用言意味属性を用いた適合文型パターンの絞り込み

前田 春奈　村上 仁一　徳久 雅人　池原 悟
鳥取大学 工学部 知能情報工学科
{hmaeta,murakami,tokuhisa,ikehara}@ike.tottori-u.ac. jp

1　はじめに

従来のパターンベース機械翻訳では, 使用するパターンの規模が小規模であったが, 池原等により 10 万件規模のパターン辞書が構築された [1]. [2] によると, 12 万件規模の単語レベル[*1]文型パターンを使用する場合, 1 入力文につき平均 50 パターンが適合することが報告されている. しかし, 適合パターンには, 正しい英文を生成するために有効なものもあれば, 不適切なものも含まれているため, 正解率の低下の原因の 1 つになっている.

翻訳の際, 適切なパターンを使用するためには, (1) パターン辞書から入力文に適合するパターンを検索する際の条件を強化すること, (2) 適合したパターンの中から最適なパターンを選択すること, の大きく 2 つの対策が考えられる. (2) については岡田や原による取り組みが進められている [3][4].

本研究では (1) を目的とする. 具体的には, 文型パターンにおける用言の変数に対して「用言意味属性」による制約条件をかける方法を試みる. そして, 制約となる用言意味属性の抽象度を変化させて, 適合率および正解率の変化を調査する. 最後に, 適合率を低下させずに正解率を向上させる条件について考察する.

2　用言意味属性を用いた適合文型パターンの絞り込み

2.1　利用するデータベースと文型パターンとプログラムと用言意味属性

実験に利用するデータベースと文型パターンとプログラムと用言意味属性を以下に示す.

(1) **評価用試験文**

実験用試験文は, 日英対訳データベース (12 万文 [5]) よりランダムに選択した 110 文を使用する. 例文を以下に示す.

- 試験文例 = 彼にはその任務を果たせるだけの能力がなかった。
- 模範訳 = He was not equal to the task.

(2) **文型パターンデータベース**

文型パターンは, 日英言語の表現対において, 線形要素を変数記号に書き換えたパターンである [6]. 本研究は文型パターン辞書に収録された単語レベルの文型パターン (12 万対) を使用する. この辞書には, 単語レベルの他, 句レベル, 節レベルが収録されており, その中の単語レベルは表現に含まれる名詞, 動詞, 形容詞, 副詞などの自立語の線形な要素を変数化している.

(3) **文型パターンパーサ**

文型パターンパーサ [7] は入力文 (日本語) と 12 万の単語レベル文型パターンを ATN[8] を用いて照合し, 入力文に適合する文型パターンを文型パターンデータベースから複数出力する.

(4) **用言意味属性**

本研究では, 日本語単文意味分類体系の用言意味属性 [2] と日本語語彙大系の用言意味属性 [9] を使用する.

日本語単文類意味分類体系の用言意味属性 (以下, IY) は用言が 4 桁の数字を用いて 4 段階に分類されている. IY は使用する桁数により第 1 分類から第 4 分類まで意味を分ける事ができ, 使用する桁数が多いほどより細かい意味が表されている. IY の例を以下に示す.

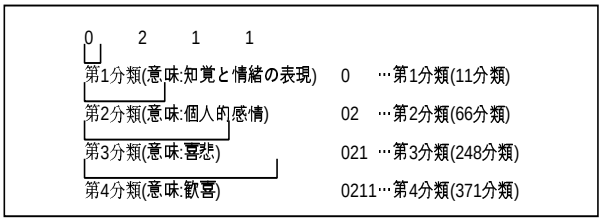

図 1: IY の意味分類例

また, 日本語語彙大系の用言意味属性 (以下, NY) は用言が 36 に分類されている.

2.2　実験方法

実験の手順を以下に示す.

＜1＞**入力文と模範訳**

入力文および模範訳を用意し, 入力文の動詞部分に対応する用言意味属性を付与する. 括弧内は用言意味属性を表す.

- **入力文** = 道路を横断する (**通過迂回**) ときは車に注意し (**叱咤非難**) なさい。

[*1] 単語レベル (122,718 パターン), 句レベル (80,160 パターン), 節レベル (12,456 パターン), /全レベル (215,334 パターン)

- **模範訳** = Watch out for the traffic when you cross the street.

＜2＞**文型パターンパーサによる文型パターンの抽出**

文型パターンパーサに文を入力し, 適合する文型パターン全てを文型パターンデータベースから抽出する.

- <文型パターン例 1>
  (**抽出された日本語パターン**)
  /y </tk $N1$ は> /$tcfk$ $N2$ を /cf $V3$(音声発話)^$rentai$！ ときは /$tcfk$ $N4$ に /cf ($V5$(叱咤非難)^$meirei$|$V5$.$meireigo$|$ND5$ をしなさい)。
  (**英語パターン**)
  ($V5$|$ND5$) to $N4$ when <you|$N1$> $V3$ $N2$.
- <文型パターン例 2>
  (**抽出された日本語パターン**)
  /y </tk $N1$ は> /$tcfk$ $N2$ を /cf $V3$(通過迂回)^$rentai$！ ときは /cf ($V4$(叱咤非難)^$meirei$|$V4$.$meireigo$|$ND4$ をしなさい)。
  (**英語パターン**)
  Be careful when <you|$N1$> $V3$ $N2$.

＜3＞**用言意味属性を用いた, 文型パターンの絞り込み**

入力文の動詞部分の用言意味属性と抽出された文型パターンにおける日本語パターンの動詞部分の用言意味属性を比較する. もし意味属性が一致しない場合は削除する.

- <削除された文型パターン (文型パターン例 1)>
  (**日本語パターン**)
  /y </tk $N1$ は> /$tcfk$ $N2$ を /cf $V3$(音声発話)^$rentai$！ ときは /$tcfk$ $N4$ に /cf ($V5$(叱咤非難)^$meirei$|$V5$.$meireigo$|$ND5$ をしなさい)。
  (**英語パターン**)
  ($V5$|$ND5$) to $N4$ when <you|$N1$> $V3$ $N2$.

＜4＞**英語パターンを使った出力英文の作成**

手順 3 の絞り込みで残った文型パターンにおける英語パターンの変数部分に, 入力文に合った英単語を人手で代入し, 出力英文とする.

- <残った文型パターン (文型パターン例 2)>
  (**日本語パターン**)
  /y </tk $N1$ は> /$tcfk$ $N2$ を /cf $V3$(通過迂回)^$rentai$！ ときは /cf ($V4$(叱咤非難)^$meirei$|$V4$.$meireigo$|$ND4$ をしなさい)。
  (**英語パターン**)
  Be careful when <you|$N1$> $V3$ $N2$.
  (**出力英文**)
  Be careful when you cross the street.

＜5＞**出力英文の評価**

すべての出力英文に対して A から C の 3 段階で評価を行う (2.3 節の評価方法参照).

- (**評価**)
  Be careful when you cross the street.(A 評価)

＜6＞**適合率および正解率の算出**

すべての出力英文に対して評価を行った後, 評価結果から適合率および正解率を算出する.

2.3 評価方法

出力英文の評価は人手によって行う. 判断基準は以下の 3 段階とする.

- A…英語パターンの変数部分に単語を代入すれば理想的な訳を作成可能
  <評価 A の例>
  (**入力文**)
  私は子供の将来を思うと切ない。
  (**模範訳**)
  I get distressed when I think of my children's future.
  (**パターン (英)**)
  $N1$ 'get'#2(^$present$|^$past$) stressed out when $N1$ $V5$ #1[$N1$^$poss$ $N3$]'s $N4$.
  (**出力英文**)
  I get stressed out when I think of my children's future.
- B…パターンの問題 (前置詞, 冠詞やパターン表記ミス) を解決すれば理想的な訳を作成可能
  <評価 B の例>
  (**入力文**)
  私は彼を頼って上京した。
  (**模範訳**)
  I came to Tokyo counting on his help.
  / (**パターン (英)**)
  <I|$N3$> N($V4$|$ND4$) where <I|$N1$> could rely on <my|$N1$^$pron$^$poss$> $N2$.
  (**出力英文**)
  I went to Tokyo where I could rely on my he.
- C…入力文に合った訳が作成不可能
  <評価 C の例>
  (**入力文**)
  母は赤ん坊をあやして笑わせた。
  (**模範訳**)
  Mother played with the baby and got him to smile.
  (**パターン (英)**)
  $N1$ $V3$.$past$ $N1$^$pron$^$poss$ $N2$ $V4$.
  (**出力英文**)
  Mother played her baby smile.

2.4 適合率および正解率の算出方法

適合率 (以下 $R1$) と正解率をそれぞれ算出する. 正解率については以下の 2 通りの評価を行う.

1. 全ての入力文に対して抽出された全ての文型パターンについて, A および B と評価された文型パターンが存在する割合 (文型正解含有率 (以下

$P1$))

2. 入力文に対して A および B と評価された文型パターンが 1 つでも存在する割合 (適合文型正解含有率 (以下 $P2$))

$R1$, $P1$ および $P2$ は以下の式でそれぞれ求める.

- $R1 = \frac{\text{自己パターン以外のパターンが 1 つ以上抽出された文数}}{\text{入力文数}}$
- $P1 = \frac{A,B \text{ と評価されたパターン}}{\text{抽出された全パターン}}$
- $P2 = \frac{1 \text{ つ以上 } A,B \text{ と評価されたパターンを持っている文数}}{\text{自己パターン以外のパターンが 1 つ以上抽出された文数}}$

2.5 実験条件

用言意味属性を用いた絞り込み実験の実験条件は以下の 6 つの条件とする.

1. 文法情報:用言意味属性を用いた文型パターンの絞り込みを行わず, 文型パターンパーサで抽出した全ての文型パターンに対して適合率および正解率を算出する.
2. IY 第 1 分類 (11 分類):文型パターンパーサで抽出した全ての文型パターンに対して、IY 第 1 分類を用いて絞り込みを行い, 絞り込まれた文型パターンパターンに対して適合率および正解率を算出する.
3. IY 第 2 分類 (66 分類):IY 第 2 分類を用いて絞り込みを行う.
4. IY 第 3 分類 (248 分類):IY 第 3 分類を用いて絞り込みを行う.
5. IY 第 4 分類 (371 分類):IY 第 4 分類を用いて絞り込みを行う.
6. NY(36 分類):日本語語彙大系で分類された用言意味属性を用いて絞り込みを行う.

3 実験結果

文型パターンパーサを用いて入力文 110 文と文型パターンデータベースを照合した所, 62 文が文型パターンを 1 つ以上抽出できた. 入力文は文型パターンデータベースには含まれていない. したがって, 本研究はオープンテストとなる.

文型パターンを 1 つ以上抽出できた 62 文の入力文中, 7 文は文に動詞が含まれていないため対象外とした. 対象となった 55 文に対して, 抽出できた文型パターンの総数は 1047 パターンであった. また, 抽出できた 1047 パターン中, 動詞部分が字面で表記されているパターンが 16 パターンあった.

動詞部分が字面で表記されている 16 パターンについては, 用言意味属性の条件に関係なく $P1$ および $P2$ の結果に含めた. 絞り込み実験の結果を表 1 に示す.

表 1: 実験結果

条件	$R1$	$P1$	$P2$	$R1 \times P1$	$R1 \times P2$
文法情報	53% (55/103)	14% (147/1047)	49% (27/55)	7.5%	26%
IY 第 1 分類 (11 分類)	44% (45/103)	19% (53/287)	51% (23/45)	8.1%	22%
IY 第 2 分類 (66 分類)	28% (29/103)	34% (33/97)	41% (12/29)	9.6%	12%
IY 第 3 分類 (248 分類)	20% (21/103)	57% (27/47)	43% (9/21)	12%	8.7%
IY 第 4 分類 (371 分類)	19% (19/103)	58% (26/45)	42% (8/19)	11%	7.8%
NY (36 分類)	41% (42/103)	19% (66/341)	50% (21/42)	7.9%	20%

(括弧内の数字は, $R1$ および $P2$ は文数を表し, $P1$ はパターン数をそれぞれ表す.)

表 1 より, $R1$, $P1$ および $P2$ において最も高い値はそれぞれ, $R1$ は 53%(文法情報), $P1$ は 58%(IY 第 4 分類), $P2$ は 51%(IY 第 1 分類) であった. また, NY を用いた場合, $R1$ は 41%, $P1$ は 19%, $P2$ は 50%であった.

文法情報と IY 第 1 分類の結果を比較すると, $R1$ は文法情報の方が高い値を示しているが, $P1$ および $P2$ については IY 第 1 分類の方が高い値を示している. よって, IY は多少ではあるが正解率を上げる効果があったと言える.

4 考察

4.1 意味属性の条件の違いによる実験結果の比較

情報検索において, 一般的には $R1 \times P1$ の値が高くなる条件を使用する. 実験結果より, $R1 \times P1$ の値は IY 第 3 分類が最も高い. しかし, IY 第 3 分類を使用した場合, 残った文型パターンの総数は 47 パターンとなり, 多くの文型パターンが絞り込みで削除されてしまった. 実際の翻訳では, さらに情報を付加して最終的に翻訳に最も適した文型パターンを選択する. したがって, 意味属性を使用した文型パターンの絞り込みの過程では, 出来るだけ翻訳に使用できる文型パターンを残す事が望ましい. よって, 用言意味属性を用いた文型パターンの絞り込みでは, IY 第 1 分類を使用するのが最も適していると考えている.

4.2 IY 第 1 分類と NY の実験結果比較

IY 第 1 分類と NY の実験結果を比較すると, 適合率および正解率は近い値を示している. よって, より少ない数で意味を分類している IY 第 1 分類が文型パターンの絞り込みに適していると言える.

また, 用言意味属性を用いた単語レベル文型パターンの絞り込みにおける適合率および正解率は, IY 第 1 分類と日本語語彙大系用言意味属性の再現率および正解率の値が近い事より, 本研究の結果から, 用言意味属性を用いた適合文型パターンの絞り込みにおいて, $R1$ は 44%, $P1$ は 19%, $P2$ は 51%程度が限界であると考えている.

4.3 字面と用言意味属性による意味制約

実験で A 評価および B 評価と判定されたパターンは合計 147 パターンあり, 147 パターン中, パターンの動詞部分が字面で表記されているパターンは 16 パターンあった. 動詞部分が字面で表記されているパターン例を以下に示す.

- **<動詞部分が字面で表記されているパターン例>**
  (**パターン (日)**)
  /y \$1ˆ{/$tcfk$ $N1$ は } /$tcfk$ $N2$ を /cf 頼って \$1 /ycf 上京した。

この 16 パターンはパターンの動詞部分に用言意味属性が付与されていないが, 字面での表記を用言の意味的制約として考えると最も厳しい意味的制約であると考えられる. よって, 実験ではパターンの動詞部分が字面で表記されている 16 パターンを, 使用する用言意味属性の条件に関係なく $P1$ および $P2$ の結果に含めた.

動詞部分が字面で表記されている 16 パターンの評価を調べてみた所, A と評価されたパターンが 3 パターン, B と評価されたパターンが 13 パターンであった. また, 入力文と同じ字面を含んだパターンを使用した場合の正解率は 100%であった. IY の第 1 分類を使用した文型パターンの絞り込み実験では正解率が 51%であった事から, 字面による意味制約は用言意味属性よりも効果的であると言える.

4.4 意味属性の付与誤り

本研究で使用した文型パターンの動詞部分に付与された意味属性に意味属性付与誤りが存在した. 以下に例を示す. 例中の日本語原文とは, パターンを作成する際に使用された日本語文である.

- (**日本語原文**)
  彼女は私を振って彼に走った。
  (**日本語パターン**)
  /y \$1ˆ{/$tcfk$ $N1$ は } /$tcfk$ $N2$ を \$1 /cf $V3$(回転振動)(て | で) \$1 /$ytck$ $N4$ に /cf 走った。

例の $V3$ には意味属性「回転振動」が付与されているが, 原文より $V3$ にはさらに意味属性「除去廃棄」を付与しなければならない. 抽出された全パターンからランダムに 25 パターン選択し, 意味属性の付与誤りを調べた所, 2 パターンに付与誤りがあった. 今後は, 文型パターンの動詞部分に付与されている意味属性をチェックし, 意味属性付与誤りを校正する必要がある.

5 おわりに

本研究では, 用言意味属性を用いた文型パターンの絞り込みを行い, 正解率を上げる為の最適な条件を明確することを目的とした. 単語レベル文型パターンにおける用言意味属性を用いた絞り込み実験の結果, 適合率 $R1$, 正解率 $P1$ および $P2$ において最も高い値はそれぞれ, $R1$ は 53%(文法情報), $P1$ は 58%(IY 第 4 分類), $P2$ は 51%(IY 第 1 分類) であった. また, IY 第 1 分類を用いた方が文法情報よりも正解率は高くなった。

今後はさらに適合率および正解率の値を向上させる方法を考えていく. 具体的には, 句レベルおよび節レベルの文型パターンを使用した用言意味属性による文型パターンの絞り込みを行い, 全レベルの文型パターンを用いた場合の適合率および正解率を求める予定である. また, 日本語単文意味分類体系は用言だけでなく名詞も意味を 371 に分類している. よって, 名詞意味属性を用いたパターンの絞り込みや, 名詞+動詞の意味属性を用いたパターンの絞り込みを行い, 適合率を低下させず正解率を上げる為に最適な意味属性の条件を明確にする予定である.

謝辞

本研究は, 科学技術振興事業団「JST」の戦略的基礎研究推進事業「CREST」における研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーによる言語の等価変換と生成技術」の支援により行いました.

参考文献

[1] 池原ほか:非線型な重文複文の表現に対する文型パターン辞書の開発, 情報処理学会, 自然言語処理研究会 2005-NL-170, pp.157-164, 2005-11.
[2] 池原:日本語単文の意味分類体系の検討, 「言語・認識・表現」研究会, 第 15 回辞書プロジェクト, 2004-9.
[3] 岡田ほか:多変量解析による最適文型パターンの選択方式, 言語処理学会第 11 回年次大会, pp.25-28, 2005.
[4] 原ほか:日英機械翻訳における多変量解析を用いた最適パターンの選択, 言語処理学会第 12 回年次大会, 2006(発表予定).
[5] 村上ほか:日本語英語の文対応の対訳データベース, 「言語・認識・表現」, 第 7 回年次研究会, 2002-12.
[6] 池原:非線形な言語表現と文型パターンによる意味の記述, 情報処理学会, 自然言語処理研究会, 2004-NL-159, pp.139-146, 2004-1.
[7] 徳久ほか:文型パターンパーサの試作, 言語処理学会第 10 回年次大会, pp.608-611, 2004.
[8] 長尾ほか:「自然言語処理」, ISBN4-00-010355-5, 岩波書店, 1996.
[9] 池原ほか:「日本語語彙大系」, ISBN4-00-130101-6, 岩波書店, 1997.