問題への挑戦, 人工知能学会誌, **16**(3), pp.422-430, 2001.
[8] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林: 日本語語彙大系, 岩波書店, 1997.
[9] 田中, 徳久, 村上, 池原: 情緒生起情報付き結合価パターン辞書の開発, 言語処理学会第 12 回年次大会発表論文集, pp.1151-1154, 2006.
[10] 徳久, 岡田: パターン理解的手法に基づく知能エージェントの情緒生起, 情報処理学会論文誌, **39**(8), pp.2440-2451, 1998.
[11] 黒住, 徳久, 村上, 池原: 結合価パターン辞書からの情緒を明示する用言の知識ベース化, 言語処理学会第 13 回年次大会発表論文集, pp.39-42, 2007.
[12] 徳久, 村上, 池原: 漫画における表情に着目した情緒タグ付きテキスト対話コーパスの構築, 自然言語処理, **14**(3), 192-217, 2007.
[13] 前田, 徳久, 村上, 池原: 情緒を表す文末表現の書き換えの試行, 電子情報通信学会ソサイエティ大会講演論文集, 基礎･境界, (投稿中).
[14] 衛藤, 池原, 佐良木, 宮崎, 池田, 新田, 白井, 柴田: 意味類型構築のための文接続表現の体系化, 情報処理学会研究報告, 2003-NL-155-6, pp.31-38, 2003.