概要

近年、機械翻訳方式として構文的・意味的に類似した文を翻訳に使用する用例翻訳が注目されている。この用例翻訳では、検索入力として与えられた文に対し、表現のもっとも類似したパターンをデータベースより検索する。そして類似パターンの対訳データを使用し、検索入力として与えられた文の翻訳を行う。よって翻訳品質を向上させるには膨大な対訳データベースが必要となってくる。現在、実用可能な精度をもつ用例翻訳を実現させるために日英対訳パターンをどれだけ収集すれば良いか調査はまだされていない。

そこで本研究では、大量の対訳データ中における日本文の類似性を調査する。

重文・複文の日本語パターンに着目し、適切な単語の置換を行うことで効率よく類似パターンを検索する。そして被覆率や包含関係の調査および評価を行う。

具体的な調査方法として、まず CREST の重文複文の例文 91680 文に対して形態素解析を行った。この形態素解析結果を用いて種々の置換を行いデータベースを作成した。そして各置換条件下のパターンの被覆率調査を行った。また、例文からランダムに 100 文選び、データベース中に包含関係を持つパターンの数を調査した。最後に包含関係にあるパターンの対訳英文の文型構造を人手で比較し評価した。

本手法により日本語の重文複文において、単語の置換による類似パターン数の変化と置換がパターンに与える影響が調査できた。

目 次

図 目 次

表 目 次

1　はじめに

従来の機械翻訳方式は、要素合成的手法と要素分解的手法の 2 通りの手法がある。要素合成的手法とは、統語構造の持つ意味を考慮にいれず、 部分の意味から全体の意味を合成しようとする手法である。要素分解的手法とは、 表現を細分化して分析すれば全体が分かるとする手法である。これらの手法は翻訳の処理能力が安定しており、多く研究者が研究している。しかし、表現の持つ意味の欠落を防げず、訳文の品質向上が難しくなってきている。

よって近年、自然言語処理の分野では柔軟な翻訳を実現するため用例翻訳が注目されている。用例翻訳の具体的な翻訳方法は、まず日英双方のパターンを翻訳対として設定する。そして、入力と一致するパターンを翻訳対のリストの中から検索し、一致したパターンの対を使用し翻訳する方法である。

この用例翻訳の翻訳品質を向上させるためには、膨大な日英対訳パターンが必要になるという問題がある。さらに、実用可能な精度を持つ用例翻訳を実現させるために必要な日英対訳パターンがどれだけ必要なのか調査はまだされていない。

そこで本研究では大規模対訳コーパスにおける日本文の構造的類似性を調査する。調査対象に CREST 重文・複文の例文の日本語文データを使用する。まずデータを形態素解析し、その結果をもとに種々の置換を行う。そして被覆率と包含関係にあるパターンの調査を行う。調査の結果、置換を行うと被覆率は 1.86%から 18.46%まで向上した。また、データベースの中から入力文としてランダムに 100 文を選ぶ。入力文の 99%にデータベース中に包含関係にある類似パターンが存在した。

以下、２章では類似パターンの調査方法、３章では被覆率の調査について、４章では包含関係にある類似パターンの調査について、そして５章で結論と今後の課題を述べる。

2　単語の並びから見た類似性

2.1　調査方法

無限に存在する表現を用例翻訳に使用するためには置換によるパターン化とその圧縮を行う必要がある。

これには、言語学的な知識が必要であり、 言語現象の体系化を行う必要がある。パターン化において 1 つのデータの大きさを（1）単語、（2）句又は節、（3）文とする方法がある。

理想的には (1),(2),(3) 全てのレベルで対応関係をつけ、用例翻訳データベースに格納することが望ましい。しかし、 マルチレベルで対応付けをすればする程翻訳処理の幅は広がるが、データベースの作成は難しくなる。

本研究では用例の追加が簡単にできる文単位で日本語パターンの構造的類似性を調査する。次に単語の置換によってパターンが一致する例を示す。

例 1：彼 は 自室 で ない と 勉強でき ない
例 2：私 は 教室 で ない と 勉強でき ない
例 3：N は N で ない と 勉強でき ない

例 1、例 2 を比較すると、名詞 (彼, 私)、(自室, 教室) が異なる。しかし名詞を N と置換すると例 3 になり同じパターンになる。本研究では単語の置換によってパターン構造が一致するパターンを構造的類似パターンとする。

そして本研究では大規模の重文・複文の日本語文において適切な置換を行い、文単位のパターン分布を調査する。

2.2 置換ルール

本論文では調査対象に CREST 重文複文の例文 91680 文を使用する。まず形態素解析プログラム altjaws を用いて形態素解析を行う。その後、表 1 の (1)〜(5) の場合で単語を品詞に置換してデータベースを作成する。

表 1: 置換ルール

(1) 単語区切り
(2) 名詞を *N* と置換
(3)(2) に表 2 の処理を行う
(4) 名詞を *N*、動詞を *V*(補助動詞は除く) と置換
(5)(3) に表 2 の処理を行う

表 2: 置換テーブル

[名詞] の [名詞] の... → [名詞] へと置換。
[名詞] の連続→ 1 個の [名詞] へと置換。
連体詞 (この、その、あの) の削除。
名詞に係る形容詞、形容動詞の削除。

各置換方法によるパターンの変化を表 3 に示す。

表 3: 各置換方法によるパターンの変化

置換方法 (1) この/本/は/値段/が/安い/から/学生/の/手/に/入る
置換方法 (2) この/ *N* /は/ *N*/が/安い/から/ *N*/の/ *N*/に/入る
置換方法 (3) *N*/は/ *N*/が/安い/から/ *N*/に/入る
置換方法 (4) この/ *N*/は/ *N*/が/安い/から/ *N*/の/ *N*/に/*V*
置換方法 (5) *N*/は/ *N*/が/安い/から/ *N*/に/*V*

3　被覆率の調査

3.1　被覆率の定義

2.2 章のルールによって作成したデータベースに対し、同一のパターンを 1 つのパターンとして数える。被覆率、エントロピーを以下に定義する。

$$\text{被覆率}\,(\%) = \frac{\text{重複したパターンの数}}{91680(\text{総例文数})} \times 100$$

重複したパターン数とは、各データベースから得たパターンの種類を総パターン数とし、総例文数と総パターン数との差である。

$$\text{エントロピー} = \sum_{k=1}^{n} P(a_k) \log_2 P(a_k) \qquad P(a_k) : \text{パターンの生起確率}$$

3.2 置換による被覆率

2.2 章で述べた置換条件 (1)〜(5) で被覆度・総パターン数・エントロピーの調査結果を表 4 に示す。

表 4: 各置換条件における被覆率の推移

置換条件	(1)	(2)	(3)	(4)	(5)
被覆率	1.86%	3.51%	5.80%	11.60%	18.46%
総パターン数	89971	88464	86365	81049	74760
エントロピー	16.45	16.40	16.29	16.00	15.71
複数回出現パターン数	1522	2517	3131	4429	5413

表4 より、置換を行うことにより単語で被覆率を計算した場合より被覆率が 16.6% 向上した。また、総パターン数は 15211 個減少し、エントロピーも 0.74 減少した。複数回出現したパターンは 3891 種類増え、本研究の置換方法が被覆率の向上に有効であると示された。

しかし、本研究では構造的類似パターンのみに着目した為パターンの意味的類似性を考慮していない。よって動詞を置換する場合、日本文の意味がかなり欠如してしまう。よってパターンは同じでも意味的には異なる場合があった。例を表 5 に示す。

表 5: 元の日本語文の違い

置換文	N/は/N/に/V/て/V/た/
入力文 対訳	私は川に沿って歩いた I walked along by the river.
データベース文 対訳	彼は事実に基づいて立案した He built on facts.

3.3 出現回数の多かったパターン

各置換条件の下被覆率の計算をした中で、出現回数の多かったパターン上位5パターンを表に示す。なお、[ ] で囲まれている部分は置換によって削除した部分である。(出典については付録の出典番号対応表を参照)

表 6: 置換条件 (1) における出現回数の多かったパターン

回数	パターン	出典番号	例
6	来年/の/こと/を/言う/と/鬼/が/笑う/		来年のことを言うと鬼が笑う
		AM007601	Next year is the devil's joke.
		AM054807	Don't count your chickens before they are hatched.
		AC044411	Talk of the next year and the oni is sure to laugh.
6	急が/ば/回れ/		急がば回れ
		AD008117	More haste,less speed.
		AG014211	Haste makes waste.
		AM003290	The more haste,the less speed.
5	色/の/白い/の/は/七/難隠/す/		色の白いのは七難隠す
		AN006946	A white complexion covers a multitude of sins.
5	所/変われ/ば/品/変わる/		所変われば品変わる
		AC029473	Every country has its own customs.
		AF016977	Customs differ from country to country.
		AM035295	So many countries,so many customs.
		AG003654	Different places have different things.
4	雷/が/鳴る/と/牛乳/が/腐る/		雷が鳴ると牛乳が腐る
		AN018352	Thunder turns milk sour.

置換を行わず単語区切りで検索したところ、同じパターンが同じ出典元から出現する場合が多い傾向にあった。

表 7: 置換条件 (2) における出現回数の多かったパターン

回数	パターン	出典番号	例
30	*N*/は/*N*/の/ある/*N*/だ/	AN095741	あれは来歴のある学校だ The school has a history.
		AD021822	彼は分別のある人だ He is a thinking man.
27	*N*/の/好い/*N*/だ/	AN094589	容貌の好い婦人だ She has personal charms.
		AN012629	思い切りの好い男だ He is a man of decisive character.
24	*N*/の/無い/*N*/だ/	AN098712	悪気の無い男だ He is a harmless fellow.
		AN004521	威厳の無い風だ He has an undignified manner.
16	*N*/は/*N*/の/ない/*N*/だ/	AM051917	彼は申し分のない学生だ He is an exemplary student.
		AC015362	あいつは根性のない男だ That fellow has no guts.
12	*N*/の/悪い/*N*/だ/	AN094590	容貌の悪い人だ He is an ugly man.
		AN072092	寝つきの悪い子だ The child will not go to sleep.

出現回数の多いパターンの中には、動詞がひとつしかないパターンが多く存在した。

表 8: 置換条件 (3) における出現回数の多かったパターン

回数	パターン	出典番号	例
519	*N*/は/*N*/だ/	AN096243	君は [立派な] 体格だ
			You have a fine physique.
		AN096247	彼は [立派な] 身代だ
			He has a large fortune.
126	*N*/だ/	AN096794	[輪廓の正しい] 顔だ
			He has regular features.
		AN071536	[人数の多い] 家族だ
			They are a large family.
62	*N*/は/*N*/を/し/ている/	AN097504	彼は [久しく] 浪人をしている
			He has been long out of employment.
		AN017371	[あの] 男は [妙な] 格好をしている
			He cuts a queer figure.
58	*N*/は/*N*/だっ/た/	AN011674	昨夜は [恐ろしい] 地震だった
			There was a dreadful earthquake last night.
		AS035041	それは [残念な] 出来事だった
			It was a regrettable happening.
56	*N*/は/*N*/*N*/だ/	AN089159	これは近年 [珍しい] 名作だ
			It is as fine a piece of work as has not been seen for many a year.
		AG000421	これはいちばん [新しい] ニュースだ
			This is the latest news.

名詞に係る形容詞・形容動詞の削除により重文複文パターンと判断しかねるパターンが存在した。

表 9: 置換条件 (4) における出現回数の多かったパターン

回数	パターン	出典番号	例
299	*N*/を/*V*/て/*V*/	AN055800	石橋を叩いて渡る to be over-prudent
		AN098489	相好を崩して笑う to grin broadly
177	*N*/を/*V*/て/*N*/を/*V*/	AN079043	両手を拡げて人を迎える to welcome one with open arms
		AN097744	賄賂を使って人を買収する to a buy off a man
167	*N*/は/*N*/を/*V*/て/*V*/た/	AN094599	彼は用務を帯びて洋行した He has gone abroad on business.
		AN049646	彼は世界を一周して来た He has been round the world.
151	*N*/は/*N*/を/*V*/て/*N*/を/*V*/た/	AN046047	彼女は寝食を忘れて病人を看護した She nursed the invalid, forgetful of everything.
		AN064905	彼は財産を使い果たして社会党に投じた He went through his fortune, and joined the socialists.
118	*N*/に/*V*/て/*V*/	AN095601	酒に酔ってよろめく A drunkard reels.
		AN097719	輪になって坐る to sit in a circle

名詞・動詞を置換すると、各パターン出現回数が大幅に増加した。しかし、パターンから元の日本文を導き出すのはかなり難しい。

表 10: 置換条件 (5) における出現回数の多かったパターン

回数	パターン	出典番号	例
519	*N*/は/*N*/だ/	AN096243	君は [立派な] 体格だ
			You have a fine physique.
		AN096247	彼は [立派な] 身代だ
			He has a large fortune.
350	*N*/を/*V*/て/*V*/	AN055800	石橋を叩いて渡る
			to be over-prudent
		AN098489	相好を崩して笑う
			to grin broadly
289	*N*/は/*N*/を/*V*/て/*N*/を/*V*/た/	AN046047	彼女は寝食を忘れて病人を看護した
			7 She nursed the invalid, forgetful of everything.
		AN064905	彼は財産を使い果たして社会党に投じた
			He went through his fortune, and joined the socialists.
278	*N*/は/*N*/を/*V*/て/*V*/た/	AN094599	彼は用務を帯びて洋行した
			He has gone abroad on business.
		AN049646	彼は世界を一周して来た
			He has been round the world.
271	*N*/を/*V*/て/*N*/を/*V*/	AN079043	両手を拡げて人を迎える
			to welcome one with open arms
		AN097744	賄賂を使って人を買収する
			to a buy off a man

出現回数が多いパターンには *N, V,* 助詞のみである場合が多かった。

4　包含関係にあるパターンの類似性

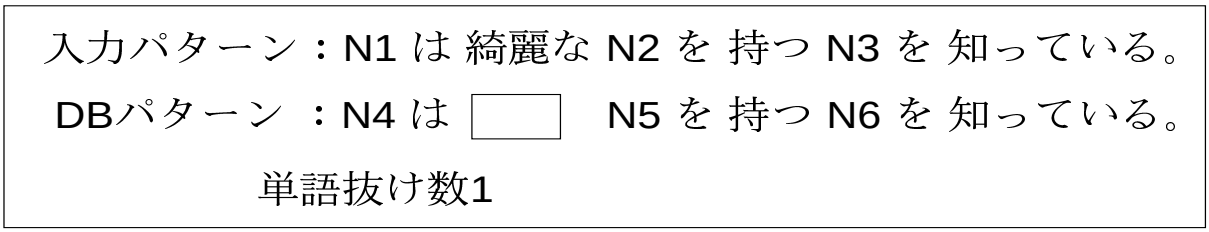
4.1　包含関係にあるパターン

被覆率の章でパターンの構造的な一致による被覆率について調査したが、置換後でも 18.46%と低い結果だった。そこでさらなる類似パターンの調査のため、新しく類似パターンの定義をする。

実際の翻訳現場で、ある入力パターンが与えられたときデータベース中に入力パターンが包含するパターンが存在すれば、データベース中のパターンの対訳が入力パターンの翻訳に有用である場合がある。この例を図 1 に示す。

よって本研究では包含関係にあるパターンも類似なパターンであると定義する。本節では各置換条件下で包含関係にあるパターンがデータベース中にどれだけ存在するのか調査を行い、類似度を削除レベルで評価する。また、各削除レベルの評価に対訳英文の文型構造の比較を用いる。

図 1: 包含関係にあるパターンの例



4.2　削除レベルの定義

入力パターンとデータベース中のパターンを比較し、データベース中のパターンが N 単語抜けで包含関係にあるペアを削除レベル N のペアと定義する。N の数値で包含関係にあるパターンの類似度を評価する。

4.3 包含関係の抽出法

包含関係を調査するためには、順序完全一致評価法を使用する必要がある。順序完全一致評価法とは、順序が完全に一致した単語のみを単語一致数として評価する方法である。すなわち、順序関係の合わない単語は無視して単語一致数に加えない。しかし, ここで順序関係が正しい単語はどれかという問題が生じる。特に単語一致数が多く、それらの順序関係が複雑に入れ替わっている場合、順序の完全一致した単語のすべての組み合わせを考慮する必要があり、その判定が複雑になる。例えば次のような例を考える。

A = 私 の 綺麗な 車 を 父 が 運転する
B = 父 が 私 の 綺麗な 車 を 運転する

ふたつのパターン A,B の間には (私, の, 綺麗な, 車, を, 父, が, 運転する) の単語が一致している。順序の完全に一致している組み合わせは (私 の 綺麗な 車 を 運転する) と (父が 運転する) のふたつである。

順序完全一致評価法ではこのふたつの組み合わせのうち一方を選んで類似度として評価をし、一方を無視する。通常、評価する一致単語は多いほどよいので、この場合は 6 単語である (私 の綺麗な 車 を 運転する) の組み合わせを選択するのが妥当である。したがって単語一致数は 6 となる。

このように単語の順序が完全に保たれており、かつ単語一致数が最大となる組み合わせを発見するアルゴリズムとして DP マッチングがある。

4.4 DP マッチング法

今、二つの単語列 A、B をそれぞれ以下に定義する。

$$A = a_1 a_2 a_3 \cdots a_{l_A}, \qquad B = b_1 b_2 b_3 \cdots b_{l_B} \quad (a_i, b_j \text{は}1\text{つの単語を表す})$$

単語列 A,B 間の最大文字一致数 *onm(A,B)* を DP マッチング法では以下に定義する。

$$onm(A,B) = \begin{cases} 0 & l_A = 0 \vee l_B = 0 \\ P_{l_A, l_B} & otherwise \end{cases}$$

$$P_{i,j} = \begin{cases} 0 & i = 0 \vee j = 0 \\ max(P_{i-1,j-1} + m_{i,j}, P_{i-1,j}, P_{i,j-1}) \end{cases}$$

$$m_{i,j} = \begin{cases} 1 & a_i = b_j \\ 0 & otherwise \end{cases}$$

次ページに 4.3 節のパターン A、B の照合を DP マッチング法で行った計算過程の例を示す。最大文字一致数 *onm(A,B)=6* が正しく計算できることがわかる。

表 11: DP マッチング法による計算過程 m

	私	の	綺麗な	車	を	父	が	運転する
父	0	0	0	0	0	1	0	0
が	0	0	0	0	0	0	1	0
私	1	0	0	0	0	0	0	0
の	0	1	0	0	0	0	0	0
綺麗な	0	0	1	0	0	0	0	0
車	0	0	0	1	0	0	0	0
を	0	0	0	0	1	0	0	0
運転する	0	0	0	0	0	0	0	1

表 12: DP マッチング法による計算過程 P

	私	の	綺麗な	車	を	父	が	運転する
父	0	0	0	0	0	1	1	1
が	0	0	0	0	0	1	2	2
私	1	1	1	1	1	1	2	2
の	1	2	2	2	2	2	2	2
綺麗な	1	2	3	3	3	3	3	3
車	1	2	3	4	4	4	4	4
を	1	2	3	4	5	5	5	5
運転する	1	2	3	4	5	5	5	6

4.5 DP マッチングツールの特性

今回 DP マッチング法を用いたプログラムに HTK[3] の HResult を使用する。このプログラムは調査対象の 2 つのパターンを比較し、削除エラー・置換エラー・挿入エラーの 3 種類のエラーを出力する。

削除エラーは入力パターンに対しデータベース中のパターンが N 単語抜けているとき出力される。挿入エラーはデータベース中のパターンに対し入力パターンが N 単語抜けているとき出力される。置換エラーは入力パターン、データベース中のパターンともに同位置の単語を置換すれば等しくなるとき出力される。

例として 4.3 節の単語列 A,B で各エラーの数値を計算する。

```
A =       私 の 綺麗な 車 を 父 が 運転する
B = 父 が 私 の 綺麗な 車 を       運転する
```

単語列 A を入力パターンとし、パターン B をデータベース中のパターンとする。パターン A,B は単語一致数６、削除エラー２、挿入エラー２、置換エラー０の関係となる。

本研究では入力パターンがデータベース中のパターンを包含している場合にのみ着目する。つまり削除エラーのみがエラーとして出力する時、調査対象のパターンを類似パターンとする。よって削除エラー数 N の関係が削除レベル N の関係となる。

4.6　包含関係にある類似パターン検索とその評価

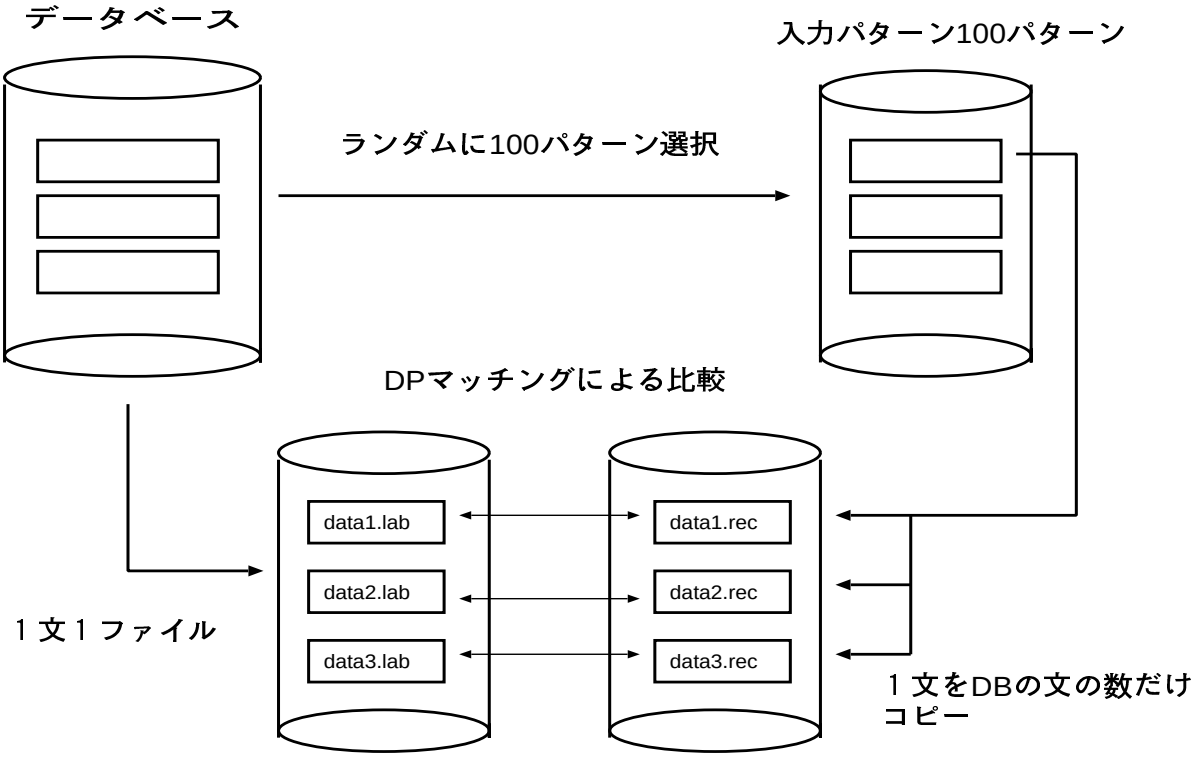
4.6.1　削除レベルの調査方法

例文からランダムに 100 文選び、置換条件 (1)〜(5) において 100 パターンのうちデータベース中に削除レベル N の関係を持つパターンの数を調査する。包含関係にあるパターンを調査するために図 2 のシステムを作成する。

包含関係にあるパターンの調査を以下の手順で行う。

1. 例文から 100 文ランダムに選択する
2. 選択した文、例文ともに形態素解析を行う
3. 形態素解析結果をもとに、選択した文と例文を各置換条件で置換し DB を作成する
4. データベース中ののパターンをそれぞれ 1 つのファイルにする
5. 100 パターンから１つパターン取りだし、データベースのパターンの数だけコピーする
6. DP マッチングで１ファイル毎に比較を行う
7. 5,6 を入力パターンの数 (100 回) くり返す

図 2: 包含関係検索システム



4.6.2　包含関係にあるパターンの評価

包含関係にあるパターンの評価として、パターンの元となる日本文の対訳英文に着目する。対訳英文の文型構造が同じであれば類似パターンであると人手で判断する。表 13 の 5 個の文型で調査を行う。

表 13: 英語の文型 5 種

第 1 文型（S＋V）
第 2 文型（S＋V＋C）
第 3 文型（S＋V＋O）
第 4 文型（S＋V＋O＋O）
第 5 文型（S＋V＋O＋C）

4.7　各削除レベルごとの調査結果

ランダムに選んだ 100 文に対し、置換条件 (1)〜(5) でデータベース内に削除レベル 5 までの関係を持つパターンの数を調査した。結果を表 14 に示す。

また、パターン元の日本文の対訳英語文の文型構造が同じであれば類似パターンだと判断する評価を人手で行なった (4.6.2 章)。この結果を表 14 の英文型一致数に示す。

ここで包含パターン保持数とは、ランダムに選んだ 100 文のうちデータベースの中に包含関係を持つ文の数を示している。平均削除エラー数とは 100 文において包含関係をもつパターンの最小削除エラー数の平均である。

表 14: 削除レベル 5 までの関係を持つパターンの数と評価

置換条件	(1)	(2)	(3)	(4)	(5)
削除レベル 1	1	1	3	20	33
英文型一致数	1	1	2	11	21
削除レベル 2	1	2	12	32	42
英文型一致数	1	1	10	19	31
削除レベル 3	0	1	10	36	56
英文型一致数	0	1	4	24	39
削除レベル 4	0	1	12	44	66
英文型一致数	0	1	7	26	48
削除レベル 5	0	1	11	53	73
英文型一致数	0	0	6	31	50
包含パターン保持数	2	6	43	85	99
平均削除エラー数	1.50	4.33	5.59	4.54	3.45

4.8　包含関係にあるパターン調査結果 (置換方法 3)

置換方法 3 における包含関係にあるパターンを各削除レベル毎に例を示す (出典番号は付録を参照)。削除レベルが低いパターンの類似度は高かった。

表 15: 包含関係にあるパターン調査結果 (置換方法 3)

削除 LV	出典番号	パターン	例
1	AN032718	あまり/好まし/N/だ/	あまり好ましくない地位だ It is not a very desirable position.
	AF002791	あまり/N/だ/	あまり [うまみの無い] 商売だ This is a business that doesn't promise very much profit.
	AC023088	N/は/N/ばかり/だ/	俗世間は [煩わし] いことばかりだ Life is full of annoyances.
	AD001156	N/は/N/だ/	彼は [我々のチームの貴重な] メンバーだ He is a valuable asset to our team.
2	AD007278	N/は/N/に/対し/て/N/を/とっ/た/	彼は敵に対して寛大な処置をとった He showed generosity toward his enemy.
	AD001703	N/は/N/に/N/を/とっ/た/	彼は彼女に親切な行動をとった He behaved kindly to[toward] her.
	AC010400	N/は/N/で/N/だ/と/思う	ぼくは自分で [負けず嫌いな] 気性だと思う I would say I've got an unyielding[open] disposition.
	AD008928	N/は/N/で/N/だ/	彼は熱病で重態だ He is very ill with a fever.
3	AG007884	N/の/言う/N/に/も/N/面/の/N/が/ある/	彼の言うことにも一面の真理がある There is a certain amount of truth to what he says.
	AM041610	N/の/言う/N/に/も/N/が/ある/	彼の言うことにも半面の真理がある There is some truth in what he says
	AG003107	N/を/飾る/N/は/ない/	身を飾るための金はない I don't have the money to dress fancy.
	AD017326	N/は/ない/	[とげのない] バラはない Every rose has its thorn.=No rose without a thorn.
4	AN068784	N/も/漏らさ/ぬ/N/だ/	水も漏らさぬ仲だ They are a pair of kid-gloves.
	AG000584	N/だ/	[危なげのない] 演技だ It's a sure performance.
	AM034947	N/は/N/を/N/離れ/た/N/だ/	ここは本土を遠く離れた島だ This is an island remote from the mainland.
	AD005568	N/は/N/N/だ/	それは十分 [正当な] 理由だ It's a good enough reason.
. 5	AM021854	N/に/は/N/を/知らせ/ない/N/が/いい/	子どもには [父親の] 死を知らせないほうがいい It is better not to tell the child of his father's death.
	AC031220	N/は/N/が/いい/	[この] 布地は [滑らかで] 手触りがいい This material is smooth and feels very soft to the touch.
	AD024811	N/酒/は/発酵し/て/N/に/なる/	りんご酒は発酵して酢になる Cider works to produce vinegar.
	AN006439	N/に/なる/	[今一つで] 十になる One more will make ten.

4.9　包含関係にあるパターン調査結果の考察

今回例文の平均単語数が 11.41(置換テーブルで置換後は 10.29) 個であるため、削除レベルが 3 以上になるとパターンの類似性が低くなった。

置換条件 (5) では入力パターン 100 パターンのうち 99 パターンが包含関係にあるパターンをデータベースから検出できた。これは、データベースの中に V/V/といった汎用的なパターンがあったためである。また、動詞を V と置換すると、包含関係にあるパターンがデータベースから複数個抽出されることが多く、英語の文型一致数が高くなった。

今回パターン間の構造のみに着目して意味は考慮にいれず類似性を調査した。単語や名詞を N と置換した条件で包含関係を調査したところ、抽出数は少ないがパターン元の日本語文はかなり構造的にも意味的にも類似する傾向にあった。しかし、置換条件を増やすと、包含関係にあるパターンを多く抽出できたがパターン元の日本語文には構造的な類似性しか見られない場合が多かった。例を表 16,17 に示す。これは動詞を置換することによりパターンの意味が大幅に減少したためと推測される。

表 16: 置換条件による包含パターンの違い (削除レベル 1)

置換方法	パターン	例
N	N/の/言う/N/に/も/N/面/の/N/が/ある/ N/の/言う/N/に/も/N/　の/N/が/ある/	彼の言うことにも一面の真理がある There is a certain amount of truth to what he says. 彼の言うことにも半面の真理がある There is some truth in what he says.
NV	N/は/N/に/V/て/N/を/V/でい/た/ N/は/N/に/V/て/N/を/V/　　た/	彼女は窓際に座って本を読んでいた She sat reading a book by the window. 彼は勝利に酔って歓声を上げた He crowed over his victory.

表 17: 置換条件による包含パターンの違い (削除レベル 2)

置換方法	パターン	例
N	N/は/N/の/N/も/せ/ず/に/引っ越し/てしまっ/た/ N/の/N/も/せ/ず/に/引っ越し/てしまっ/た/	彼女は何の挨拶もせずに引っ越してしまった She moved out without even saying good-bye. 何の挨拶もせずに引っ越してしまった She moved away without saying a word.
NV	N/は/N/で/負けず嫌いな/N/だ/と/V/ N/は/N/N/だ/と/V/	ぼくは自分で負けず嫌いな気性だと思う I would say I've got an unyielding[open] disposition. それは不正行為だと思う I call that dishonest.

5　おわりに

5.1　結果

本論文では意味属性などの情報を用いず、形態素の表層的な単語列情報のみで類似日本語パターンを調査を行うため、いくつかの置換方法を提案した。

この方法を CREST 重文複文の例文 91680 文を対象に、被覆率・包含関係にあるパターン数の変化を調査した。また、包含関係にあるパターンの対訳英文の文型構造の比較によりパターンの構造的類似性を評価した。結果、被覆率は 18.46%と低いが、包含関係にあるパターンは入力文として使用した 100 文ほぼすべてにパターンが存在した。

今回用いた例文約９万文では、用例翻訳を実現させるために必要となるデータ量を推定するまでにはいたらなかった。しかし、各置換方法の日本語パターンに与える影響が顕著に現れていた。

5.2　今後の課題

今後はさらなる日英対訳データを増やすとともに、入力文を増やし、より多くのパターンについて包含関係パターンの調査を行う必要がある。また、連体詞や副詞の置換等行っていない置換作業をすることでさらに被覆率を向上させることができると考えられる。動詞の置換に関しては、用言意味属性を用いた類似パターンの検索・評価を行う必要がある。

本論文では対訳英文側の文型の類似性を入力文の対訳英語パターンを基準に評価した。しかし、異なる英語文型でも意味的には対応する場合があり、その場合についても調査する必要がある。

今回 DP マッチングの削除エラーのみに着目して調査を行ったが、置換エラーも視野にいれたところさらに多くの類似なパターンが検出できた。よって置換エラーも視野にいれた調査が必要である。

謝辞

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 研究室の方々に対し深く感謝します。altjaws は NTT との共同研究の下に使わせて頂きました。

また、参考にさせて頂いた文献の著者の方々に対して感謝します。

参考文献

[1] 市原 池原:文字の並びから見た類似文検索手法, 情報処理学会第 57 回全国大会,Vol.3,pp.237-238(1998).

[2] 村上 池原 徳久:日本語英語の文対応の対訳データベースの作成,「言語、認識、表現」第 7 回年次研究会,2002 年 11 月

[3] Hidden Markov Model Toolkit(HTK)
http://htk.eng.cam.ac.uk/

[4] 南 敏:情報理論:産業図書 (1988〜1994)

付録

出典番号表

出現回数の多かったパターン TOP100

単語
名詞を N へ置換
名詞を N へ置換、置換テーブル (表 2) の処理
名詞を N へ置換、動詞を V へ置換
名詞を N へ置換、動詞を V へ置換、置換テーブル (表 2) の処理

削除レベル 5 までの関係にあるパターン

入力文 100 文

単語
名詞を N へ置換
名詞を N へ置換、置換テーブル (表 2) の処理
名詞を N へ置換、動詞を V へ置換
名詞を N へ置換、動詞を V へ置換、置換テーブル (表 2) の処理

出典番号表

本論文で記載した出典番号は以下の表に準拠する。なお*は任意の数字が入る。

表 18: 出典番号表

出典の記載	名称
AC******	アンカー和英辞典
AD******	アンカー英和辞典
AE******	学研英和辞典
AF******	基本語用例辞典
AG******	英語表現辞典
AI******	英文ビジネスレター英文文例辞典
AM******	講談社和英辞典
AN******	斉藤英和大辞典
AO******	小倉書店 英語文例 文例辞典
AS******	ビジネス技術実用英語大辞典
AU******	佐良木コーパス
AV******	白井コーパス
AW******	比較構文
AX******	鳥取大学池原研究室 澤田康子コーパス ：因果関係構文
AY******	英語教師用データベース

出現回数の多かったパターンTOP100

(単語区切り)

出現回数の多かったパターンTOP100

(名詞をNへ置換)

出現回数の多かったパターンTOP100

(名詞をNへ置換、置換テーブル(表2)の処理)

出現回数の多かったパターンTOP100

(名詞をNへ置換、動詞をVへ置換)

出現回数の多かったパターンTOP100

(名詞をNへ置換、動詞をVへ置換、置換テーブル(表2)の処理)

削除レベル5までの関係にあるパターン

入力文