日英対訳パターンの自動抽出に向けて

道祖尾　太祐[†]　村　上　仁　一[†]
徳　久　雅　人[†]　池　原　　悟[†]

意味のまとまりを持つ表現を用いた機械翻訳方式が注目されている．その実現には原言語と目的言語の表現を対としたパターンを大量に作成する必要がある．ここで，表現対の意味的対応は人手で判断するため，その対応関係のありうる表現対に絞り込むべきである．そこで，本研究では，パターン候補の自動作成手法を提案する．具体的には，日英対訳コーパスに N-gram 統計処理を用いて日本語文および英語文から別々に部分表現を取り出し，対訳の関係を利用して表現の対を作りパターンの候補とする．初期の実験として，単文対からなる日英対訳コーパスから候補を作成したところ，約 6 割の候補は人手による容易な判別でパターン化できることを確認した．
キーワード：対訳パターン, 対訳コーパス, N-gram, 自動抽出

Towards automatic extraction of Japanese/English expression pattern-pairs

Daisuke Sainoo,[†] Jin'ichi Murakami,[†] Masato Tokuhisa[†]
and Satoru Ikehara[†]

Large semantical expression pattern-pairs are required for a pattern based machine translation. As the semantical relation between a source language expression and a target one must be inspected by hand, it is necessary to assist in extraction of possible pattern-pairs. This paper proposes an automatic extraction method for candidates of the pairs from Japanese/English parallel corpus as follows: First Japanese and English expression are independently picked up from the corpus. Next these expressions are connected each other based on the corpus number. In our preliminary experiment, candidates were obtained from Japanese/English corpus that consist of simple sentence pairs, and approximately 60 percent of the candidates were adequate to the semantical pairs.
Keyword: expression pattern pair, parallel corpus, N-gram, automatic extraction

1. はじめに

日英機械翻訳において，翻訳知識の獲得は重要な課題の一つである．結合化文法を用いた日英機械翻訳システム ALT-J/E では，日本語と英語の文型パターン対という言語間の等価あるいは近似的な表現に変換する知識が用いられ，高い精度の日英翻訳が実現されている．このパターンは，用言を中心として単文レベルで作成されたものである[1)]．

一方，単文内，単文間，あるいは重文・複文において同様の変換知識を作成する場合，用言にこだわらず，単語を組み合わせた意味的にまとまりを持つ表現を単位として翻訳する方法が必要とされる．この方法を用いた翻訳を実現するためには，同じ意味を持つ “日本語表現” と “英語表現” を対にした “日英対訳パターン” を作成する必要がある．日本語表現の例として「〜するために」があり，英語表現の例として「in order to」がある．これらの日英対訳パターンは人手で作成する必要があるが，大量の日英対訳パターンを人手で作成することは困難である．

そこで，本研究では，日英対訳パターンを人手で作成することを補助するため，日英対訳パターンの候補を自動的に抽出する方法を提案する．具体的には，対訳コーパスから，N-gram 統計処理方法[2)] により日本語表現を抽出する．また，同様に対訳コーパスから， N-gram 統計処理方法により英語表現も抽出する．そして，同じ対訳文から抽出されている日本語表現と英語表現を探すことで，連続する単語から成る日本語表現と英語表現の日英対訳パターンの候補を抽出する．そして，人手で評価を行い，本手法の性能調査を行う．

2. 日英対訳パターンの自動的な抽出方法

本研究では，日英対訳コーパスから，同じ意味を持つ日本語表現と英語表現を対にした日英対訳パターンの候補を自動的に抽出する方法を提案する．

具体的には，表 1 のように，日本文一文と英文一文が対応している対訳コーパスの日本文，英文の双方に対して，ある一定回以上出現した表現を自動的に発見し，抽出する．そして，得られた日本語表現と英語表現を意味的に対応づける方法である．

対訳コーパスの日本文，英文の双方から抽出される表現は，連続的な文字列である連鎖型共起表現と，離れた場所にある文字列である離散型共起表現が対象である．

† 鳥取大学工学部
Faculty of Engineering, Tottori University
{sainoh,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

以下，本章では，連鎖型共起表現を例にとって，抽出された日本語表現と英語表現を意味的に対応づける方法を示す．

表 1 対訳コーパスの例 (1)
Table 1 Example of parallel corpus(1)

文番号	日本文	英文
(1)	風呂が熱い．	the bath is too hot.
(2)	彼は立派な文を書く．	he writes a fine style.
(3)	これは別の品です．	this is a different article.

2.1 日本語表現と英語表現を対応づける方法

対訳コーパスの日本文から抽出された日本語表現と，対訳コーパスの英文から抽出された英語表現に対して，それぞれの表現を含む文の文番号を比較する．そして，同じ文番号の日本語表現と英語表現は日英対訳パターンである可能性が高いと仮定し，日英対訳パターンの候補を自動的に抽出する．

図 1 の例では，日本語表現「bc」を含む日本文と英語表現「BC」を含む英文が同じ文番号である (文番号 (1))．また，日本語表現「gh」を含む日本文と英語表現「GH」を含む英文が同じ文番号である (文番号 (2))．

文番号 (1) の文では，日本語表現「bc」と英語表現「BC」が日英対訳パターンである可能性が高い．同様に，文番号 (2) の文では，日本語表現「gh」と英語表現「GH」が日英対訳パターンである可能性が高い．

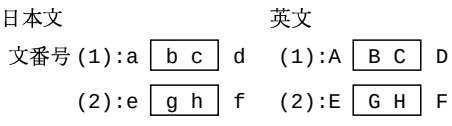

図 1 対訳コーパスの例 (2)
Fig. 1 Example of parallel corpus(2)

2.2 日本語表現と英語表現の対応づけにおける問題点

前節において，文番号を比較して日英対訳パターンの候補を発見する方法を示したが，文番号を比較するだけでは日英対訳パターンの候補を発見できない場合がある．

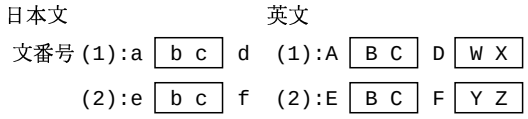

図 2 対訳コーパスの例 (3)
Fig. 2 Example of parallel corpus(3)

図 2 の例では，日本語表現「bc」を含む日本文と，英語表現「BC」と「WX」を含む英文が同じ文番号である (文番号 (1))．また，日本語表現「bc」を含む日本文と，英語表現「BC」と「YZ」を含む英文が同じ文番号である (文番号 (2))．

文番号 (1) の文で日英対訳パターンを考えると，日本語表現「bc」は，英語表現「BC」と「WX」のどちらと日英対訳パターンであるか判断できない．同様に，文番号 (2) の文では，日本語表現「bc」は，英語表現「BC」と「YZ」のどちらと日英対訳パターンであるか判断できない．

2.3 対処方法

そこで本研究では，日本語表現の文番号と英語表現の文番号を比較し，文番号が一致している割合が高い表現同士を日英対訳パターンの候補とする．

ここで，文番号が一致している割合を文番号の一致率と定義し，以下の式で求める．

文番号の一致率 ＝ 文番号の一致数 ／ 抽出回数

抽出回数とは，対訳コーパスから抽出した表現の抽出回数のことである．

2.4 対処方法の例

日英対訳パターンの候補を抽出する方法を，図 2 において説明する．

日本語表現「bc」は，文 (1)，(2) に含まれている．
英語表現「BC」は，文 (1)，(2) に含まれている．
英語表現「WX」は，文 (1) に含まれている．
英語表現「YZ」は，文 (2) に含まれている．

図 2 の日本語表現と英語表現の文番号を表 2 に，文番号の一致率を表 3 に示す．

表 2 日本語表現と英語表現の文番号
Table 2 Sentence number of Japanese expression and English expression

日本語表現	英語表現
「bc」 … 文番号 (1), (2)	「BC」 … 文番号 (1), (2)
	「WX」 … 文番号 (1)
	「YZ」 … 文番号 (2)

表 3 文番号の一致率
Table 3 Corresponding percentage of sentence number

日	英	文番号が一致している割合
「bc」	「BC」	100%(文 (1),(2) のうち，(1),(2) が一致)
「bc」	「WX」	50%(文 (1),(2) のうち，文 (1) が一致)
「bc」	「YZ」	50%(文 (1),(2) のうち，文 (2) が一致)

そして，文番号の一致率が高い日本語表現「bc」と英語表現「YZ」を日英対訳パターンの候補とする．

3. 日本語表現および英語表現を自動的に抽出する方法

3.1 N-gram 統計処理方法

日本語表現や英語表現を自動的に抽出する方法として，N-gram 統計処理方法[2] が提案されている．この方法は，複数の文から共通の文字列を自動的に発見し，抽出する方法である．

N-gram 統計処理方法には，以下の二つの方法がある．

- 連鎖型共起表現 N-gram 統計処理方法[2]
- 離散型共起表現 N-gram 統計処理方法[2]

前者は，連続的な共通の文字列を抽出する方法である．例えば,「するために」がある．後者は，離れた場所にある共通の文字列を抽出する方法である．例えば,「全く〜ない」がある．

本研究では，連鎖型共起表現 N-gram 統計処理方法によ

り日本語表現や英語表現の抽出を行う．

3.1.1 連鎖型共起表現 *N*-gram 統計処理方法

図 3 において，文 (1) は「ABCDE」，文 (2) は「FBCG」という文字列である．文 (1) と文 (2) の共通の連続的な文字列は「BC」である．連鎖型共起表現 *N*-gram 統計処理方法は，共通の連続的な文字列である「BC」を抽出する方法である．

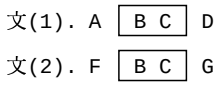

図 3 連鎖型共起表現の例
Fig. 3 Example of uninterrupted collocation

3.1.2 連鎖型共起表現の抑制方法

連鎖型共起表現 *N*-gram 統計処理方法には抽出の抑制について，以下の三つの方法がある．

(1) 無抑制型
(2) 強抑制型
(3) 弱抑制型

図 4 を用いて三つの方法を説明する．文 (1) は「ABCDEF」，文 (2) は「GBCDH」，文 (3) は「IJBC」という文字列である．

図 4 では,「BCD」,「BC」,「CD」が共通の連続的な文字列である．

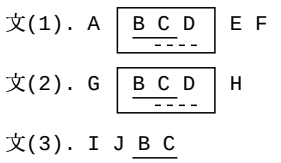

図 4 連鎖型共起表現 *N*-gram 統計処理方法の例
Fig. 4 Example of uninterrupted collocation by *N*-gram

(1). 無抑制型

無抑制型は，共通の連続的な文字列を全て抽出する方法である．ただし，無抑制型は，共通の連続的な文字列を全て抽出するため，意味を持たない断片的な表現が多数抽出される可能性がある．

図 4 の例では,「BCD」,「BC」,「CD」が抽出される．

(2). 強抑制型

強抑制型は，共通の連続的な文字列のうち，他の文字列と重なっている文字列を抽出しない方法である．

図 4 の例では,「BCD」のみが抽出される.「BC」と「CD」は「BCD」と重なっているので抽出されない．

(3). 弱抑制型

弱抑制型は，他の文字列と重なっている文字列は抽出せず，他の文字列と重なっていても他の場所で独立している文字列は抽出する方法である．

図 4 の例では,「BCD」,「BC」が抽出される.「BC」は「BCD」と重なっているが，文 (3) で独立して出現しているので,「BCD」と共に「BC」も抽出される.「CD」は「BCD」と重なっていて，他の場所で独立していないので抽出されない．

4. 実　　験

4.1 日英対訳パターンの候補を抽出する手順

本研究で提案する日英対訳パターンの候補を自動的に抽出する手順を以下に述べる．

(1) 対訳コーパスから日本語表現を抽出
(2) 対訳コーパスから英語表現を抽出
(3) 日本語表現と英語表現の対応づけを自動的に行い，日英対訳パターンの候補を抽出
(4) 日英対訳パターンの候補を人手で判断

本論文で提案した方法の有効性を調べるため，連鎖型共起表現と離散型共起表現のうち，連鎖型共起表現に対して実験を行う．

4.2 実 験 条 件

4.2.1 日本語表現と英語表現の表現の単位

連鎖型共起表現 *N*-gram 統計処理方法により日本語表現や英語表現の抽出を行う場合，意味的にまとまりを持つ表現を抽出するため，以下の二つの方法がある．

- 単語単位
  N-gram 統計処理方法で日本語表現や英語表現を抽出する場合，文字を単位として抽出する方法と，単語を単位として抽出する方法がある．文字を単位として抽出する方法では，意味をもたない文字列が多く発見される．例えば,「ストーブ」と「ストライク」では,「スト」が共通の連続的な文字列である．しかし,「スト」は意味を持たない文字列である．この問題を避けるため，本研究では，単語を単位として文字列を抽出する．
- 名詞の置換
  日本語表現や英語表現を抽出する場合，意味としてまとまりを持つ表現を抽出するために，品詞ごとに単語を置換する方法[3] が提案されている．
  本研究では，日本語表現と英語表現としての構造を残すため，名詞のみを置換する．

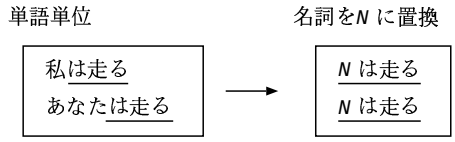

図 5 名詞の置き換え
Fig. 5 Change of noun

図 5 において，単語単位で抽出した場合,「は走る」が抽出される．しかし,「は走る」は意味としてのまとまりに欠ける．名詞を *N* に置換することで,「*N* は走る」が抽出され，意味としてまとまりを持つ表現が抽出できる．

名詞を置換すると，表現の意味や構文形成に重要な文字列を見失う可能性がある．そこで，本実験では，単語単位と名詞置換の二つの方法を実験する．

4.2.2 実験に用いる対訳コーパス

本研究では，複数の対訳辞書[4] から単文を抽出した対訳コーパスを使用する．

単語単位の実験では，対訳コーパス 8,500 文を使用する．表 4 に例を示す．

表 4 対訳コーパスの例 (4)
Table 4 Example of parallel corpus(4)

文番号	日本文	英文
(1)	言は簡を尊ぶ.	brevity is the soul of wit.
(2)	彼はきっと来る.	he is sure to come.
(3)	確かに寒い.	it sure is cold.
(4)	迷信は消えない.	superstitions survive.
(5)	彼は禁酒した.	he swore off drinking.

名詞を N に置換した場合の実験では，対訳コーパス 28,000 文を使用する．

表 4 の名詞を N に置換した対訳コーパスを表 5 に示す．この対訳コーパスでは，英文の be 動詞の現在形 (is, am, are) は全て is に，過去形 (was, were) は全て was にする．また，my, your 等の所有格も N's に統一する．

表 5 対訳コーパスの例 (5)
Table 5 Example of parallel corpus(5)

文番号	日本文	英文
(1)	N は N を尊ぶ.	N is the N of N .
(2)	N はきっと来る.	N is sure to come.
(3)	確かに寒い.	N sure is cold.
(4)	N は消えない.	N survive.
(5)	N は禁酒した.	N swore off N .

4.2.3 連鎖型共起表現の抑制方法

本研究では，連鎖型共起表現 N-gram 統計処理方法の強抑制型と弱抑制型を用いて，単文から日本語表現や英語表現抽出を抽出する．無抑制型は，意味を持たない断片的な表現が多数抽出される可能性があるため，使用しない．

4.2.4 日英対訳パターンの候補の抽出

本研究では，最終的には英対訳パターンの候補が同じ意味を持つかどうかを人手で判断する．そのため，日英対訳パターンの可能性が高いものを効率良く抽出する必要がある．そこで，日英対訳パターンである可能性が高いものを抽出するため，以下の条件を満たすものを日英対訳パターンの候補として抽出する．

(1) 文番号の一致数が 3 以上の日英対訳パターン
(2) 文番号の一致率が 50%以上の日英対訳パターン

4.3 評 価 方 法

日英対訳パターンの候補を自動的に抽出した後，人手で評価を行い，日英対訳パターンを決定する．評価は，上位 50 個に対して人手で行い，三つに分類する．

○：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，正解率を以下の二つの式で求める．

正解率 (1) ＝ ○の数 ／ 評価対象の総数
正解率 (2) ＝ ○と△の数 ／ 評価対象の総数

正解率 (1) は，完全に対訳であると判断されるものを日英対訳パターンとするので，厳しい評価である．また，正解率 (2) は，ほぼ対訳であると判断されるものも日英対訳パターンとするので，ゆるやかな評価である．

5. 実 験 結 果

本研究では，対訳辞書から抽出された単文を用いて実験を行った．

パターンの抽出は，日本語表現，英語表現共に，連鎖型共起表現 N-gram 統計処理方法の強抑制型と弱抑制型で行った．

それぞれの結果について述べていく．

5.1 単語単位の場合

単語単位では，対訳コーパス 8,500 文に対して実験を行った．

5.1.1 単語単位の強抑制型の実験結果

連鎖型共起表現 N-gram 統計処理方法の強抑制型で日本語表現と英語表現を抽出し，実験を行った．その結果，日英対訳パターンの候補が 74 個抽出された．

上位 50 個について評価を行った．その結果，完全に対訳であると判断した日英対訳パターンが多かった．

完全に対訳であると判断した例，ほぼ対訳であると判断した例，対訳ではないと判断した例を表 6 に示す．また，正解率を表 10 に示す．

表 6 評価の例 (1)
Table 6 Example of evaluation(1)

“○”	彼はまだ	he is still
	これは私の	this is my
	雨がやんだ	it stopped raining
	彼に会った	I met him
	私はのどが痛い	I have a sore throat
“△”	は丘の上にある	on the hill
	と結婚した	she married a
	の中を歩いて来た	I had to walk it
	今からでも遅くない	it is not too late
“×”	の天才だ	he is a
	へ行っている	the ship is
	は品切れです	is out of
	を読んでいる	he is well

5.1.2 単語単位の弱抑制型の実験結果

連鎖型共起表現 N-gram 統計処理方法の弱抑制型で日本語表現と英語表現を抽出し，実験を行った．その結果，日英対訳パターンの候補が 161 個抽出された．

上位 50 個について評価を行った．その結果，完全に対訳であると判断した日英対訳パターンよりも，ほぼ対訳であると判断した日英対訳パターンの方が多かった．

完全に対訳であると判断した例，ほぼ対訳であると判断した例，対訳ではないと判断した例を表 7 に示す．また，正解率を表 10 に示す．

5.2 名詞を N に置換した場合

名詞を N に置換した場合では，対訳コーパス 28,000 文に対して実験を行った．

表 7 評価の例 (2)
Table 7 Example of evaluation(2)

“○”	彼はまだ	he is still
	これは私の	this is my
	彼に会った	I met him
	雨がやんだ	it stopped raining
	お手紙ありがとう	thank you for your letter
“△”	で学校へ	to school by bus
	と言っている	they complain of the
	も知らない	does not know how to
	様子が好い	he is good-looking
“×”	の天才だ	he is a
	立てられぬ	people will talk
	をはさんだ	in the door
	て来ない	sick or something

5.2.1 名詞を N に置換した場合の強抑制型の実験結果

名詞を N に置換した後，連鎖型共起表現 N-gram 統計処理方法の強抑制型で日本語表現と英語表現を抽出し，実験を行った．その結果，日英対訳パターンの候補が 130 個抽出された．

上位 50 個について評価を行った．その結果，ほぼ対訳であると判断した日英対訳パターンが多かった．

完全に対訳であると判断した例，ほぼ対訳であると判断した例，対訳ではないと判断した例を表 8 に示す．また，正解率を表 10 に示す．

表 8 評価の例 (3)
Table 8 Example of evaluation(3)

“○”	N は N が短い	N is short in N
	N と N は	N and N
	N と N とは	N and N
	N は N を生やしている	N wears a N
	N は N が得意だ	N is good at N
“△”	N は N の N の N である	N is the N of
	N をもっている	N has a
	N を N に行った	N went to
	この N では N	in this N
“×”	あの N は N の N だ	N is a
	この N は N が	N of N
	N は N に甘い	to N’s N
	N は N が遠い	N’s N is

5.2.2 名詞を N に置換した場合の弱抑制型の実験結果

名詞を N に置換した後，連鎖型共起表現 N-gram 統計処理方法の弱抑制型で日本語表現と英語表現を抽出し，実験を行った．その結果，日英対訳パターンの候補が 438 個抽出された．

上位 50 個について評価を行った．その結果，ほぼ対訳であると判断した日英対訳パターンが多かった．

完全に対訳であると判断した例，ほぼ対訳であると判断した例，対訳ではないと判断した例を表 9 に示す．また，正解率を表 10 に示す．

5.3 正 解 率

表 10 に全ての実験の正解率をまとめる．

単語単位の場合の強抑制型は，日英対訳パターンの抽出数が少なかったが，完全に対訳であると判断した日英対訳パターンが一番多かった．

表 9 評価の例 (4)
Table 9 Example of evaluation(4)

“○”	N と N は	N and N
	N が安い	N is low
	N が短い	N is short
	N は N が短い	N is short in N
“△”	N を買った	N bought a N
	N をもっている	N has a
	N を N に行った	N went to
	N を 変えた	changed N’s N
“×”	N が痛い	N have a
	なかなかの N だ	N is a
	N に立っていた	N of the N
	N を読んでいる	N is well

表 10 実験結果
Table 10 Result of experiment

	○	△	×	正解率 (1)	(2)	抽出数
単語単位 (強抑制型)	22	10	18	44%	64%	74
単語単位 (弱抑制型)	12	17	21	24%	58%	161
名詞を置換 (強抑制型)	7	23	20	14%	60%	130
名詞を置換 (弱抑制型)	4	27	19	8%	62%	438

名詞を置換した場合の弱抑制型は，日英対訳パターンの抽出数が多くても，完全に対訳であると判断した日英対訳パターンが少なかった．

しかし，評価をゆるやかにした場合は，ほぼ同じ結果が得られたことが，正解率 (2) から分かる．

6. 考 察

6.1 本手法の有効性

本研究では，日英対訳パターンの候補を自動的に抽出する方法を提案し，その性能調査のため，連鎖型共起表現 N-gram 統計処理方法を用いて，単文から日本語表現や英語表現を抽出した．

全体的に，完全に対訳であると判断した日英対訳パターンが少なく，名詞を N に置換した弱抑制型の場合では 8%だった．しかし，評価をゆるやかにした場合の正解率 (2) の平均は 61%だった．

正解率 (2) は，完全に対訳であると判断した日英対訳パターンと，ほぼ対訳であると判断した日英対訳パターンの評価である．よって，評価をゆるやかにすることで日英対訳パターンの候補を自動的に抽出できる見通しとなった．

また，日本語表現や英語表現を抽出する場合，どの条件を使用するかは，入手する対訳コーパスに依存する可能性があるため，今後の検討が必要だと考えられる．

6.2 日英対訳パターンの作成

本実験において，ほぼ対訳であると判断した日英対訳パターンが多かった．しかし，ほぼ対訳であると判断した日英

対訳パターンを人手で修正すれば，完全に対訳な日英対訳パターンの作成が可能である．

表 11 は，実験において，ほぼ対訳であると判断した例である．日本語表現と英語表現を比較して人手で修正を行った．下線部分が修正を行った箇所である．

表 11 修正の例
Table 11 Example of modification

	この N では N	in this N
(修正後)	この N では N	$\underline{N}$ in this N
	N を買った	N bought a N
(修正後)	$\underline{N}$ は N を買った	N bought a N
	N をもっている	N has a
(修正後)	$\underline{N}$ は N をもっている	N has a $\underline{N}$

本研究の目的は，人手での日英対訳パターンの作成を補助するために，日英対訳パターンの候補を自動的に抽出することである．ほぼ対訳であると判断した日英対訳パターンについては，人手での修正が必要であるが，日英対訳パターンの作成を補助できると考えられる．

6.3 今後の課題

6.3.1 離散型共起表現

修正する日英対訳パターンの数が多ければ人手での修正は困難となるので，日本語表現や英語表現を抽出する段階で，より意味的にまとまりを持った表現を抽出する必要がある．

よって，今後は，連鎖型共起表現 N-gram 統計処理方法だけではなく，離散型共起表現 N-gram 統計処理方法を用いて，日本語表現や英語表現を抽出する必要がある．また，単文の他，重文や複文からも日本語表現や英語表現を抽出することも必要となる．

6.3.2 連鎖型共起表現と離散型共起表現

日本文と英文では，文を生成する品詞の順番が異なる場合があるため，連鎖型共起表現と離散型共起表現が同じ 意味を持つ可能性がある．

日本文が「私は野球をする.」，英文が「I play baseball.」である例を用いて説明する (図 6)．

図 6 の日本文では主語と動詞が離れた場所に存在し，英文では主語と動詞が連続して存在している．そして，離れた場所にある日本語表現 (離散型共起表現) と連続した英語表現 (連鎖型共起表現) が同じ意味を持っている．

(日本文) 私は野球をする.
主語 動詞
(英文) I play baseball.
主語 動詞

図 6 品詞の位置
Fig. 6 Position of part of speech

本研究では，日本語の連鎖型共起表現と英語の連鎖型共起表現から日英対訳パターンを作成した．今後は，以下の三つの場合で日英対訳パターンを作成し，信頼性のある日英対訳パターンを作成する必要がある．

(1) 日本語の連鎖型共起表現と英語の離散型共起表現
(2) 日本語の離散型共起表現と英語の連鎖型共起表現
(3) 日本語の離散型共起表現と英語の離散型共起表現

7. おわりに

本研究では，日英対訳パターンの作成を補助するため，N-gram 統計処理方法を用いて対訳コーパスから日本語表現や英語表現を自動的に抽出した．次に，日本語表現の文番号と英語表現の文番号を検索した．そして，文番号を比較し文番号が一致している割合を求めることで，日英対訳パターンの候補を自動的に抽出する方法を提案した．

本手法の性能調査のため，日本語表現と英語表現は連鎖型共起表現を用いて，単語単位の場合と，名詞を N に置換した場合で実験を行った．そして，自動的に抽出された日英対訳パターンの候補を人手で評価し，正解率を求めた．

その結果，本手法により，日英対訳パターンの候補を自動的に抽出できる見通しとなり，本手法の有効性が確認された．

今後の課題としては，機械翻訳では多くの対訳パターンが必要とされるので，日英対訳パターンを増やす必要がある．また，より信頼性のある日英対訳パターンを作成する必要がある．

本研究では，初期の取り組みとして，連鎖型共起表現 N-gram 統計処理方法を用いて実験を行ったが，今後は，離散型共起表現 N-gram 統計処理方法を用いて日本語表現や英語表現の抽出を行うことも必要である．

参 考 文 献

1) 池原悟，ほか 7 名，日本語語彙大系，岩波書店, 1997.
2) 池原悟, 白井諭, 河岡司, “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, No.11, pp.2584-2596, 1995.
3) 斎藤健太郎, “大規模コーパスからの重文複文の統語構造の自動抽出”, 鳥取大学工学部知能情報工学科卒業論文, 2000.
4) 村上仁一, “英日対訳データベースの現状”,「言語，認識，表現」第 7 回年次研究会プログラム, 2002.