

概要

パターンに基づく日英機械翻訳方式の実現に向けて、単語レベル、句レベル、節レベルの3レベルで構成された重文複文を対象とする大規模な日英対訳パターン辞書が構築された。すでに、入力文と文型パターンを照合し、入力文に適合する文型パターンを検索する文型パターンパーサが試作されている。適合したパターンの中には英文生成に不都合なパターンも含まれている。大規模なパターン辞書を用いると、入力文に適合するパターン数が多いため、翻訳に適したパターンを選択する必要がある。

最適なパターンを選択する方法として、単語レベルのパターンを対象とした場合、多変量解析を用いた方法が有効である。しかし句レベルのパターンを対象とした場合、適合するパターン数が単語レベルに比べて非常に多いため、同様の方法は効果が不明である。

そこで、本研究では句レベルの場合について多変量解析による選択方式を適用し、有効性を調査する。まず入力文と適合パターンの情報を多変量解析により分析して評価関数を求める。求めた評価関数を用いて適合パターン選択実験を行うことで最適なパターン選択を行う。実験の結果、評価関数の値が0.75以上である場合は70%以上の正解率を得た。また単語レベルで行われた実験と同様の実験を行い、結果を比較したところ、単語レベルと句レベルではほぼ同等な結果となったことから、句レベルについての多変量解析を用いた選択方式の有効性を示すことができた。

目次

1	はじめに	2
1.1	本研究の目的	3
2	多変量解析を用いた選択方式	4
2.1	文型パターン	4
2.2	文型パターン辞書と文型パターンパーサ	5
2.2.1	文型パターン辞書	5
2.2.2	文型パターンパーサ	5
2.3	多変量解析	6
2.3.1	単語レベルの評価関数	6
2.4	評価関数	7
3	回帰係数の設定	11
3.1	回帰係数の作成条件	11
3.2	評価パラメータ x_i と評価値 y の事例	11
3.3	回帰係数の作成結果	13
4	評価関数を用いた適合パターン選択実験	14
4.1	実験目的と調査対象	14
4.2	実験手順	14
4.3	オープンテスト	14
4.4	実験結果	15
5	考察	17
5.1	推定値の信頼性	17
5.2	推定値による正解パターン選択失敗の原因	17
5.3	単語レベルとの比較	18
5.4	評価パラメータの考察	19
5.4.1	動詞意味属性距離について	19
5.4.2	寄与率	19
5.4.3	意味属性距離の問題点	20
5.5	推定精度	21

目 次

1	適合パターンの検索	5
2	推定値 1 位の正解率 (クローズドテスト)	15
3	推定値 1 位の正解率 (オープンテスト)	15
4	推定値 1 位の正解率 (クローズドテスト)	16
5	推定値 1 位の正解率 (オープンテスト)	16
6	意味属性距離	20
7	推定値の分布	21

表 目 次

1	要素記号の一覧	8
2	各評価における作成訳の例	12
3	入力文の比較	17
4	入力文の比較	18
5	入力文の比較	18
6	寄与率の比較	19
7	入力文の比較	21

1 はじめに

日英機械翻訳には、用例翻訳や統計翻訳等の手法があるが、翻訳の際に言語表現を分割することで全体の意味が失われてしまうことが問題であった。

そこでパターンに基づく日英機械翻訳方式の実現に向けて、単語レベル、句レベル、節レベルの3レベルで構成された大規模な日英対訳パターン辞書が構築された(各レベルの規模は単語レベル 123,451, 句レベル 95,496, 節レベル 123,451 パターンである)[1]。すでに、入力文と文型パターンを照合し、入力文に適合する文型パターンを抽出する文型パターンパーサが試作されている [2]。しかし大規模なパターン辞書を用いると、入力文に適合するパターン数が多いため、翻訳に適したパターンを選択する必要がある [3]。最適なパターンを選択する方法として、単語レベルのパターンを対象とした場合、多変量解析を用いた方法が有効である [4]。しかし句レベルのパターンを対象とした場合、入力文に対するカバー率が高く、適合するパターン数が単語レベルに比べて非常に多いため、同様の方法は効果が不明である。

そこで、本研究では句レベルの場合について多変量解析による選択方式を適用し、有効性を調査する。まず入力文と適合パターンの情報を多変量解析により分析して評価関数を求める。求めた評価関数を用いて適合パターン選択実験を行うことで最適なパターン選択を行う。本研究では以下、第2章では多変量解析を用いた選択方式について、第3章では評価関数作成について、第4章では適合パターン選択実験について、第5章では考察、第6章ではまとめを述べる。

1.1 本研究の目的

過去の研究によって、単語レベルについては多変量解析を用いた選択方式が有効であることが示されている [4]。句レベルパターンは単語レベルパターンよりもパターンの抽象度が高く、カバー率が上がり易いため、適合するパターンが非常に大きくなる。そのため単語レベルで示された多変量解析を用いた選択方式が、句レベルについても効果があるかは未確認である。本研究では句レベルについて適合パターン選択の精度を求め、多変量解析の選択方式の有効性の確認を行う。多変量解析による最適パターン選択の目標は、入力文により検索される適合パターンと、その英文の評価をもとに作成された評価関数による適合パターンの推定の評価値 (推定値) に基づき選択した結果、英文として許せる評価の範囲に含まれることである。そのため複数の適合パターンから英文生成に適したパターンを選択することが目的であり、適合パターンを1つに絞りこむまでの精度は必要としていない。本研究での最適パターンとは、入力文の意味が失われていない英文を生成できるパターンを指す。

2 多変量解析を用いた選択方式

2.1 文型パターン

文型パターンは可読性と網羅性を意識して設計されており, 字面, 変数, 関数, 記号で記述されている [5]. また, 日英文型パターン辞書 (次節参照) の原文を, 単語レベル, 句レベル, 節レベルの 3 レベルにパターン化した構造を持つ. ただし変数化すると対訳の訳出が困難である部分は変数化されず, 字面や関数の形式で残される. 原文を (L) とし, 句レベルのパターンを (P) とする. 文型パターンは, 原文についての日本語文型パターン (PJ) と英語文型パターン (PE) の 2 通りで記述されており, 変数を介して字面, 変数, 関数に単語, 句ごとに対応箇所が存在する. 以下に原文 (LJ) と翻訳例 (LW), 原文に対応する PJ と PE の例を示す.

[原文と文型パターンの例 (1)]

- < LP > あの高い枝には手が届かない。
- < LE > I can't reach that high branch.
- < PJ > < $N1$ は > /cf $NP2$ には/tcfk 手が/cf $V3$.*hitei*.
- < PE > $N1$ can't $V3$ $NP2$.

[原文と文型パターンの例 (2)]

- < LP > 急用ができたのでお先に失礼します。
- < LE > An emergency has come up. Please excuse me.
- < PJ > $N1$ が/cf できたので! $VP2$.*teinei*.
- < PE > $N1$ have come up. $VP2$.

2.2 文型パターン辞書と文型パターンパーサ

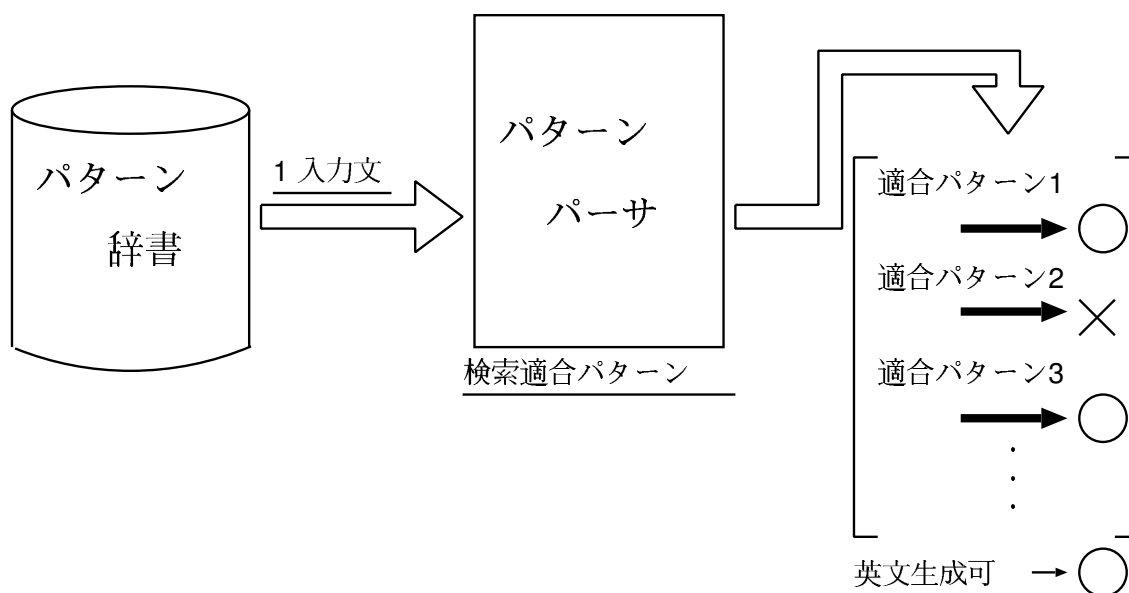
2.2.1 文型パターン辞書

文型パターン辞書には翻訳対象である日英両言語の文型パターンが収録されており、その規模は単語レベル12万パターン、句レベル9.5万パターン、節レベル1.2万パターンである。

2.2.2 文型パターンパーサ

入力文と文型パターンを照合し、入力文に適合する文型パターンを抽出する文型パターンパーサが試作されている。適合するパターンは1入力文につき1パターンではなく、複数適合する場合が多い。適合パターンを用いて日本語入力文から英文を生成したとき、英文として意味の通じないものが生成される場合がある。そのため適合したパターンから英文生成に適したパターンを選択する必要がある。

図 1: 適合パターンの検索



2.3 多変量解析

多変量解析は複雑なデータを解析して有効な情報を見つけるための統計的な手法である [6]. 本実験では入力文に適合した複数のパターンそれぞれに複数のパラメータを付与する, そのためデータ数が多いので, 多変量解析の重回帰分析を用いる. 回帰式は (式 1) のようになる.

$$y = a + \sum_{n=1}^9 b_n x_n$$

回帰式は, 英文評価の評価値 y , 多変量解析によって求められる切片 a と回帰係数 b , パラメータ x_i で構成される. b_i は評価値と評価関数の値 (後述) の残差の 2 乗の総和を最小にする係数である. また, 本実験では 9 つのパラメータを用いるので, $x_1 \sim x_9$ となる.

2.3.1 単語レベルの評価関数

以下に単語レベル実験に用いられた回帰式を示す. パラメータは $x_1 \sim x_7$ の 7 つとなる. 句レベルではこれに名詞句の平均意味属性の逆数と動詞句の平均意味属性距離の逆数の 2 つが加わる.

$$\hat{y} = -0.403 + 0.122x_1 - 0.194x_2 + 0.498x_3 + 0.027x_4 + 0.208x_5 + 0.195x_6 + 0.130x_7 \quad (1)$$

2.4 評価関数

評価値を求める関数であることより、本実験で作成する回帰式のことを評価関数と呼ぶ。評価値 y 、と9つのパラメータ $x_1 \sim x_9$ について説明する。

- 評価値 y

文型パターンパーサによって入力文から適合パターンを得る。適合パターンに対して人手で英文生成を行い英文の品質についての評価を行う。その評価を数値化したものが評価値である。評価値は適合パターン選択実験で、評価関数の値と比較して、適切な適合パターンが選択されているかを確認するために用いる。

- x_1 : パターン適合率

入力文の文字数のうち適合パターンに対応した文字の割合である。ただし単語単位で計算し、パターンに適合している単語と入力文の単語で適合している単語の総文字数の除算で求める。以下の例では、“彼”と“N1”、“頭を打つ”と“VP2”、“喪失した”と“V3.kako”が対応しており、パターン適合率は0.8(11文字/14文字)である。

入力文

彼は頭を打つて記憶を喪失した。

適合パターン

/N1はVP2て/V3.kako。

- x_2 : パターン字面適合率

入力文と、適合パターンに共通する字面の一致する割合。単語単位で計算し、一致単語数と入力文の総単語数の除算で求める。以下の例ではパターン字面適合率は0.3(3単語/10単語)である。

入力文

彼は頭を打つて記憶を喪失した。

適合パターン

N1はVP2て/N3を/V4.kako。

- x_3 :パターン元字面適合率

入力文と、適合パターンを作成する際に用いた原文に共通する字面の一致する割合をパターン元字面適合率とする。単語単位で計算し、一致単語数と入力文の総単語数の除算で求める。以下の例ではパターン字面適合率は0.5(5単語/10単語)である。

入力文

彼は頭を打つて記憶を喪失した。

適合パターンの原文

裁判官は公訴を却下して被告人を放免した。

- x_4 :記号の適合率

適合パターン中の記号が使用される割合で、表1の記号を用いる。以下の例では1.0(2個/2個)である。

入力文

彼は頭を打つて記憶を喪失した。

適合パターン

$\$1^{\{N1\}}$ は}VP2(て/で) $\$1/N3$ を $\$1/V4$.kako。

表 1: 要素記号の一覧

記号名	表記	意味
選択要素記号	(... ...)	いずれかの要素列と適合
任意要素記号	[...]	文型選択上, 任意の要素
補間要素記号	< ... >	ゼロ代名詞等
順序任意要素 指定記号	{... ...}	順序入れ換え可能な範囲 (例 各要素の順序)
位置変更可能 要素指定記号	$\$n^{\{定義\}}$ $\$n$	指定位置に入れ換え可能 (例 副詞の位置)

- x_5 : 変数の適合率

適合パターンに含まれる変数が、入力文との適合に使用される割合である。変数が複数ある場合、異なる品詞ごとにパターンに含まれる数を調べ、入力文で使用される割合の平均を求める。以下の例では $N(2/3)$, $VP(1/1)$, $V(1/1)$ より平均適合率 0.9 である。

入力文

彼 は 頭 を 打って 記憶 を 喪失した。

適合パターン

N1 は / [N3 の] N4 を /V5 て /V6.kako。

- x_6 : 名詞の平均意味属性距離の逆数

入力文と、適合パターンの元となった原文との間で、変数を介して対応する名詞箇所に関して意味属性距離を調べ、平均値の逆数を使用する。平均意味属性距離が 0 の場合は 1 とする。

入力文

彼 は 頭 を 打って 記憶 を 喪失した。

適合パターン

N1 は VP2 て N3 を V4.kako。

適合パターンの原文

裁判官 は 公訴 を 却下して 被告人 を 放免した。

- x_7 : 動詞の平均意味属性距離の逆数

入力文と、適合パターンの元となった原文との間で、変数を介して対応する動詞箇所に関して意味属性距離を調べ、平均値の逆数を使用する。平均意味属性距離が 0 の場合は 1 とする。

入力文

彼 は 頭 を 打って 記憶 を 喪失した。

適合パターン

N1 は VP2 て N3 を V4.kako。

適合パターンの原文

裁判官 は 公訴 を 却下して 被告人 を 放免した。

- x_8 : 名詞句の平均意味属性距離の逆数

入力文と、適合パターンの元となった原文との間で、変数を介して対応する名詞句箇所に関して、意味の中心である名詞の意味属性距離を調べ、平均値の逆数を使用する。意味属性距離が0の場合は1とする。

入力文

いくつかの安全策 を取る ことが一番だ。

適合パターン

$/NP1$ を $/V2^{rentai}$ ことが $/AJV3$ 。

適合パターンの原文

物事の本質 を理解する ことが大切だ。

- x_9 : 動詞句の平均意味属性距離の逆数

入力文と、適合パターンの元となった原文との間で、変数を介して対応する動詞句箇所に関して、意味の中心である動詞の意味属性距離を調べ、平均値の逆数を使用する。以下の例では”頭を打って”と”VP2”が対応しており、”打って”が意味の中心になっている。意味属性距離が0の場合は1とする。

入力文

彼は 頭を打って 記憶を喪失した。

適合パターン

$N1$ は $VP2$ て $N3$ を $V4.kako$ 。

適合パターンの原文

裁判官は 公訴を却下し て被告人を放免した。

3 回帰係数の設定

3.1 回帰係数の作成条件

日英文型パターン辞書から、単語レベルで用いたクローズドテストと同じランダム 200 文 [4] について、句レベルパターンを持つ原文 55 文を入力文として選択し、各文の適合パターンを収集した。ただし適合パターンが入力文から作られたパターンは対象外とする。1 入力文につき最大 10 件の適合パターンを使用することで 492 件を得た。文型パターンパーサは *jpp20040816* を使用する。英文の生成は文型パターンパーサの出力から、入力文に適合した日英の文型パターンを用いて、人手により生成を行う。回帰分析には MicrosoftOffice の回帰分析ツールを使用する。

3.2 評価パラメータ x_i と評価値 y の事例

英文の生成は、適合したパターンに対応する英語パターンを用いて、人手により行う。その英文の品質を評価する。評価は以下の $A \sim D$ の 4 段階で行い、評価関数作成の際には評価に応じた値を使用し、適合パターンの評価値とする。

評価 A : 1

情報や文法に問題がない

評価 B : 0.66

重要でない情報が欠如しているが簡単に修正可能

評価 C : 0.33

入力文を部分的に訳せている

評価 D : 0

入力文の訳としては使用不可能

作成英文の評価をを表 2 に示す.

表 2: 各評価における作成訳の例

評価 A	
入力文 対訳	彼にはその任務を果たせるだけの能力がなかった。 He was not equal to the task.
適 P(日)	N1 には!VP2 ^{rentai} だけの/N3 が /(無かつ なかつ) た。
適 P(英)	N1 had little N3 to VP2.
作成訳	He had little ability to fulfill the duty.

評価 B	
入力文 対訳	この辺には気の利いた料理屋がない。 There are no respectable restaurants in this neighbourhood.
適 P(日)	/ < N1 は > /NP2 が/AJ3。
適 P(英)	N1 AJ3 NP2.
作成訳	No respectable restaurants in this neighbourhood.

評価 C	
入力文 対訳	将来は建築関係に進みたいと思っている。 I'm thinking about going into architecture in the future.
適 P(日)	\$1^{N1 は }!VP2.tai と\$1/思っている。
適 P(英)	N1 hope to VP2.
作成訳	In the future hope to going into architecture.

評価 D	
入力文 対訳	警官が来て騒ぎを鎮めた。 The policemen came and got the things under control.
適 P(日)	N1 が VP2(て で)/N3 を/V4.kako。
適 P(英)	N1 VP2.past across N3.
作成訳	The policemen came across the things.

3.3 回帰係数の作成結果

求めた評価パラメータと評価値から、重回帰分析によって回帰係数を求める。回帰係数から求めた評価関数を式 2 に示す。

$$y = -0.566 + 0.611x_1 - 0.061x_2 + 0.195x_3 - 0.014x_4 \\ + 0.143x_5 + 0.108x_6 + 0.309x_7 - 0.024x_8 + 0.093x_9 \\ \text{(式 2)}$$

式 2 から、評価パラメータ x_1 (パターン適合率) の回帰係数が最も高い値となっていることがわかる。回帰係数の効果については、回帰係数そのものの値の大小では判断できない。パラメータが評価関数に与える影響については考察で述べる寄与率で示すことができるが、回帰係数からパラメータの影響を調べるには、評価値の精度を下げていると思われるパラメータと係数を除いた条件で多変量解析する必要がある。今回の研究では、最も係数の値の小さい x_2 のパターン字面適合率について、多変量解析の際にパラメータを除いて実験を行ったが、評価関数の精度にはほぼ変化がなく、精度を低下させる原因とは言えないことがわかった。

4 評価関数を用いた適合パターン選択実験

4.1 実験目的と調査対象

本実験では適合パターン選択の精度を求め、多変量解析の選択方式の有効性の確認を行う。多変量解析による最適パターン選択の目標は、評価関数による適合パターンの推定の評価値(推定値)に基づき選択した結果、英文として許せる評価の範囲に含まれることである。オープンテストでは、3.1節と同様の方法で新たに集めた35文を対象とする。

4.2 実験手順

各対象文に対して以下を行う。

1. 対象文の適合パターンをパターンパーサで検索
2. 評価関数を適用(推定値を得る)
3. 推定値の最も高い適合パターンを選択
4. 適合パターンから英文を作成
5. 3.2節と同一の基準で英文を評価

4.3 オープンテスト

入力文35文から前節の手順1により341件の適合パターンを得た。手順2で各件の適合パターンの推定値を算出する。手順3で各入力文における最大10件の適合パターンの中で、最も推定値の高いパターンを選択する。手順4では手順3で求めた適合パターンを用いて英文を作成する。最後に手順5で英文を4段階に評価する。

4.4 実験結果

各入力文で推定値上位1位の英文が評価AまたはBの場合、適合パターンの選択に成功とし、その適合パターンを正解適合パターンとする。評価CまたはDであるものを不正解適合パターンとする。推定値を0.1の階級区間に分け、各区間において推定値の1位を持つ入力文が、正解適合パターンを持つ割合を調べた。55文のクローズドテスト結果を図1に、35文のオープンテスト結果を図2に示す。実線が正解適合パターン正解率を表し、点線は評価の条件を変えた場合の正解率を表す。図より、推定値が0.7以上のときは正解適合パターン選択の正解率が50%以上で行えると推測できる。また推定値が0.4未満のときは正解適合パターン選択の正解率は0%である。

図 2: 推定値1位の正解率 (クローズドテスト)

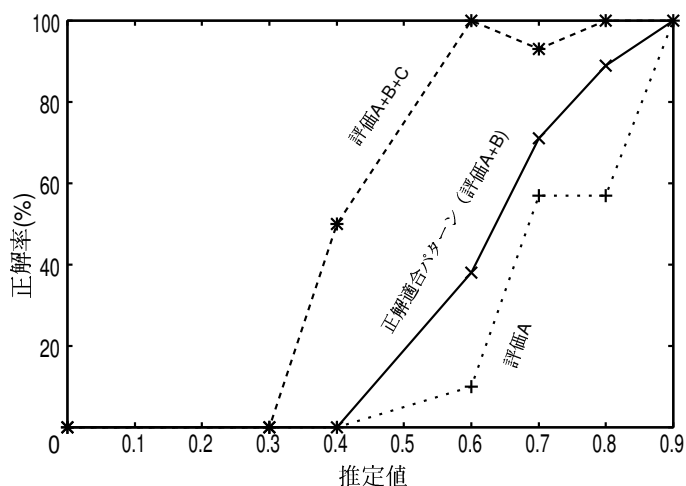
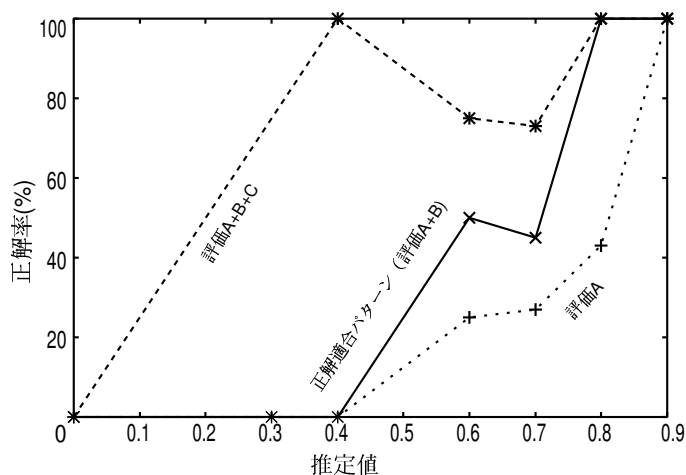


図 3: 推定値1位の正解率 (オープンテスト)



推定値ごとの入力文数を調査した。推定値が0.75以上のときは正解率が高いことから、正解適合パターンを持つ入力文も、推定値0.7~0.8の間で特に多い結果となった。

図 4: 推定値 1 位の正解率 (クローズドテスト)

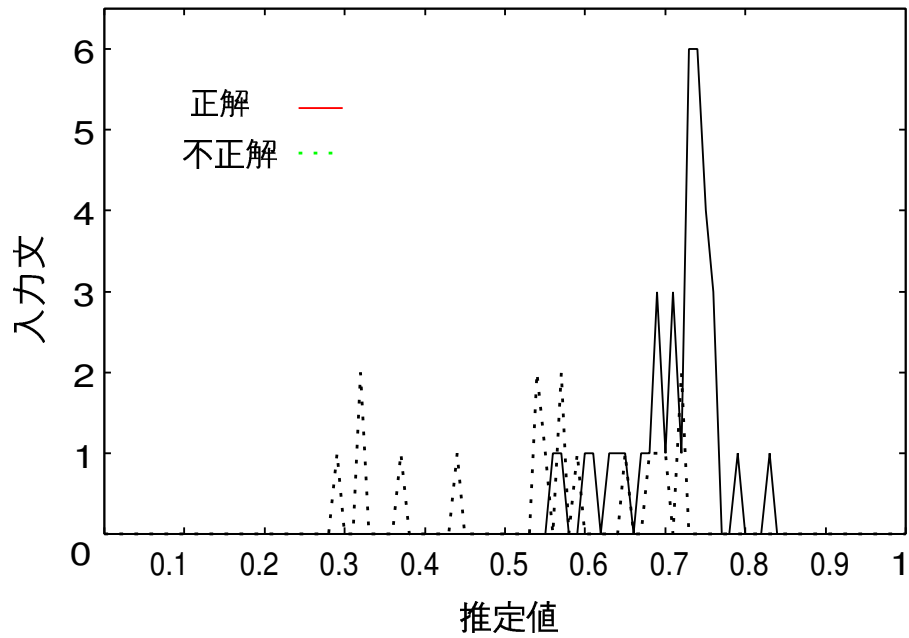
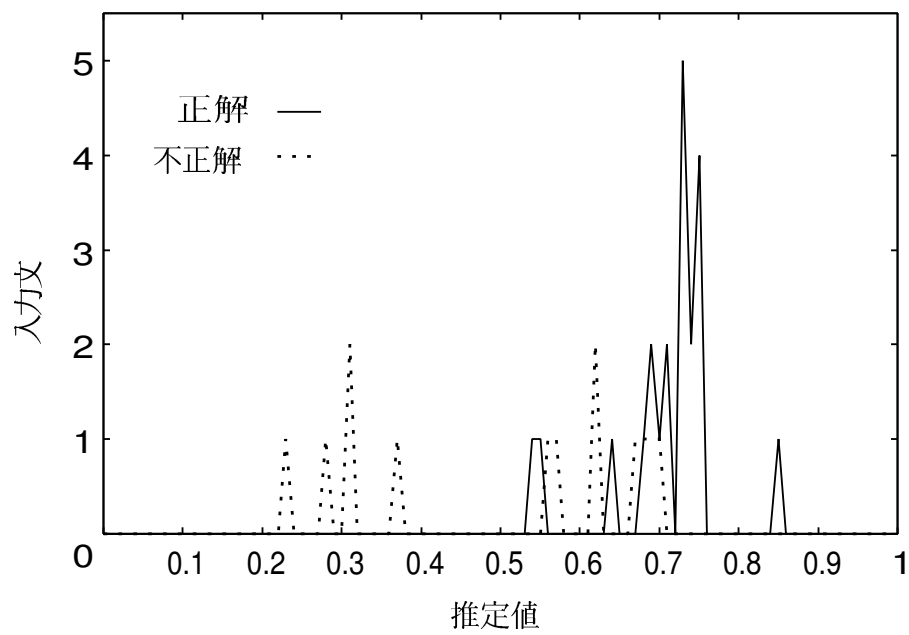


図 5: 推定値 1 位の正解率 (オープンテスト)



5 考察

5.1 推定値の信頼性

入力文 35 文のうち, 評価 A を最低 1 つ持つ文は 22 文 (212 件) あった. その 22 文に対して推定値が上位 1 位の適合パターンの評価を調べてみたところ, 11 文 (50%) において評価 A の適合パターンを選択した. また 18 文 (82%) において評価 A または B を選択した (表 3).

表 3: 信頼性

	評価 A	評価 A または B
1 位	50%(11/22 文)	82%(18/22 文)
1 位 ~ 2 位	68%(15/22 文)	95%(21/22 文)
1 位 ~ 3 位	73%(16/22 文)	100%(22/22 文)
1 位 ~ 4 位	100%(22/22 文)	100%(22/22 文)

5.2 推定値による正解パターン選択失敗の原因

実験結果より, 各入力文に対して推定値が上位 1 位の適合パターンの評価を調べてみた所, 不正解適合パターンが上位 1 位となる入力文は 11 文であった. 推定値の上位 1 位が評価 C, D である入力文について調査した所, 最大 10 件全てのパターンが評価 C または D であった入力文が 6 文あった.

不正解適合パターンのみとなる入力文は, 入力文とパターンとのパターン適合率がどれも 50 % 以下で, 入力文の一部しか適合しないため, 正解パターンの選択に失敗したと推測できる.

5.3 単語レベルとの比較

過去に行われた単語レベルの実験結果との比較を行う。単語レベルでの実験に使用されたクローズドテスト, オープンテストでの入力文の比較を表4に示す。調査対象入力文は適合パターンを持つ入力文の数である。各入力文ごとに30パターンまで調査に使用したため, 総適合パターン数と調査パターン数は異なる。

表 4: 入力文の比較

	クローズド	オープン
入力文数	200 文	200 文
調査対象入力文	72 文	82 文
総適合パターン数	1847 パターン	1716 パターン
平均適合パターン数	26 パターン	21 パターン
調査パターン数	765 パターン	680 パターン
正解適合パターンを持つ入力文の数	29 文	29 文
正解パターンの割合	13%(96/765)	14%(95/680)

5.1 節と同じ実験が単語レベルにおいても行われており, 結果が表5のようになる [4].

表 5: 単語レベル実験結果

	評価 A
1 位	83%(24/29 文)
1 位 ~ 2 位	90%(26/29 文)
1 位 ~ 3 位	100%(29/29 文)

結果を比較すると単語レベル, 句レベルともに推定値の上位3位までで正解適合パターンを選択できている。調査対象文にも差は少ないため, 多変量解析を用いた選択手法において, 単語レベル句レベルでほぼ同等の結果を得たと言える。

5.4 評価パラメータの考察

5.4.1 動詞意味属性距離について

推定値の上位1位が不正解適合パターンを選択したパターンについて、各評価パラメータを調査したところ、動詞意味属性距離の値が近いという共通の特徴があった。動詞の意味属性距離が近いことで推定値の増加が起こり、不正解適合パターンが推定値の上位になった。

5.4.2 寄与率

評価関数の寄与率を求めた。寄与率によってパラメータの重要性がある程度得られる。作成した評価関数に加え、各評価パラメータ単独で回帰分析し、寄与率を求めた結果を表6に示す。

表 6: 寄与率の比較

本稿作成評価関数	15.7%
パターン適合率のみ	8.5%
パターン字面適合率のみ	0.2%
パターン元字面適合率のみ	1.1%
変数の適合率のみ	0.8%
記号の適合率のみ	0.7%
名詞意味属性のみ	1.2%
動詞意味属性のみ	5.5%
名詞句意味属性のみ	0.9%
動詞句意味属性のみ	0.5%

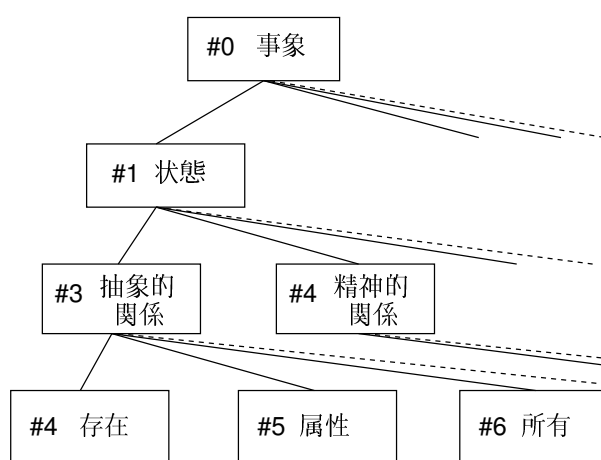
表5より、パターン適合率(8.5%)と動詞の意味属性距離(5.5%)について寄与率が高い。評価関数に与える動詞意味属性距離の影響が大きいことが、不正解適合パターンが推定値の上位となった原因と考えられる。

5.4.3 意味属性距離の問題点

意味属性距離は日本語語彙大系 [7] に記載されている「一般名詞意味属性大系」と「用言意味属性大系」を用いる。一般名詞意味属性大系は約 40 万語の名詞を、最大 12 段の木構造で 2710 の意味属性大系に分類している。また用言意味属性大系は約 6000 語の用言を、最大 4 段の木構造で 36 の意味属性に分類している。一般名詞意味属性大系、用言意味属性大系ともに、木構造を基本構造とし、各ノードにおいて上位の意味属性の性質を下位の意味属性が継承する。

本研究では意味属性距離の逆数をパラメータとして用いる。距離は単純に各ノード間を 1 として計算する。以下の図 6 のような場合について考察する。

図 6: 意味属性距離



1 つ目の例として図中の #4 から #0 の距離 3, #4 から #5 の距離 2 を選ぶと距離の差は 1 である。また 2 つ目として #4 から #1 の距離 2, #4 から #0 の距離 3 を選ぶと、これも距離の差は 1 となる。本研究では 2 つの例を等しい距離の差として計算に用いる。

1 つ目の例では、あるノードから出る枝は右と左の両方を通っている。一方 2 つ目の例では全て左の枝を選んでいる。同方向に枝が伸びている「状態」と「存在」の距離が 2 となり、左、右と枝が伸びている「存在」と「所有」も同じく距離が 2 となるが、意味として「存在」と「所有」の方が遠い距離であるように感じる。この例のように、あるノードから一方方向に伸びる枝と、左右に広がっていく枝がある場合、距離の間隔に差をつけることで、より効果的なパラメータとして作用する可能性が考えられる。枝の伸びる方向によって、条件を付ける等の調整が必要であるかどうかの検討が今後の課題である。

5.5 推定精度

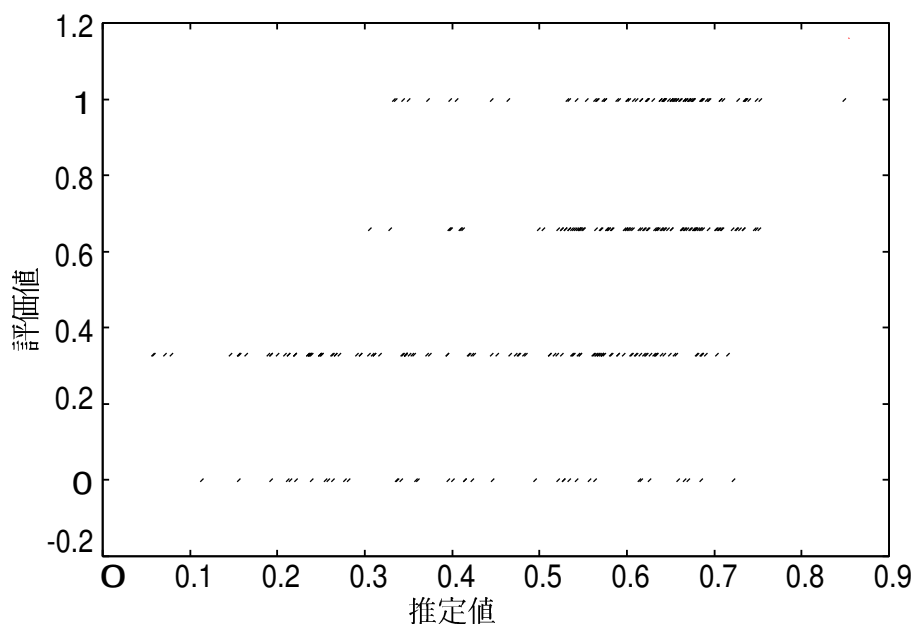
評価値が離散的であるため単純な相関係数が使えないと考えて、推定値と評価値の差が0.33未満(評価の差が1つ未満)ならば「合」とした。推定精度は77%(合の数/推定数 = 27/35)であった。また、各文で上位10位までの推定値について推定精度を求めたところ、79%(270/341)であった。推定値をもとに適合パターンを選択し、そこから英文を生成すると、その推定値の品質で英文の得られる可能性が79%であることを意味する。

表 7: 実験結果

値の差	推定値 1 位	推定値上位 10 位
0.33 未満 (合)	77%(27/35 文)	79%(270/341 件)
0.66 未満	20%(7/35 文)	18%(63/341 件)
0.66 以上	3%(1/35 文)	2%(8/341 件)

オープンテスト 341 件の推定値が、A ~ D の 4 段階でどのように分布しているかを調べた。結果を図 7 に示す。分布の様子から推定値が 0.5 付近より大きい場合の評価が A にかたまっていることがわかる。

図 7: 推定値の分布



6 おわりに

本研究では句レベルについて多変量解析による選択方式を用いることで、最適な文型パターンを選択した。実験の結果、推定値が0.7以上のとき正解適合パターンを選択する正解率が50%以上となった。また、推定値が0.4未満のときは正解適合パターン選択の正解率は0%である。今後は入力文と適合パターン数の収集量の増加による選択精度の調査と精度の向上が課題である。また節レベルについて多変量解析による選択方式が有効であるかを適用する必要がある。

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」の支援により行った。

本論文作成に際して、多大なる検討と助言をしてくださった池原悟教授ならびに村上仁一助教授、徳久雅人助手そして計算機工学C研究室の方々に厚く御礼申し上げます。また、参考にさせて頂いた論文、本の著者の方々に御礼申し上げます。

参考文献

- [1] 池原悟: 等価的類推思考の原理による機械翻訳方式, 信学技報, TL2002-34, pp.7-12, 2002.
- [2] 徳久雅人, 池原悟, 村上仁一: 文型パターンパーサの試作, 言語処理学会年次大会発表論文集, pp.608-611, 2004.
- [3] 池原悟, 徳久雅人, 竹内(村本)奈央, 村上仁一: 日本語重文・複文を対象とした文法レベル文型パターンの被覆率特性, 自然言語処理, Vol.11, No.4, pp.147-178, 2004.
- [4] 岡田敏, 村上仁一, 徳久雅人, 池原悟: 多変量解析による最適文型パターンの選択方式, 言語処理学会年次大会発表論文集, pp.25-28, 2005.
- [5] 池原悟ほか: 機械翻訳のための日英文型パターン記述言語, 信学技報, TL2002-48/NLC2002-90, pp.1-6, 2003
- [6] 上田太一郎, 苅田正雄, 本田和恵; 実践ワークショップ Excel 徹底活用 多変量解析, 秀和システム, 2003.
- [7] 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.

付録

1 使用した入力文

2 オープンテスト用データ

入力文 110 文と入力文の模範訳

表は以下の通りに構成されている。

入力文 100 文と入力文に対応する模範訳 300 文 (1 入力文に対し 3 模範訳) を記載している。

適合文型パターンの評価

表は以下の通りに構成されている。