

「意味」による文型パターン検索方式の最適化

中村 聡 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

{s022036,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1 はじめに

現在、日英機械翻訳の分野において、大規模な日英文型パターン辞書が作成されている。この辞書は、「類推原理」による「意味的等価変換方式」に基づいて、非線形な言語表現構造に対する文型パターンを意味的に類型化したものである。日英の原文と単語レベル、句レベル、節レベルの文型パターンと真理項が収録されている。

真理項は、統語的分類と意味的分類に大別できる。統語的分類には、文種別、統語構造、節間キーワードがあり、意味的分類には、節間の意味分類、従属節の意味、主節の意味がある。

従来のパターン翻訳方式では、入力文の表現構造に適合する文型パターンをパターン辞書から検索して使用しているが、これらの真理項を利用すれば入力文と“意味”が類似した文型パターンの検索ができる可能性がある。

そこで本研究では、真理項を用いた英文パターン検索実験を行い、意味的に適切な英文パターンを出力するための最適な真理項の検索条件について調査する。

2 意味検索パターン辞書の概要

以下に使用する文型パターン辞書と辞書の内部に含まれる真理項を説明する。

2.1 パターン対の構成

本研究で使用する日英対訳文型パターン辞書には、単語レベル(12万件)、句レベル(9万件)、節レベル(1.2万件)の文型パターン(合計22万件)と、真理項が収録されている。本研究では、このうち単語レベルのパターンを実験の対象とする。日英対訳単語レベル文型パターン辞書(以下文型パターン辞書)の例を以下に示す。

1. 日本語パターン

$/y \$1^{\wedge} \{ /cf \#1[GEN2] /k N3 \text{ は } \} /tcfk NUM4$
で $\$1 /ytk NUM5$ だ。

(日本語パターンの例文)

このタバコは20本入りで1ドルだ。

2. 英語パターン

$\#1[AJ2] N3 \text{ be } NUM5 \text{ for } NUM4.$

(英語パターンの例文)

These cigarettes are a dollar for twenty.

3. 真理項

- (a) — 文種別 : 1
- (b) — 統語構造 : 並列
- (c) — 節間KW : 連用中止
- (d) — 節間の意味 : 順接的並列

(e) — 従属節の意味 : 判断指定の表現

(f) — 主節の意味 : 判断指定の表現

2.2 真理項の詳細

本研究で用いる真理項とは、文章の性質を分類する情報のことである。真理項の分類は、統語的構造と意味的分類に大別できる。統語的構造の分類には、文種別、統語構造、節間キーワードがあり、意味的分類には、節間の意味分類、従属節の意味、主節の意味がある。各分類の説明を以下に示す。

1. 文種別

本項目は、文接続と埋め込み文の数により以下の5つに分類されている。

(a) 文種別1: 文接続を1ヶ所もつ文。最も単純な重文に相当する。

(b) 文種別2: 文接続を2ヶ所もつ文。

(c) 文種別3: 埋め込み文を1つもつ文。最も単純な複文に相当する。

(d) 文種別4: 埋め込み文を2つもつ文。

(e) 文種別5: 文接続を1ヶ所もつ、埋め込み文を各1つもつ文。

本研究では、翻訳評価を明確にする為に、文構成の理解と翻訳が容易な文種別1(重文)と文種別3(複文)に限定して研究を行う。以下に具体例を示す。

• 例文1: 文種別1

わたしは沖縄出身なので 暑さは平気です
副詞節 主節

• 例文2: 文種別3

先生になろうと 考えた。
補足節 述部

2. 統語構造

本項目は、節や部、詞の修飾方法で分類されている。例文3では、「彼の部屋」と「私の部屋」は別々の存在であり、1文で同時に表現していることから、並列節が、主節を修飾していることが判る。

• 例文3: 並列節

この部屋は彼の部屋であれば私の部屋です

3. 節間KW

本項目は、節と節を繋ぐ字面の分類である。

例文4では、「~を受け」と用言で節が区切られているため、「連用中止」と分類する。

(a) 例文4: 「連用中止」

ご逝去の知らせを受け、たいへん驚きました。

例文5では、「見る」の活用形の語尾、「見て」

の「て」により節が区切られているため、「て」が分類される。

- (b) 例文 5: 「て」
頃合いを見 て 株を売ればよい。

4. 節間の意味

本項目は約 140 個の分類の中から、節間位置、前方の節の繋がり先、節間の役割で分類している。例文 6 では、節を繋ぐ字面「から」の役割は「理由根拠」の分類であり、「から」の前に、「大金のことです」という副詞節があることを示す。

- (a) 例文 6: 「副詞節，理由根拠」
大金のことです から 早速には調達しかねます

例文 7 では、節間字面である「には」は、「一般名詞や固有名詞」を受ける性質があり、節間字面の前に、「自分のすること」という名詞修飾節がある事を示す。

- (b) 例文 7: 「名詞修飾節，一般名詞，固有名詞」
自分のすること には 完全主義者です。

5. 従属節の意味

本項目は従属節に付加されている意味を分類する。例文 8 は、真理項が複数該当する場合である。「大雨になり = 雨雲が発生」と考え「発生」の分類、「今晩は～になり，・・・が鳴る」は「今晩は」と限定するため、「別の日とは違う事 = 変化している」と考えられる。その結果、「変化の現象」，「事象の表現」の 2 つの項目になる。

- (a) 例文 8: 「発生」，「変化の現象と事象の表現」
今晩は 大雨になり、ところによっては雷が鳴るでしょう
例文 9 では、「おしまいだ」の表現から「終着する」という意味が考えられるため、「移動行為の表現」と分類する。

- (b) 例文 9: 「来る，戻る，到着」，
「移動行為の表現」
話もここまで来れば おしまいだ

又、本項目の分類例を示す。

- 第 1 段分類の例
 - ・ 知覚と情緒の表現
 - ・ 知的な行為の表現
 - ・ 社会的活動の行為
 - ・ 対物的行為
- 第 2 段分類の例
 - ・ 知覚感覚
 - ・ 個人的感情
 - ・ 開始終了
 - ・ 対象依存の行為
- 第 3 段分類の例
 - ・ 約束交渉
 - ・ 物の性状既定
 - ・ 調査研究
 - ・ 通知報道
- 第 4 段分類の例
 - ・ 性格
 - ・ 方向
 - ・ スポーツ
 - ・ 温度変化

6. 主節の意味

本項目の分類方法は従属節と同じである。従属節での説明の補足として、分類項目の概要を具体例と共に示す。

- ・ 例文 10: 「苦痛恐怖」，「性状既定の表現」
冬の寒い朝は寝床を離れるのが つらい
分類項目の概要

分類項目は、4 段の木構造を形成している。木構造による分類例を以下に示す。

- (a) 第 1 段分類
「日常の行為」，「現象事象の表現」，
「性状既定の表現」他 7 つ
- (b) 第 2 段分類: 「性状既定の表現」の下位の場合
「存在既定」，「感覚感情既定」他 4 つ
- (c) 第 3 段分類: 「感覚感情既定」の下位の場合
「感覚の既定」，「感情的性状既定」他 2 つ
- (d) 第 4 分類: 「感情的性状既定」の下位の場合
「喜び」，「苦痛恐怖」他 11 個
- (e) 「苦痛恐怖」分類の表現の一部
苦しい，つらい，恐ろしい，怖い

2.3 "意味" 検索プログラム

真理項を検索項目として、文型パターン辞書から入力文と真理項が一致する文型パターンを出力するプログラムを「意味」検索プログラムと呼んでいる。

この「意味」検索プログラムは、2 種類の入力が可能である。ひとつは、日本文を形態素解析によって各要素に分解し、真理項を抽出する。そして、抽出した真理項のうち、指定した真理項が一致する文型パターンを文型パターン辞書から出力する。もうひとつは、真理項を直接指定し、文型パターン辞書から文型パターンを出力する。

本研究では、入力文は、文型パターン辞書に使用される日本文を利用するため、後者の方式を文型パターン収集の方法に利用する。日本文を入力とした例を以下に示す。

1. 入力文
この刀は正宗の作だという
2. 入力真理項
文種別 : 3
統語構造 : 補足節
節間 KW : と
節間の意味 : 間接引用
従属節の意味: 判断指定の表現
主節の意味 : 音声発話, 言語表現一般, 伝搬拡散
3. 出力された英語文型パターンの一列
 - ・ $N1 V4 \#5 (^{present|^{past}}) to be$
 $N3 of N2.$
 - ・ $N1 'is' \#5 (^{present|^{past}}) V4.past to be$
 $NUM3 of N2.$
 - ・ $N1 V5 that N3 be AJ(N4) from N2.$
 - ・ $N1 V4.p passive to be the birthplace$
 $\#2[of N3].$

3 真理項を用いたパターン検索

3.1 実験の目的

本研究では、真理項を用いて文型パターンの検索実験を行い、意味的に適切な英文パターンを出力するための最適な真理項の検索条件について調査する。しかし真理項にはさまざまな検索条件が考えられるため、単純な条件で実験する。

3.2 真理項を用いたパターン検索の方法

以下に実験の手順を示す。

1. 入力文を用意する。
2. 入力文と文型パターン辞書を比較し、真理項を得る。
3. 抽出した入力文真理項における検索条件を設定する。
4. 英文パターンを得る。
5. 入力文と出力された英文パターンを比較し、評価する。

3.3 入力文の選定

文種別 1(重文)、文種別 3(複文) に属す文をそれぞれ 10 文選び入力文とする。

3.4 真理項の検索条件

文型パターンの検索条件は以下の 5 種類について行う。

1. 統語構造のみ
2. 節間 K Wのみ
3. 節間の意味のみ
4. 従属節の意味のみ
5. 主節の意味のみ

3.5 評価方法

英文パターンの評価で、入力文と英文パターンを比較し、英文パターンの変数、関数を人手で日本文の単語から英文の単語に翻訳し、得られた英文を評価する。以下に、評価基準と具体例を示す。なお、() 内は模範訳で [] 内は人手で作成した英文である。

1. 評価 A...入力パターンと意味、
文構造が共に類似
入力文：たびたびお手をかけてすみません
(I'm sorry to trouble you so often.)
英文パターン：<I|N1> be sorry to trouble
<you|N3^obj> so AJ(ADV2).
[I am sorry to trouble you so often.]
2. 評価 B...入力パターンと意味が通じる
入力文：親父が借金を残して死んだ
(His father left debts behind him.)
英文パターン：N1 V4.past, V3^ing behind
N2.
[His father died, leaving behind debts.]
3. 評価 C...辛うじて入力パターンの意味が
汲み取れる
入力文：彼は当惑面をして帰って来た
(He returned with a puzzled countenance.)
英文パターン：In N1^poss opinion, N3 should
be N2.
[In his opinion, he should be a puzzled countenance.]
4. 評価 D...入力パターンと意味がかけ離れている
入力文：人の性は善であると言われている
(It is said that men are born good.)
英文パターン：N1 V5 #2[AJ(N3)] pretty
N4.
[People say pretty born.]

3.6 データの例

以下に具体的な検索データ例を示す。この例では、従属節の意味のみを真理項の検索条件としている。

1. 入力文
私は彼が無罪だと思う。
[I assume him innocent.]
2. 真理項
文種別 : 3
統語構造 : -
節間 K W : -
節間の意味 : -
従属節の意味 : 判断措定の表現
主節の意味 : -
3. 検索英文パターンの一部
英文パターン：N1 V5#6(^present|^past) to
V4 AJ2 N3.
[I supposed to assume his innocence.]
4. 評価 ... B
主語と目的語の関係に誤りが無く、動作が入力文と同じである。しかし、「supposed」により、「assume」の動作が消極的になるため、評価を B とする。

3.7 実験結果

翻訳に利用する文型パターンを表中の 5 種類の真理項の条件で検索する。出力される英文パターンが多いため、1 つの条件で出力される英文パターンからランダムに 10 パターン選択し、1000 パターン (文種別 2 種類 × 入力文 10 文 × 条件 5 種類 × 英文パターン 10 件) 評価する。結果を表 1, 2 に示す。なお、表内の数字は英文パターン数である。

表 1 文種別 1 の評価結果

検索条件	A	B	C	D
統語構造	1	6	26	67
節間 K W	4	13	22	61
節間意味	2	13	26	59
従属節	2	15	19	60
主節	3	12	23	62

表 2 文種別 3 の評価結果

検索条件	A	B	C	D
統語構造	1	14	31	54
節間 K W	1	12	23	64
節間意味	1	12	27	60
従属節	6	25	25	44
主節	3	17	29	51

実験結果より、文種別 1 では、いずれの真理項の検索条件も、大きな差が見られないが、文種別 3 では「従属節の意味」が重要であることが判る。

4 複数の真理項を組み合わせたときの実験

本研究の目的は、意味的に適切な文型パターンを抽出する為の最適な真理項の条件の調査である。しかし、真理項のすべての組み合わせを検索条件として実験することは困難である。

そこで、前章で評価に使用した 1000 パターンを対象に、各パターンに付与されている 6 種類の真理項が、ど

れだけ入力文の真理項と一致するかを調べることにより、複数の真理項を指定した場合に抽出される英文パターンの品質を推定する。

以下に、指定する真理項の組を示す。これらの真理項の組は、1000 件のパターンの真理項の調査結果を参考に、品質の良いパターンが検索できそうな条件として決めた。

1. 条件
文種別 + 統語構造 + 節間キーワード + 節間の意味分類 + 従属節の意味 + 主節の意味
2. 条件
統語構造 + 従属節の意味 + 節間の意味 + 節間キーワード
3. 条件
統語構造 + 節間の意味
4. 条件
第 1 従属節 + 主節の意味

4.1 カバー率、正解率の計算方法

評価には、カバー率 $R1$ 、正解率 $P1$ 、 $P2$ を用いる。カバー率 $R1$ 、正解率 $P1$ 、 $P2$ の計算方法を以下に示す。

$$\text{カバー率 } (R1) = \frac{\text{出力がある入力文の数}}{\text{入力文の数}}$$

$$\text{正解率 } (P1) = \frac{\text{全パターン中に含まれる A,B 評価の数}}{\text{全出力英語パターン数}}$$

$$\text{正解率 } (P2) = \frac{\text{A,B 評価が 1 つでもある入力文の数}}{\text{出力がある入力文の数}}$$

4.2 推定結果

実験結果を表 3 および表 4 に示す。

表 3 文種別 1 の推定値

各条件	カバー率 $R1$	正解率 $P1$	正解率 $P2$ *1
条件	80%(8/10)	52.9%(9/17)	50.0%(4/8)
条件	100%(10/10)	20.4%(10/49)	60.0%(6/10)
条件	30%(3/10)	14.8%(4/27)	66.7%(2/3)
条件	20%(2/10)	50.0%(1/2)	50.0%(1/2)

表 4 文種別 3 の推定値

各条件	カバー率 $R1$	正解率 $P1$	正解率 $P2$ *1
条件	80%(8/10)	36.3%(12/33)	62.5%(5/8)
条件	90%(9/10)	22.5%(14/62)	100%(9/9)
条件	30%(3/10)	0.0%(0/4)	0%(0/3)
条件	0%(0/10)	(0/0)	(0/0)

この結果をみると、6 つの真理項をすべて一致させて検索することで、正解率 $P1$ が高く、条件 および ではカバー率 $R1$ と正解率 $P2$ が共に高いことが判る。つまり、大量の翻訳候補の中に 1 つでも翻訳に利用できる適切なパターンが含まれていることを意味している。

4.3 実際の検索条件による実験

前節の分析結果を実証するために、文種別 1 および文種別 3 において其々 2 文ずつ、出力された英文パターンから、条件、条件、条件 の条件毎に 10 文を選択する。そして選択された英文パターンの評価を行った。評価結果を表 5 および表 6 に示す。

*1 入力文に適したパターンからランダムに 10 件選んだとき、その中に A、B 評価が 1 つ以上ある場合。

表 5 文種別 1 の実測値

各条件	カバー率 $R1$	正解率 $P1$	正解率 $P2$ *1
条件	100%(2/2)	46.6%(14/30)	50%(1/2)
条件	100%(2/2)	3.3%(1/30)	50%(1/2)
条件	100%(2/2)	13.3%(4/30)	50%(1/2)

表 6 文種別 3 の実測値

各条件	カバー率 $R1$	正解率 $P1$	正解率 $P2$ *1
条件	100%(2/2)	33.3%(10/30)	100%(2/2)
条件	100%(2/2)	16.6%(5/30)	100%(2/2)
条件	100%(2/2)	10.0%(3/30)	100%(2/2)

この表 5、表 6 をみると、表 3、表 4 に近い値が得られている。よって表 3、表 4 は、ある程度信頼性のおける値であると考えている。

5 考察

表 1、2 の結果からわかる通り、単一の真理項を検索条件とした場合、全検索文から翻訳に使えるパターンは少ないことが判る。しかし、表 5、6 の結果より、6 つの全真理項 (文種別 + 統語構造 + 節間キーワード + 節間の意味分類 + 従属節の意味 + 主節の意味) を指定した場合、カバー率 $R1$ が 80% と、ある程度高い値が得られることが示された。また統語構造 + 従属節の意味 + 節間の意味 + 節間キーワードを指定した場合、正解率 $P2$ が 60.0% から 100% と高い値が得られることがわかった。

6 おわりに

本研究では、真理項を用いた英文パターン検索実験を行い、意味的に適切な英文パターンを出力するための最適な真理項の検索条件について調査した。

その結果、6 つの全真理項 (文種別 + 統語構造 + 節間キーワード + 節間の意味分類 + 従属節の意味 + 主節の意味) を指定した場合、カバー率 $R1$ が 80% と、ある程度高い値が得られることが示された。また統語構造 + 従属節の意味 + 節間の意味 + 節間キーワードを指定した場合、正解率 $P2$ が 60.0% から 100% と高い値が得られることがわかった。真理項のパラメータには木構造になっているものがあるため、今後これらの検索方法を考えていきたい。

7 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」の支援によるものである。

参考文献

- [1] 池原、文型パターンの分類方法の検討、研究資料、2005.
- [2] 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.
- [3] 藤本、埋め込み文に対する日英翻訳規則の作成について、言語処理学会第 10 回年次大会、pp.245-248、2004.
- [4] 前田ほか、パターンを使用した重文複文の日英翻訳の精度、言語処理学会第 10 回年次大会、pp.237-240、2004.