N-gramを利用した日英対訳パターンの自動抽出

道祖尾 太祐　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
{sainoh,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1. はじめに

機械翻訳において，熟語や連語のような意味的にまとまりを持つ表現を単位として翻訳を行う方法が注目されている．この場合，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” を大量に作成する必要がある．この日英対訳パターンの意味的対応は人手で判断するため，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，人手での作成を補助するため，日英対訳パターンの候補を自動抽出する方法が必要となる．

従来，対訳コーパスから日英対訳パターンの候補を自動抽出する方法が提案されてきた [1][2]．しかし，自動抽出された日英対訳パターンは，単語や合成名詞が多かった．

そこで本稿では，熟語や連語のような意味的にまとまりを持つ表現を対象とした日英対訳パターンの候補を自動抽出する方法を提案する．

2. 従来の手法

熊野ら [1] は，言語間で対応付けが行える単位をユニットと定義し，日本語ユニット中の各内容語に対して機械翻訳対訳辞書から得られる訳語候補群と，英語ユニット中の内容語との対応度である対応確信度を求め，対応関係を推定している．

北村ら [2] は，日本語表現と英語表現を対応付けるため，まず，相互情報量と Dice 係数を比較し，Dice 係数の方が優れていることを示した．そして，対訳コーパスに複数回出現する任意の長さの単語列に対して Dice 係数を基にした式を定義し対応関係を推定している．しかし，自動抽出された日英対訳パターンは，単語や合成名詞が多かった．

そこで，本稿では，熟語や連語のような意味的にまとまりを持つ表現を対象とした日英対訳パターンの作成を目指し，まず，対訳コーパスの日本文および英文のそれぞれから複数回出現する日本語表現および英語表現を抽出する．そして，両者が抽出された文の相互関係を用いて日本語表現と英語表現を意味的に対応付ける方法を示す．

3. 日英対訳パターンの候補の自動抽出手順

日英対訳パターンの候補を自動抽出する手順を以下に示す．

1. 日本語表現を抽出
2. 英語表現を抽出
3. 日本語表現と英語表現の対応付け

それぞれについて以下に説明する．

3.1 日本語表現の抽出

本稿では，対訳コーパスから日本語表現を抽出する場合，連鎖型共起表現 N-gram 統計処理方法 [3] を用いる．連鎖型共起表現 N-gram 統計処理方法は，複数の文から連続的な共通の文字列を抽出する方法である．図 1 からは「ab」が抽出される．

文(1)　a b c

文(2)　a b d

図 1: 連鎖型共起表現の抽出

連鎖型共起表現 N-gram 統計処理方法は，抽出の抑制について，無抑制型，強抑制型，弱抑制型の 3 つの方法がある [4]．

強抑制型は意味を持たない文字列を抽出する可能性が低いため，本稿では強抑制型を用いて日本語表現を抽出する．

3.2 英語表現の抽出

対訳コーパスから英語表現を抽出する場合，日本語表現を抽出する場合と同様に，連鎖型共起表現 N-gram 統計処理方法の強抑制型を用いる．

3.3 日本語表現と英語表現の自動的な対応付け

3.3.1 文番号の一致率

本稿では，日本語表現および英語表現の抽出は連鎖

型共起表現 N-gram 統計処理方法を用いる．しかし，N-gram 統計処理方法は二言語間から同時に表現を抽出することはできないため，別々に抽出した日本語表現と英語表現を対応付ける必要がある．そこで，日本語表現および英語表現が抽出された文の相互関係を用いて日本語表現と英語表現を意味的に対応付ける．

具体的には，日本語表現を含む文の文番号と英語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，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して自動抽出する [5]．本稿では，文番号が一致している割合を “文番号の一致率” と定義し，式 (1) で求める．

文番号の一致率＝文番号の一致数／抽出回数…(1)

抽出回数とは，対訳コーパスから抽出された表現の抽出回数である．

3.3.2 日本語表現と英語表現を対応付ける手順

図 2 において，日英対訳パターンの候補を抽出する方法を説明する．図 2 では，日本語表現「bc」を含む日本文と，英語表現「BC」,「EF」を含む英文が対訳である (文番号 (1))．また，日本語表現「bc」を含む日本文と，英語表現「BC」,「IJ」を含む英文が対訳である (文番号 (2))．

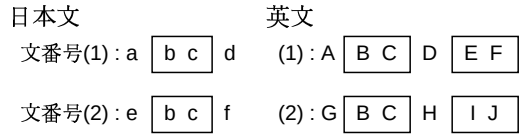

図 2: 対訳コーパスの例 (1)

図 2 の日本語表現と英語表現の文番号を表 1 に，文番号の一致率を表 2 に示す．

表 1: 日本語表現と英語表現の文番号

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

表 2: 文番号の一致率

日本語表現	英語表現	文番号の一致率
「bc」	「BC」	100% (文番号 (1), (2) のうち，(1), (2) が一致)
「bc」	「EF」	50% (文番号 (1), (2) のうち，(1) が一致)
「bc」	「IJ」	50% (文番号 (1), (2) のうち，(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「BC」を日英対訳パターンの候補とする．

4. 実験

本稿では，人手での日英対訳パターンの作成を補助するため，文番号の一致率を求め，文番号の一致率が高いものを日英対訳パターンの候補として自動抽出する．そして，抽出した日英対訳パターンの候補に対し，人手で評価を行う．

4.1 実験条件

(1) 対訳コーパス

複数の対訳辞書 [6] から抽出した単文を対訳コーパスとして使用する．対訳コーパスの文数による違いを探るため，10,000 文, 50,000 文, 80,000 文に対して実験を行う．例を表 3 に示す．

表 3: 対訳コーパスの例 (2)

日本文	英文
犬は忠実な動物である．	A dog is a faithful animal.
ピカソの絵を買った．	I bought a Picasso.
彼女はスペイン語がわかる．	She understands Spanish.

(2) 品詞の置換

日本語表現および英語表現を抽出する場合，意味的にまとまりを持つ表現を抽出するため，品詞ごとに単語を置換する方法 [7] が提案されている．本稿では，様々な種類の日英対訳パターンの候補を抽出するため，以下の 4 つの場合に対して実験を行う．

1. 文字を単位とした字面の場合
2. 名詞を置換した場合
3. 名詞・動詞を置換した場合
4. 名詞・動詞・副詞を置換した場合

(3) 日英対訳パターンの候補の抽出

日英対訳パターンである可能性が高いものを効率良く抽出するため，文番号の一致数が 2 以上で，かつ，文番号の一致率が 50%以上のものを日英対訳パターンの候補として抽出する．

4.2 評価方法

日英対訳パターンの候補を自動抽出した後，人手で評価を行い，日英対訳パターンを決定する．評価はランダムに選んだ 50 個に対して行い，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

評価 “△” には，人手で修正を行うことで，日英対訳パターンが作成できる可能性があるものも含むこととする．そして，正解率を式 (2) で求める．

正解率＝ “○” と “△” の数／評価対象の総数…(2)

5. 実験結果

5.1 正解率

本稿では，複数の対訳辞書から抽出した単文を用いて実験を行った．また，実験で抽出した日英対訳パターンの候補のうち，ランダムに選んだ 50 個に対して人手で評価を行い，正解率を求めた．

単文 10,000 文から抽出した日英対訳パターンの候補数と正解率を表 4 に示す．また同様に，50,000 文の候補数と正解率を表 5 に，80,000 文の候補数と正解率を表 6 に示す．

表 4: 10,000 文の正解率

条件	○	△	×	正解率	候補数
字面	10	37	3	94% (47/50)	112
名	9	36	5	90% (45/50)	133
名・動	13	29	8	84% (42/50)	113
名・動・副	10	28	12	76% (38/50)	108

表 5: 50,000 文の正解率

条件	○	△	×	正解率	候補数
字面	9	37	4	92% (46/50)	1,317
名	7	35	8	84% (42/50)	1,058
名・動	5	32	13	74% (37/50)	672
名・動・副	5	29	16	68% (34/50)	613

表 6: 80,000 文の正解率

条件	○	△	×	正解率	候補数
字面	14	34	2	96% (48/50)	2,735
名	7	33	10	80% (40/50)	2,181
名・動	6	36	8	84% (42/50)	1,296
名・動・副	1	35	14	72% (36/50)	1,198

実験で抽出した日英対訳パターンの候補は，10,000 文ではほぼ 110 個だった．50,000 文と 80,000 文では字面の場合の候補が最も多く，品詞の置換を行うと候補が少なくなった．正解率は，字面の場合が高く，品詞の置換を行うと低くなった．

全体では，80,000 文の字面の場合で抽出した日英対訳パターンの候補が多く，正解率も高かった．品詞の置換を行った場合は抽出した日英対訳パターンの候補が減り，正解率も下がった．

5.2 日英対訳パターンの例

完全に対訳であると判断した評価 “○” の例を表 7 に，ほぼ対訳であると判断した評価 “△” の例を表 8 に，対訳ではないと判断した評価 “×” の例を表 9 に示す．

表 7: 評価 “○” の例

条件	日本語表現	英語表現
字面	旅行に行った	went on a trip
名	N は N が得意だ	N is good at N
名・動	N は N に V れていた	the N was V-ed in N
名・動・副	N の N は $AdvV$ れている	N's N is V-ed Adv

表 8: 評価 “△” の例

条件	日本語表現	英語表現
字面	この場合には	in this case the
名	N は N を 取りもどした	N recovered N's N
名・動	N から この N を V た	N V-ed this N from N
名・動・副	$AdvN$ の N が悪い	N is wrong with N's N

表 9: 評価 “×” の例

条件	日本語表現	英語表現
字面	の安定は政府の	is the duty of
名	N が N 斤	N have put on
名・動	お N が V	V N for N's N
名・動・副	N が $AdvV$ ぬ	N can Adv V the N

表 7 や表 8 から，意味的にまとまりを持つ日英対訳パターンを抽出できたことが分かった．

6. 考察

6.1 連鎖型共起表現を対象にした日英対訳パターン

6.1.1 正解率の推移

品詞の置換を行った日英対訳パターンは，字面の場合に比べ抽出した日英対訳パターンの候補が少なく，正解率も低かった．品詞の置換を行うことで字面の情報が失われ，日本語表現と英語表現の意味的対応の判断が困難だったためだと考えられる．しかし，様々な種類の日英対訳パターンが必要となる場合があるため，品詞の置換を行った日英対訳パターンは，今後，日英対訳パターンを作成する上で有効な指針になることが期待できる．

6.1.2 日英対訳パターンの作成

本実験では，ほぼ対訳であると判断した評価 “△” の日英対訳パターンが多かった．評価 “△” の日英対訳パターンは，人手で修正を行うことで完全に対訳な日英対訳パターンを作成することができる．

評価 “△” の日英対訳パターンを人手で修正し日英対訳パターンを作成する場合，人手での修正は困難な

場合があるが，日英対訳パターンの作成を補助できると考えられる．

表 8 に対して人手で修正を行った結果を表 10 に示す．下線部分が修正箇所である．

表 10: 修正の例

日本語表現	英語表現
この場合には	in this case (“the” を削除)
N は N のN を取りもどした	N recovered N’s N
N はN からこの N を V た	N V-ed this N from N
N は$AdvN$ の N が悪い	N is Adv wrong with N’s N

6.2 離散型共起表現を対象にした日英対訳パターン

6.2.1 日英対訳パターンの候補数

本手法では,「between 〜 and」のような離れた場所にある離散型共起表現を対象にした日英対訳パターンの抽出も可能である．

単文 80,000 文の字面の場合では，日英対訳パターンの候補を 3 個抽出した．結果を表 11 に示す．数は少ないが，日英対訳パターンを得ることができた．

表 11: 離散型共起表現の結果

評価	日本語表現	英語表現
△	は、いくつかの点で 〜と異なる	differs from 〜 in several ways
△	大佐は〜位が下である	a colonel is 〜 a general
○	翻訳は原文の 〜を存している	the translation retains the 〜 of the original

6.2.2 日本語表現と英語表現の数

離散型共起表現を対象にした日英対訳パターンは，連鎖型共起表現を対象にした場合に比べ，抽出した日英対訳パターンの候補数が少なかった．そこで，単文 80,000 文の字面の場合に対して，連鎖型共起表現を対象にした場合と離散型共起表現を対象にした場合の比較を行った．結果を表 12 に示す．

表 12: 連鎖型共起表現と離散型共起表現の比較

	日本語表現の数	英語表現の数	候補数
連鎖型	13,629	12,691	2,735
離散型	1,331	318	3

表 12 から，離散型共起表現を対象とした場合は，連鎖型共起表現を対象とした場合に比べ，対訳コーパスから別々に抽出した日本語表現と英語表現の数が少ないことが分かった．今後は，離散型共起表現を増やすため，日本語表現および英語表現をさらに効率良く抽出する必要があると考えられる．

7. おわりに

本稿では，人手での日英対訳パターンの作成を補助するため，まず，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および英語表現を抽出した．次に，文番号の一致率を求めることで日英対訳パターンの候補を自動抽出した．そして，抽出し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．

連鎖型共起表現を対象に，品詞の置換を行い実験を行った結果，様々な種類の意味的にまとまりを持つ日英対訳パターンの候補を自動抽出することができた．それぞれの日英対訳パターンの候補数と正解率を求めた結果，80,000 文の字面の場合が抽出した日英対訳パターンの候補が多く，正解率は 96%だった．また，全ての正解率の平均を求めたところ，約 8 割の候補が人手での判断により，日英対訳パターンの作成が可能であることが分かった．抽出した日英対訳パターンは，ほぼ対訳であると判断した評価 “△” が多かった．評価 “△” の日英対訳パターンは人手で修正を行うため，今後，評価 “△” が減り，完全に対訳であると判断される日英対訳パターンが増えれば，効率良く日英対訳パターンを作成することができると考えられる．

一方，離散型共起表現を対象に実験を行った結果，自動抽出した日英対訳パターンの候補は少なかった．今後は，離散型共起表現を対象にした日英対訳パターンの候補を増やす必要があると考えられる．

参考文献

[1] 熊野明, 平川秀樹, “対訳文書からの機械翻訳専門用語辞書作成”, 情報処理学会論文誌, Vol.35, No.11, pp.2283-2290, 1994.

[2] 北村美穂子, 松本裕治, “対訳コーパスを利用した対訳表現の自動抽出”, 情報処理学会論文誌, Vol.38, No.4, pp727-736, 1997.

[3] 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.

[4] 波宏誠, “N-gram 統計を応用した日本語共起表現辞書の作成”, 鳥取大学工学部知能情報工学科卒業論文, 1999.

[5] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟, “日英対訳パターンの自動抽出に向けて”, 情報処理学会研究報告, 2003-NL-153, pp.113-118, 2003.

[6] 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.

[7] 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.