概要

機械翻訳において、意味としてまとまりを持つパターンを用いた翻訳方法が注目されている。このパターンを用いた翻訳を実現するためには、日本語パターンと英語パターンを登録した日英対訳パターンを作成する必要がある。しかし、膨大な日英対訳パターンの作成は人手では困難であるので、自動的に作成する方法が必要となる。

そこで本研究では、対訳コーパスを用いて、日本文・英文から N-gram 統計処理によりパターンを抽出した。そして、パターンを含む文の文番号を比較し、対応率を計算することで、パターンを自動的に対応づけた。

上記の手法をアンカー英和辞典の 5,000 文に適用し、対訳パターンを抽出した。抽出した対訳パターンから、きちんと対訳となっているかを人手で判断し、対訳パターンを得た。

本手法により、日英対訳パターンを自動抽出できる見通しとなった。

目 次

1 はじめに **1**

2 パターンの抽出方法 **2**

2.1 連鎖型共起表現と離散型共起表現 2

2.2 連鎖共起表現 N-gram 統計処理方法 3

3 品詞の置き換え **4**

4 パターンの対応 **5**

4.1 文番号の比較 5

4.2 対応率 6

5 実験方法 **7**

5.1 本手法の手順 7

5.1.1 プログラムの流れ 8

5.1.2 対応率の算出 10

5.1.3 対訳パターンの候補を抽出 11

5.1.4 評価 12

6 実験結果 **13**

7 考察 **17**

7.1 抽出された対訳パターンについて 17

7.2 正解率について 17

7.3 評価の向上について 18

8 おわりに **19**

図 目 次

1　手法の手順 7

表 目 次

1 実験結果 (1) 14
2 実験結果 (2) 15
3 実験結果 (3) 16

1　はじめに

機械翻訳において、意味としてまとまりを持つパターンを用いた翻訳方法が注目されている。このパターンを用いた翻訳を実現するためには、日本語パターンと英語パターンを登録した日英対訳パターンを作成する必要がある。しかし、膨大な日英対訳パターンの作成は人手では困難であるので、自動的に作成する方法が必要となる。

そこで本研究では、対訳コーパスを用いて、日本文・英文からパターンを抽出し、パターンを含む文の文番号を比較することで、パターンを自動的に対応づける方法を提案した。そして、アンカー英和辞典の 5,000 文に適用することで、対訳パターンを抽出した。その結果、31 個の対訳パターンを得ることができた。

以下、2 章ではパターンの抽出方法について、3 章では置き換え処理について、4 章ではパターンを対応させる方法について、5 章では評価方法について、6 章では実験結果について、そして 7 章では考察について、最後に 8 章で結論を述べる。

2　パターンの抽出方法

パターンを抽出する方法として、N-gram を使用し、連鎖型共起表現と離散型共起表現を抽出する方法が提案されている [1]。

2.1　連鎖型共起表現と離散型共起表現

連鎖型共起表現とは、連語やフレーズのように連続した文字列を構成するものである。

離散型共起表現とは、係り結びや呼応関係のように複数の文字列が、文中の離れた場所に現れるものである。

例 1 に連鎖型共起表現の例を、例 2 に離散型共起表現の例を示す。

(例 1)

歩い てきた

走ってきた

(例 2)

もし、晴れ なら、運動会はある だろう

もし、雨 なら、中止 だろう

例 1 の「てきた」が連鎖型共起表現で、例 2 の「もし、~なら、~だろう」が離散型共起表現となっている。

離散型共起表現は、連鎖型共起表現の文字列が文中で共起したものと考えることができる。

そこで、本研究ではまず、パターンを連鎖型共起表現に限定し、日英対訳パターンの作成を目指す。

2.2　連鎖共起表現 N-gram 統計処理方法

連鎖共起表現 N-gram 統計処理方法とは、例 1 の「てきた」のように、原文データに現れる共通の連鎖型共起表現を抽出する方法である。

例 1 では、「てきた」の内部の「てき」や「きた」も共通の連続的な文字列である。このように、一度抽出された文字列の内部に含まれる文字列を部分文字列という。この部分文字列の抽出方法には、3 つの方法がある。

- 無抑制型 … 部分文字列も抽出する。
- 強抑制型 … 部分文字列は抽出しない。
- 弱抑制型 … 部分文字列が他の場所に独立していれば、抽出する。しかし、一度抽出された文字列の内部だけに含まれているときは、抽出しない。

例 1 から抽出される文字列を以下に示す。

- 無抑制型 … 「てきた」の他、部分文字列である「てき」、「きた」が抽出される。
- 強抑制型 … 部分文字列である「てき」、「きた」は抽出されず、「てきた」だけが抽出される。
- 弱抑制型 … 部分文字列である「てき」、「きた」は「てきた」の内部だけに含まれているので抽出されず、「てきた」だけが抽出される。

弱抑制型で部分文字列が抽出される場合は、例 3 の「きた」のように、部分文字列が他の場所に独立している場合である。例 3 では、「てきた」、「きた」が抽出される。「てき」は「てきた」の内部だけに含まれているので抽出されない。

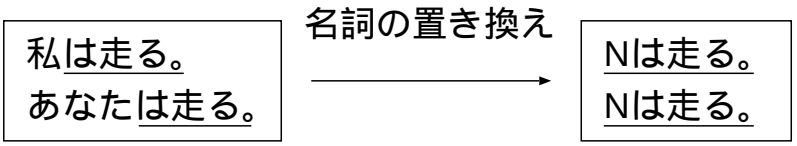
(例 3)

歩いてきた
走ってきた
春がきた

3　品詞の置き換え

パターンを抽出する場合、効率良くパターンを抽出するために、品詞ごとに単語を置き換え、パターンを抽出する方法が提案されている [2]。パターンとしての構造を残すため、本研究では名詞だけを N に置き換える。

(例4). 原文データ



例 4 の原文データでは、連鎖共起表現 N-gram 統計処理方法により、「は走る。」が抽出される。

名詞を N に置き換えると、「N は走る。」が抽出される。

例 4 のように、名詞を N に置き換えることで、効率良くパターンを抽出することができる。

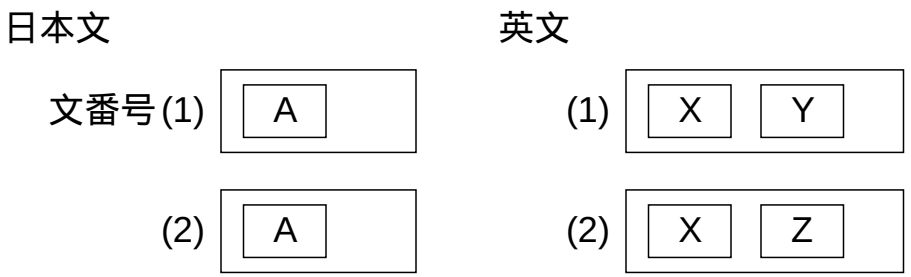
4　パターンの対応

4.1　文番号の比較

本研究では、文番号を比較することで対訳パターンを検索する。そして、文番号が一致していれば対訳パターンとする。

以下に対訳コーパスの例を示す。例 5 では、パターン A を含む文番号 (1) の日本文とパターン X, Y を含む文番号 (1) の英文が対訳で、パターン A を含む文番号 (2) の日本文とパターン X, Z を含む文番号 (2) の英文が対訳となっている。

(例5). 対訳コーパスの例



文番号 (1) の文で対訳パターンを考えると、パターン A はパターン X, Y のどちらと対訳となっているのか判断できない。同様に、文番号 (2) の文では、パターン A はパターン X, Z のどちらと対訳となっているのか判断できない。

原文データが増えれば、対訳パターンを含む文も増えるので、対訳であるパターン同士の一致数が多くなると考えられる。

そこで、文番号の一致数を求め、一致数が一番多いパターン同士を対訳パターンとする。

例 5 の文番号の一致数は以下のようになる。

A : X = 一致数 2(文 (1),(2) が一致)

A : Y = 一致数 1(文 (1) が一致)

A : Z = 一致数 1(文 (2) が一致)

そして、一致数の多いパターン A とパターン X を対訳パターンとする。

4.2 対応率

文番号の一致数が多くても一致している割合が低ければ、対訳パターンではないと考えられる。しかし、前節の方法では、一致数の多さのみを考慮しているため、一致している割合が低くても対訳パターンとなってしまう。そこで、効率良く対訳パターンを抽出するため、対応率を定義する。そして、対応率が高いパターン同士を対訳パターンとする。

対応率 (%) = 一致数 / 抽出回数

以下に例を示す。例 6 では、パターン A とパターン X が対訳で、パターン B とパターン Y が対訳となっている。そして、原文データからパターン A が 10 回、パターン B が 4 回抽出されたとする。

(例 6) 対訳パターンの例

A : X = 一致数 3

B : Y = 一致数 3

対応率は以下のようになる。

A : X = 対応率 30 % (3 / 10)

B : Y = 対応率 75 % (3 / 4)

→ 対応率の高いパターン B とパターン Y を対訳パターンとして抽出する。

前節の方法では、一致数の多さから、対応率の低いパターン A とパターン X も対訳パターンとして抽出される。

例 6 のように、対応率を求め、対応率が高いパターン同士を対訳パターンとすることで、効率良く対訳パターンを抽出することができると考えられる。

また、例 5 のように、複数の候補が考えられる場合には、対応率が一番高いパターン同士を対訳パターンとする。

対訳パターンを抽出した後、本当に対訳パターンとなっているかの評価を人手で行うため、対訳パターンである可能性が高いものだけを抽出する必要がある。

5　実験方法

5.1　本手法の手順

本手法の手順の流れを以下に示す。

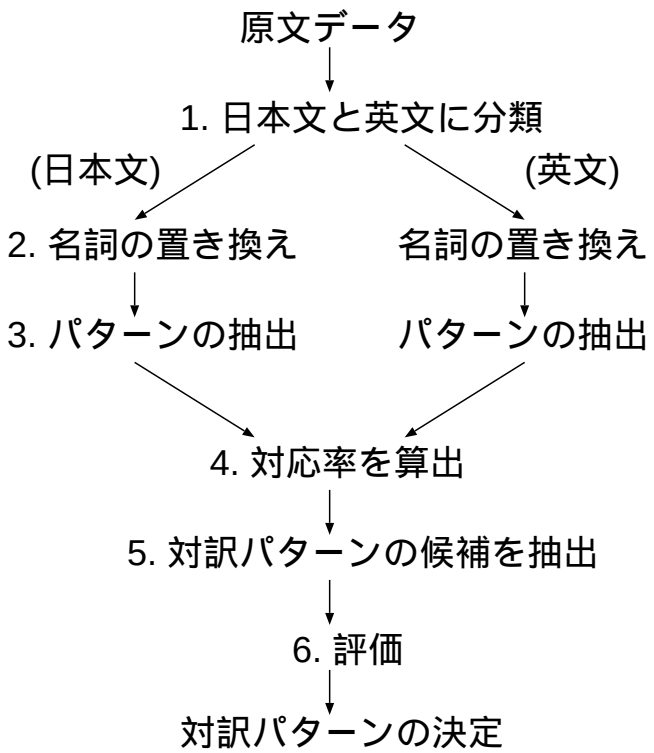


図 1: 手法の手順

1. ~ 5. はプログラムを作成し、処理を行う。以下に、処理の説明を行う。

5.1.1 プログラムの流れ

本研究では、原文データとして対訳コーパスを使用する。日本文と英文に分けて実験を行うため、最初に対訳コーパスから日本文と英文に分ける必要がある。

例 7 のような原文データの場合を考える。

例7. 原文データの例

原文データ

文(1) 私は学校へ行く。
I go to school.
文(2) あなたは学校へ行く。
You go to school.
文(3) 私は寝る。
I sleep.
文(4) あなたは寝る。
You sleep.

原文データを日本文と英文に分類する。

日本文データ

私は学校へ行く。
あなたは学校へ行く。
私は寝る。
あなたは寝る。

英文データ

I go to school.
You go to school.
I sleep.
You sleep.

日本文データと英文データの名詞を*N* に置き換える。

N は*N* へ行く。
N は*N* へ行く。
N は寝る。
N は寝る。

N go to *N* .
N go to *N* .
N sleep.
N sleep.

そして、名詞を N に置き換えたデータから、パターンの抽出を行う。対応率を求めるときに文番号を比較するので、パターンを含む文の文番号の検索も行う。

ここで、パターンの抽出は、連鎖共起表現 N-gram 統計処理方法を用いて、強抑制型で行った。強抑制で行うことにより、まとまりのない断片的な表現が削除できる。

例 7 では、例 8 のようなパターンが抽出される。

(例 8) 例 7 から抽出されたパターン

日本文側

・N は N へ行く。 … 抽出回数 2 (文 (1)、(2) から抽出)

・N は寝る。 … 抽出回数 2 (文 (3)、(4) から抽出)

英文側

・N go to N. … 抽出回数 2 (文 (1)、(2) から抽出)

・N sleep. … 抽出回数 2 (文 (3)、(4) から抽出)

5.1.2　対応率の算出

日本文側のパターンの文番号と英文側のパターンの文番号を比較して、一致数を求める。そして、対応率を求め、対応率が高いパターン同士を対訳パターンとする。

例 9 に、例 7 における対応率を示す。

(例 9) 例 7 における対応率

「N は N へ行く。」:「N go to N.」 … 一致数 2(文 (1),(2) が一致)

・対応率 100 %(2 / 2)

「N は N へ行く。」:「N sleep.」 … 一致数 0

・対応率 0 %

→「N は N へ行く。」と「N go to N.」を対訳パターンとする。

「N は寝る。」:「N go to N.」 … 一致数 0

・対応率 0 %

「N は寝る。」:「N sleep.」 … 一致数 2(文 (3),(4) が一致)

・対応率 100 %(2 / 2)

→「N は寝る。」と「N sleep.」を対訳パターンとする。

5.1.3　対訳パターンの候補を抽出

対応率を求めることで得られた対訳パターンから、対応率が 50 %以上で一致数が 2 以上のものを対訳パターンの候補として抽出する。

- 対応率を 50 %以上としたのは、対応率が 50 %より低い場合は、対訳パターンである可能性が低いと判断したためである。
- 一致数を 2 以上としたのは、文番号の一致数が 1 の場合は、偶然文番号が一致している可能性があると判断したためである。

例 9 で得られた二つの対訳パターンは、共に対応率が 50 %以上で一致数が 2 以上となっているので、二つとも対訳パターンとして抽出される。

ここで、例 10 のような対訳パターンが得られた場合を考える。

(例 10)

パターン A : パターン W … 対応率 80 % (4 / 5)
パターン B : パターン X … 対応率 20 % (1 / 5)
パターン C : パターン Y … 対応率 75 % (3 / 4)
パターン D : パターン Z … 対応率 100 % (1 / 1)

例 10 では、対応率が 50 %以上で一致数が 2 以上であるパターン A とパターン W、パターン C とパターン Y が対訳パターンとして抽出される。

対訳パターンを抽出した後、人手で評価を行うため、対訳パターンである可能性が低いものを削除した方が効率良く実験を行うことができる。

5.1.4　評価

対応率が 50 %以上で一致数が 2 以上である対訳パターンの中から、パターンとは言えないものを人手で削除し、対訳パターンを絞り込む。そして、人手で評価を行い、評価を 3 つに分類する。

◎ : 完全に対訳であると判断されるもの

○ : ほぼ対訳であると判断されるもの

× : 対訳ではないと判断されるもの

そして、次の式で正解率を求める。

正解率 (%) = ◎ と○ の数 / 評価対象のパターン数

対応率を求めた結果、例 11 のように、一つのパターンに対して複数のパターンが対訳パターンとして抽出される場合もある。例 11 では、パターン A はパターン X とパターン Y のどちらとも対訳となっている可能性がある。

(例 11)

パターン A : パターン X … 対応率 50 % (2 / 4)

パターン A : パターン Y … 対応率 50 % (2 / 4)

例 11 のような場合は、それぞれに対して評価を行い、対訳パターンの個数は 2 とする。

6　実験結果

本研究では、アンカー英和辞典 5,000 文、新聞記事 5,000 文、アンカー英和辞典 20,000 文に対して実験を行った。

ここで、原文データの例としてアンカー英和辞典の一部を以下に示す。

(例 12)

英文	日本文
He is a poet and diplomat	彼は詩人でもあり外交官でもある
Tom and Betty are of an age	トムとベティーは同い年だ
A dog is a faithful animal	犬は忠実な動物である
I know a man who says so	私はそう言っている人を知っている
He has a knowledge of English	彼は英語を少々知っている

アンカー英和辞典 5,000 文に対して実験を行った結果、対応率が 50 %以上で一致数が 2 以上である対訳パターンがプログラムにより 93 個抽出された。

93 個の中から、「立っていた」のように、パターンではないもの 51 個を人手で削除し、42 個で評価を行った。「立っていた」は 1 単語だけなので、パターンではないと判断し削除した。

以下に、削除したものの例を示す。

(例 13)

立っていた : N stood there ･･･ 対応率 66.7 % (2 / 3)

平等である : All N are equal ･･･ 対応率 100 % (2 / 2)

落ちていた : N was dripping ･･･ 対応率 100 % (2 / 2)

本研究では、日英対訳パターンの作成を目指しているので、例 13 のようにほぼ対訳となっていても、パターンでなければ削除した。

42 個に対する評価の結果を表 1 に示し、完全に対訳であるパターンの例を例 14 に、ほぼ対訳であるパターンの例を例 15 に示す。

表 1: 実験結果 (1)

原文データ	評価対象	完全に対訳	ほぼ対訳	対訳ではない	正解率
英和辞典 5,000 文	42	5	26	11	73.8 % (31 / 42)

(例 14) 完全に対訳であるパターンの例

N は N をとる : N catch N ••• 対応率 50 % (2 / 4)

N に会いたい : want to see N ••• 対応率 66.7 % (2 / 3)

N は N に終わった : N ended in N ••• 対応率 66.7 % (2 / 3)

N は N を信じ : N believe in N ••• 対応率 66.7 % (2 / 3)

N に沈んだ : below the N ••• 対応率 100 % (2 / 2)

(例 15) ほぼ対訳であるパターンの例 (下線部分が対訳)

N は N に沿って : along the N ••• 対応率 66.7 % (2 / 3)

N を渡った : N crossed the N ••• 対応率 100 % (2 / 2)

その N をあきらめた : N gave up the N ••• 対応率 100 % (2 / 2)

・

・

・

新聞記事 5,000 文に対して実験を行った結果、対訳パターンが 223 個抽出された。パターンではないもの 166 個を削除し、57 個で評価を行った。その結果を表 2 に示し、完全に対訳であるパターンの例を例 16 に、ほぼ対訳であるパターンの例を例 17 に示す。

表 2: 実験結果 (2)

原文データ	評価対象	完全に対訳	ほぼ対訳	対訳ではない	正解率
新聞記事 5,000 文	57	14	24	19	66.7 % (38 / 57)

(例 16) 完全に対訳であるパターンの例

N に予定される N : N scheduled for N ･･･ 対応率 50 % (2 / 4)

N に設立した N : N established in N ･･･ 対応率 66.7 % (2 / 3)

N を取り巻く N は : N surrounding N ･･･ 対応率 66.7 % (2 / 3)

・
・
・

(例 17) ほぼ対訳であるパターンの例 (下線部分が対訳)

$\underline{N \text{ を発売する}}$: release two $\underline{N}$ ･･･ 対応率 66.7 % (2 / 3)

$\underline{\text{いわゆる } N \text{ の}}$ N : $\underline{\text{the so-called } N}$ ･･･ 対応率 100 % (3 / 3)

N からの $\underline{N \text{ に頼ってきた}}$: $\underline{\text{relied on } N}$ ･･･ 対応率 100 % (2 / 2)

・
・
・

アンカー英和辞典 20,000 文に対して実験を行った結果、対訳パターンが 451 個抽出された。パターンではないもの 229 個を削除し、222 個で評価を行った。その結果を表 3 に示し、完全に対訳であるパターンの例を例 18 に、ほぼ対訳であるパターンの例を例 19 に示す。

表 3: 実験結果 (3)

原文データ	評価対象	完全に対訳	ほぼ対訳	対訳ではない	正解率
英和辞典 20,000 文	222	47	130	45	79.7 % (177 / 222)

(例 18) 完全に対訳であるパターンの例

N は N に着いた : N arrived N ⋯ 対応率 50 % (5 / 10)

N は N と結婚した : N married N ⋯ 対応率 62.5 % (5 / 8)

N は N の N がある : N has a N for N ⋯ 対応率 50 % (3 / 6)

・

・

・

(例 19) ほぼ対訳であるパターンの例 (下線部分が対訳)

$\underline{N \text{ が好きに}}$ なった : $\underline{\text{to like } N}$ ⋯ 対応率 50 % (2 / 4)

N の $\underline{N \text{ がいい}}$: $\underline{N \text{ is better}}$ ⋯ 対応率 50 % (2 / 4)

$\underline{N \text{ は}} N$ に $\underline{\text{反する}}$: $\underline{N \text{ is against}}$ ⋯ 対応率 50 % (2 / 4)

・

・

・

以上の結果のように、文番号の一致数は少なかったが、日英対訳パターンを自動的に抽出することができた。

7　考察

7.1　抽出された対訳パターンについて

本研究では、最初に英和辞典 5,000 文と新聞記事 5,000 文に対して実験を行った。その結果、日英対訳パターンを自動的に抽出することができた。抽出された対訳パターンは、対応率が高くて文番号の一致数が多ければ、信頼できる対訳パターンと言える。そのため、さらに一致数を増やす必要があると考えられる。

原文データを増やすと、対訳パターンを含む文も増え、文番号の一致数が増えると考えられる。よって、一致数を増やすためには大規模なコーパスへの適用が必要である。

そこで、英和辞典 20,000 文に対して実験を行ったが、一致数はあまり増えなかった。そのため、さらに大規模なコーパスへの適用が必要であると考えられる。

しかし、英和辞典 5,000 文では対訳パターンが 31 個抽出されたのに対し、英和辞典 20,000 文では 177 個抽出された。このように、原文データを増やすことで対訳パターンを増やすことができた。

7.2　正解率について

英和辞典 5,000 文では 73.8 %、20,000 文では 79.7 %の正解率が得られた。原文データを増やした際、約 6 %正解率が上昇し、効率良く対訳パターンの抽出ができた。

しかし、さらに大規模なコーパスへ適用した場合に、正解率が下がることも考えられる。正解率が低い場合は× の評価が多いということだが、対訳パターンも存在しているので、効率は悪くなるが対訳パターンの抽出は可能だと考えられる。

しかし、大規模なコーパスへ適用して対訳パターンを抽出しても、最初に抽出された対訳パターンと同じものが繰り返し抽出され、対訳パターンが増えないことも考えられる。よって、今後は、正解率が何%位で対訳パターンが増えなくなるのかを探っていくことも必要だと考えられる。

7.3　評価の向上について

本研究では、完全に対訳であるパターンが少なかったので、今後は評価を上げる手法が必要だと考えられる。その手法の提案を例を用いて行う。

本研究では、例 20 のように、対応率が低く評価の対象とならなかったパターンが存在する。

(例 20)

と N の N に：a great $\underline{N \text{ of } N}$ ･･･ 対応率 25 %

と N の N に：$\underline{N \text{ of } N}$ about the N ･･･ 対応率 25 %

・

・

・

例 20 では、「と N の N に」と対訳となっているパターンとして、複数の候補があったが、どの候補も対応率が低く、対訳をパターン決定できなかった。そこで、候補のパターンの中から、共通な「N of N」をパターンとして取り出す。その結果、対訳である「と N の N に」と「N of N」の対応率が上がり、対訳パターンとして抽出されると考えられる。そして、「と N の N に」と「N of N」の評価も高くなると考えられる。

ここで提案した手法により、評価の向上とともに、新しい対訳パターンの抽出が可能になると考えられる。

8　おわりに

本研究において、文番号を比較し、対応率を算出することで、日英対訳パターンを自動的に対応づける方法を提案した。そして、アンカー英和辞典 5,000 文に対して実験を行い、対訳パターンを抽出した。抽出された対訳パターンから、きちんと対訳になっているかを人手で評価した結果、完全に対訳であるパターン：5 個、ほぼ対訳であるパターン：26 個、対訳ではないパターン：11 個という結果が得られ、31 個の対訳パターンが得られた。また、正解率は 73.8 %となり、日英対訳パターンを自動的に抽出できる見通しとなった。

しかし、機械翻訳では多くの対訳パターンが必要とされるので、対訳パターンを増やす必要がある。また、評価を上げ、より信頼できる日英対訳パターンを作成することも重要である。よって、大規模なコーパスへの適用と、前章で提案した手法が必要とされる。

参考文献

[1] 池原, 白井, 河岡：大規模コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌 Vol.36(1995)

[2] 斎藤：大規模コーパスからの重文複文の統語構造の自動抽出, 鳥取大学卒業論文