タイトル	アイヌ語-日本語対訳コーパスを対象とした局所着目型学習による対訳語の自動抽出
著者	越前谷，博；荒木，健治；桃内，佳雄
引用	北海学園大学工学部研究報告，32：41-63
発行日	2005-02-21

アイヌ語－日本語対訳コーパスを対象とした局所着目型学習による対訳語の自動抽出

越前谷　博[†]・荒木健治[‡]・桃内佳雄[†]

Automatic Extraction of Bilingual Word Pairs Using Local Focus-based Learning from an Ainu-Japanese Parallel Corpus

Hiroshi ECHIZENYA[†] · Kenji ARAKI[‡] · Yoshio MOMOUCHI[†]

あらまし

対訳辞書の品質向上のために対訳コーパスから対訳語を抽出することは、より自然な訳語や新たな表現の訳語を得るために非常に有効である。しかし、これまでの対訳コーパスから対訳語を自動抽出する研究では、大規模な対訳コーパスが不可欠となる。様々な言語を対象とした場合、常に大規模な対訳コーパスが得られるとは限らないため、この問題は深刻である。我々は、この問題点を解決するために新たな手法として、局所着目型学習を用いた対訳語の自動抽出手法を提案する。本手法は、対訳文中の局所部分を対象に語の対応関係を決定するため探索範囲を限定できる。さらに、言語間のコロケーションを利用することで、抽出対象の対訳語そのものの出現頻度が低い場合でも対訳語を効率よく自動抽出することが可能である。性能評価実験では、名詞および動詞対訳語の平均出現頻度が1.96である小規模なアイヌ語－日本語対訳コーパスを対象に名詞対訳語と動詞対訳語の自動抽出を試みた。実験の結果、再現率として54.0%、適合率として60.8%が得られた。この結果は、統計的手法の再現率に比べ10%以上高い値であり、本手法の有効性を示すものである。さらに、既存の辞書には存在しない、より自然な訳語や新たな表現の訳語の抽出も確認された。

Abstract

Using a parallel corpus is effective to obtain many high-quality equivalents. Most of methods that extract bilingual word pairs requires a large parallel corpus. However, it is difficult to obtain such a large parallel corpus easily for various languages. This paper proposes a new learning method for extraction of bilingual word pairs to solve this serious problem. The method, called Local Focus-based Learning (LFL), limits the search range to decide corresponding words focusing on parts of sentences. Moreover, it can efficiently extract bilingual word pairs using the information of collocations between source and target languages even if frequency

†北海学園大学工学部電子情報工学科
Dept. of Electronics and Information, Hokkai-Gakuen University
‡北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

of appearances of bilingual word pairs is very low. The system based on LFL described in this paper extracts noun and verb bilingual word pairs from a small Ainu-Japanese parallel corpus. Evaluation experiments indicated that the recall was 54.0% and the precision was 60.8%.The recall of our system was more than 10% higher than the recall of statistical methods. Consequently, we confirmed the effectiveness of LFL. Moreover, natural equivalents that do not exist in Ainu-Japanese dictionaries were also extracted.

1 はじめに

近年、様々な分野でのグローバル化が進み、英語ほど国際的に広く使われていない言語にも強い関心が寄せられている［1］。種々の言語を扱うために対訳辞書は重要かつ基本的な言語資源である。そして、対訳辞書の品質向上のために、対訳コーパスより対訳語を得ることは有効である。対訳コーパスを利用することによって、既存の辞書の訳語には存在しない、より自然な訳語や新たな表現の訳語が得られる。しかし、その作業を人手で行うことは膨大な労力を伴うため、システムによって自動化することが望ましい。こうした背景から、対訳コーパスより対訳知識を自動抽出するための研究が盛んに行われている。これらの研究は手法の観点より大きく２つに分類される。一つは基本対訳辞書などの既存の言語知識に基づく解析的なアプローチ［2,3］である。このアプローチは、静的な知識に強く依存しているため、問題点として多様な言語現象に対処することの困難さが挙げられる。そこで、対訳コーパスに対して統計処理を行うことで対訳知識を抽出する統計的なアプローチ［4,5,6,7,8,9,10］が提案されている。このアプローチの問題点は、大規模な対訳コーパスが必要となることである。様々な言語を対象とした場合、大規模な対訳コーパスを容易に入手できるとは限らないため、この問題は深刻である。したがって、静的な言語知識に強く依存することなく、かつ小規模な対訳コーパスからでも効率よく対訳知識を自動抽出可能な手法が必要となる。

我々は従来より、英語－日本語対訳コーパスを対象として、既存の対訳辞書や規則などの静的な言語知識に強く依存することなく、対訳知識を自動抽出するための学習手法［11,12,13］を提案している。これらの手法は、与えられた小規模な対訳コーパスのみから対訳知識を自動抽出する。しかし、これらの手法では、対訳知識を抽出するために表層レベルで類似した複数の対訳文が必要となる。これは、学習能力が不十分であり、データスパースネスの問題が十分に解消されるには至っていないことを意味する。

そこで、本論文において我々は、出現頻度の低い対訳語が多くを占める、小規模な対訳コーパスからでも対訳語の自動抽出が可能な新たな学習手法として、局所着目型学習（Local Focus-based Learning、略してLFL）を提案する。本手法は、対訳文の局所部分を対象に語の対応関係を決定するため探索範囲を限定できる。さらに、言語間のコロケーションを用いることで、抽出対象の対訳語の出現頻度に依存することなく対訳語を効率よく自動抽出できる。本論文で

は、このLFLをアイヌ語－日本語対訳コーパスに適用した。アイヌ語－日本語対訳コーパスを対象とした理由は２つある。一つは、様々な表現の訳語を有する対訳辞書という観点から、既存のアイヌ語－日本語辞書は改善の余地が多分にあると考えられるためである。また、アイヌ語－日本語対訳コーパスを大量に入手することは困難であるため、小規模な対訳コーパスから出現頻度の低い対訳語を自動抽出する手法が要求される。性能評価実験の結果、名詞および動詞対訳語の平均出現頻度が1.96である288文のアイヌ語－日本語対訳コーパスに対して、本手法により再現率54.0%、適合率60.8%の精度で名詞対訳語と動詞対訳語を抽出できることが確認された。再現率は統計的な手法に比べ10%以上高く、この結果は本手法の有効性を示すものである。本論文では、始めに、対訳語抽出のための新たな学習手法LFLの基本的な考え方について述べ、続いてアイヌ語の簡単な説明とLFLのアイヌ語－日本語対訳コーパスへの適用について述べる。最後に、性能評価実験の結果に基づき、本手法の有効性について述べる。

2　基本的な考え方

LFLは、出現頻度に強く依存することなく、対訳文の局所部分を対象に言語間のコロケーションを用いて対訳語を自動抽出する学習手法である。そして、これは、「対訳文の局所部分を対象として、部分的な対応関係が成立する場合、それらに隣接する部分もまた対応関係が成立する可能性が高い」という仮定に基づいている。この仮定を導入する際には、対訳文中の局所部分とその中の部分的な対応関係を決定する方法として、２つの処理を用いている。さらに、これらの処理は連動しているため、より効率よく対訳語を抽出することができる。本論文では、２つの処理をそれぞれ対訳文ペア方式、テンプレート方式と呼ぶ。

2.1　局所着目型学習（LFL）の概要

図１にLFLの概略図を示す。対訳文ペア方式は、対訳文同士の共通部分とそれに隣接する部分を局所部分として対応関係を決定する方式であり、n：m（n,mは自然数）の対訳語を抽出する。さらに、対訳文ペア方式では、コロケーションテンプレートを獲得する。コロケーションテンプレートとは、局所部分に関する言語間のコロケーション情報を有する、対訳語を自動抽出するためのルールである。テンプレート方式は、コロケーションテンプレートが適用可能な対訳文において、コロケーションテンプレートとの間の共通部分とそれに隣接する部分を局所部分として対応関係を決定する方式であり、１：１の対訳語を抽出する。

2.1.1　対訳文ペア方式

対訳文ペア方式による対訳語およびコロケーションテンプレートの抽出処理の詳細を述べ

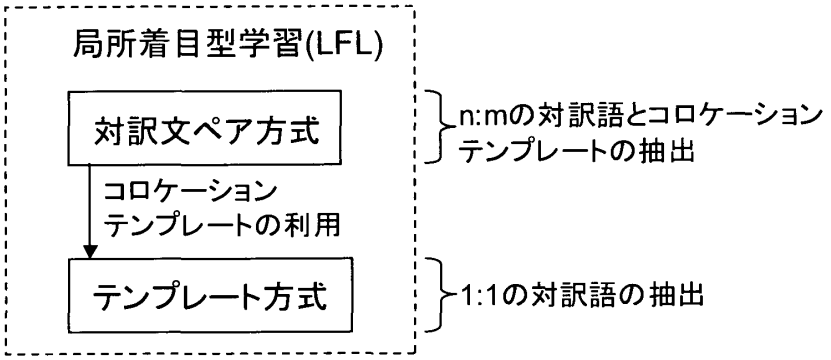

図1 局所着目型学習（LFL）の概略図

る。

(1)対訳コーパスから、原言語文、目的言語文の両方において、共通部分が存在する対訳文のペアを選択する。

(2)選択された対訳文のペアに対し、以下の処理のいずれかを行う。

(a)原言語文と目的言語文において、共通部分が2つ存在する場合、共通部分に挟まれた部分を原言語文と目的言語文から抽出する。

(b)原言語文もしくは目的言語文において、共通部分が1つ存在し、その共通部分の左側に隣接する語が文頭の1語の場合にはその1語を抽出する。そして、対応する原言語文もしくは目的言語文からは共通部分の左側に隣接する語から文頭の語までを抽出する。

(3)原言語文から抽出された部分と目的言語文から抽出された部分を組み合わせ対訳語を得る。

(4)対訳文中の抽出部分を変数に置き換える。

(5)原言語文と目的言語文のそれぞれにおいて、個々の共通部分と変数の組を抽出する。

(6)抽出された変数と共通部分の組において、原言語文から抽出された部分と目的言語文から抽出された部分を組み合わせ、コロケーションテンプレートを得る。

図2に、対訳文ペア方式とテンプレート方式の処理概要を示す。図2の対訳文ペア方式の処理概要において、対訳文1と対訳文2の原言語文では“Z*”と“Θ”、目的言語文では“δ”と“η θ”が共通部分である。これらの共通部分に挟まれた部分として、対訳文1の原言語文から“H”、目的言語文から“ε ζ”が抽出される。対訳文2では、原言語文から“Ξ”、目的言語文から“o”が抽出される。したがって、抽出された部分を組み合わせることによって、(H；ε ζ) と (Ξ；o) が対訳語として得られる。さらに、対訳文中の抽出部分を変数“@”に置き換える。そして、変数“@”と左側の共通部分、変数“@”と右側の共通部分の組を原言語文と目的言語文から抽出する。したがって、原言語文から“Z @”と“@ Θ”、目的言語文から“δ @”と“@ η θ”が抽出される。これらを組み合わせることで、コロケーションテンプレートとして (Z @；δ @)、(@ Θ；@ η θ)、(Z @；@ η θ)、(@ Θ；δ @) が得られる。また、本章では、対訳語とコロ

*ギリシャ文字の使用は、様々な言語を一般化した表現として用いている。

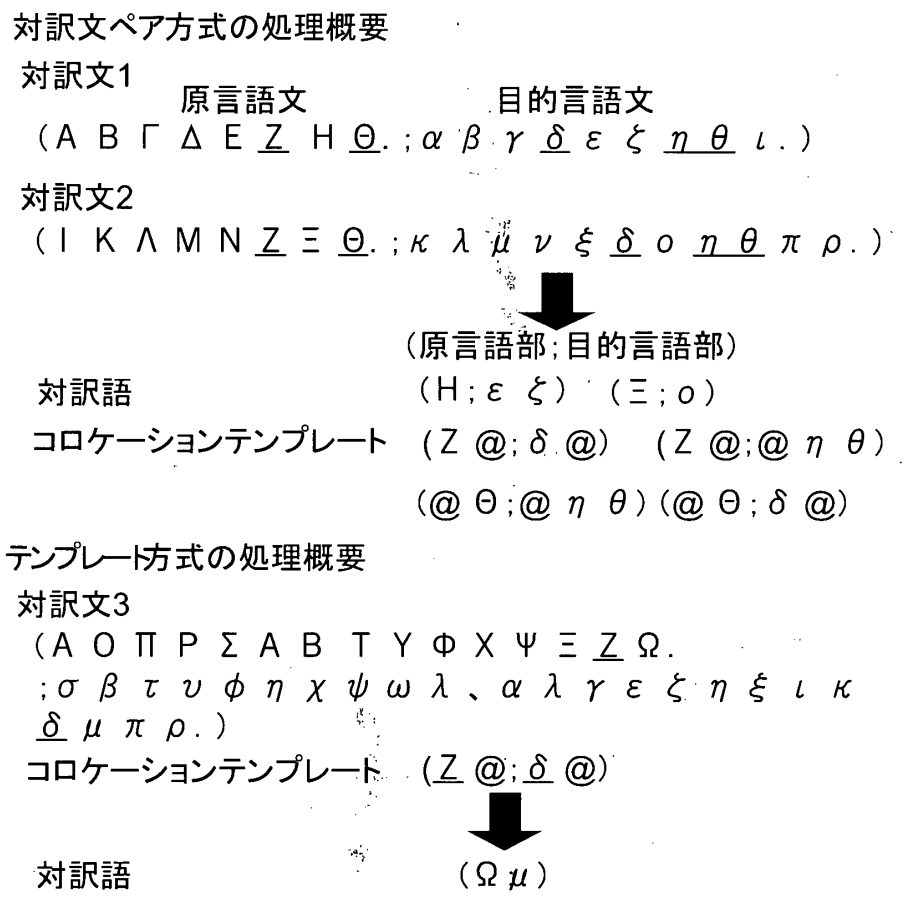

図２　局所着目型学習（LFL）の処理概要

ケーションテンプレートに対し、原言語文から抽出されたものを原言語部、目的言語文から抽出されたものを目的言語部と呼ぶ。

2.1.2　テンプレート方式

テンプレート方式による対訳語の抽出処理の詳細を述べる。

(1)コロケーションテンプレートの原言語部と目的言語部共に変数以外の全ての部分を共通部分に持つ対訳文を選択する。

(2)選択された対訳文の原言語文と目的言語文のそれぞれに対し、以下の処理のいずれかを行う。

(a)コロケーションテンプレートの原言語部もしくは目的言語部において左端に変数が存在する場合、対訳文の原言語文もしくは目的言語文から共通部分の左側に隣接する１語を抽出する。

(b)コロケーションテンプレートの原言語部もしくは目的言語部において右端に変数が存在する場合、対訳文の原言語文もしくは目的言語文から共通部分の右側に隣接する１語を抽出する。

(3)原言語文から抽出された1語と目的言語文から抽出された1語を組み合わせ、対訳語を得る。

図2のテンプレート方式の処理概要では、対訳文ペア方式より獲得された4つのコロケーションテンプレートを対象に、対訳文3に対して適用可能なコロケーションテンプレートを選択する。この場合、(Z @；δ @) の原言語部と目的言語部の変数を除いた部分"Z"と"δ"の両方が共に対訳文3に出現するため、(Z @；δ @) が適用可能なコロケーションテンプレートとして選択される。そして、コロケーションテンプレート (Z @；δ @) は、原言語部、目的言語部共に変数"@"が右端に存在するため、対訳文3の原言語文、目的言語文のそれぞれから共通部分の右側に隣接する1語を抽出し、それらを組み合わせる。その結果、(Ω；μ) が対訳語として得られる。

2.2 対訳語抽出の仮定に対する予備実験

先に述べた「対訳文の局所部分を対象として、部分的な対応関係が成立する場合、それらに隣接する部分もまた対応関係が成立する可能性が高い」という仮定の信頼性を確認するために予備実験を行った。予備実験は、後述する4章のシステム構成に基づき構築したシステムに、アイヌ語－日本語対訳文40文 [25] を与え、対訳語を抽出した。そして、対訳文ペア方式に基づき抽出された対訳語においては、原言語文の共通部分と目的言語文の共通部分の対応関係が正しい場合、どれだけの精度で対応関係の正しい対訳語を抽出できたのかを求めた。また、テンプレート方式に基づき抽出された対訳語においては、コロケーションテンプレートの原言語部と目的言語部の対応関係が正しい場合、どれだけの精度で対応関係の正しい対訳語を抽出できたのかを調べた。その結果、正しい対訳語の抽出精度は56.0%となり、仮定の信頼性は比較的高いことが確認できた。

3 アイヌ語の概略

本節では、アイヌ語について表記、語順、品詞の観点から簡単な説明を行う [24,25,26]。

3.1 表記について

表記については、元来アイヌ語は文字を持たない言語であるが、カタカナやローマ字による文章化が行われている。おおよそ、ローマ字は語を表記し、カタカナは発音を表す読み仮名として用いられる。文章中の語は、単語を単位として分かち書きされる。また、ローマ字表記上のきまりとしては以下のものがある。

(1)固有名詞のみ大文字で始める。他は、たとえ文頭でも小文字を用いる。

(2)ハイフン「-」は、一語であっても二語のアクセントで発音される語に対して用いられる。

例：umma-ru＜馬道＞

(3)人称接辞と語幹との間には「＝」を入れる。これは表記上のきまりであって、発音の切れ目ではない[†]。 例：ku=tekehe＜私の手＞

3.2 語順について

アイヌ語の語順は、基本的な文型がSOV（主語＋目的語＋動詞）という点で日本語と同じであるが、形態論的には膠着語である日本語とは異なり、抱合語[‡]に属するとされている。図3にアイヌ語の肯定文、否定文、疑問文の具体例を示す。否定文は日本語と異なり否定を表す語を動詞の前に置く。図3では“somo”という副詞が否定を表す語であり、動詞“ki”に係っている。また、疑問文であることを明示する場合には、疑問を表す語を動詞句の後に置く。図3では“ya”という終助詞が疑問を表す語である。

肯定文
アイヌ語文: ku= pon　hi　ta ramma　ku= siyeye.
　　単語訳 [私・小さい　とき　に　　よく　　私・病気する]
日本語文:　私は小さい時、よく病気しました.

否定文
アイヌ語文: k= eramasu ka somo　ki.
　　単語訳 [私・～を好む　　も　ない　　～する]
日本語文:　私は好きではありません.

疑問文
アイヌ語文: tan cep hunna koyki　ya?
　　単語訳 [この　魚　　誰　　～をとる　か]
日本語文:　この魚は誰が獲ったのか？

図3　アイヌ語文の具体例

3.3 品詞について

アイヌ語には、日本語と同様の品詞として、動詞、名詞、副詞、助詞、助動詞、接続詞、連体詞が存在するが、その詳細は必ずしも一致していない。例えば、アイヌ語の動詞には時制による語形変化や語尾変化は無いが、単数、複数を区別するものがある。名詞においては、所属形と概念形では語尾が異なる。また、形容詞はアイヌ語では動詞に分類されるため、アイヌ語の品詞分類に形容詞は存在しない。

[†]本論文では、辞書への登録時の便宜上、あらかじめ「＝」の後にスペースを入れている。
[‡]抱合語とは、動詞を中心に、その前後に人称接辞や目的語が結合または挿入されて、一語が文のような形態をとる言語である（広辞苑）。

4 システム構成

与えられたアイヌ語－日本語対訳コーパスから対訳語を自動抽出する、LFLに基づくシステムの構成図を図4に示す。本手法は、語を最小単位として処理するため、日本語のような膠着語に適用する場合、分かち書きされた文に変換する必要がある。したがって、日本語文に対しては日本語形態素解析ツール「ChaSen[§]」を用いて、分かち書きを行っている。図4の対訳知識抽出部では、対訳文に対し、対訳文ペア方式とテンプレート方式からなるLFLを用いて対訳語とコロケーションテンプレートを自動抽出する。対訳知識評価部では対訳文を参照することで、対訳知識抽出部で抽出された対訳語とコロケーションテンプレートの対応関係の正誤を評価する。そして、その結果得られた評価値を全ての対訳語とコロケーションテンプレートに付与し、辞書に登録する。

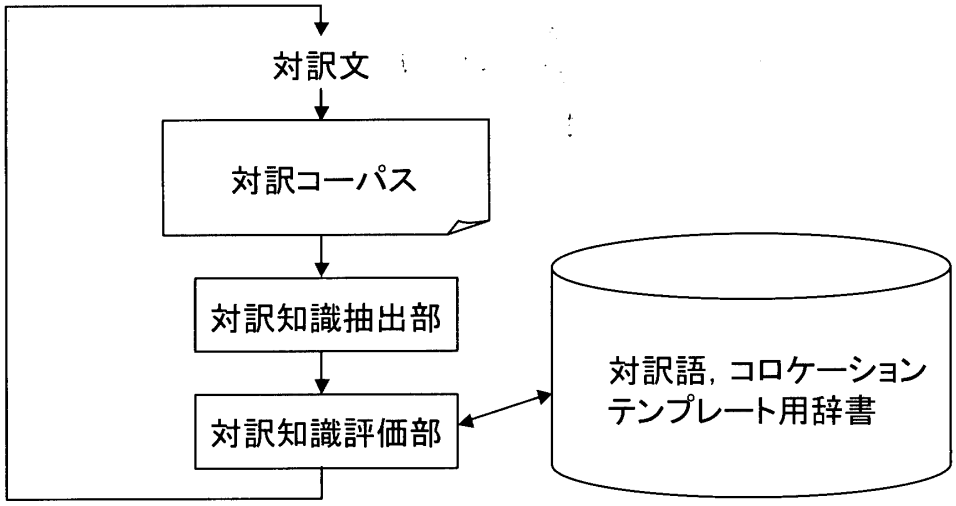

図4 システム構成図

5 処理過程

5.1 対訳知識抽出部

5.1.1 対訳文ペア方式によるアイヌ語－日本語対訳語とコロケーションテンプレートの抽出

対訳文ペア方式では、アイヌ語－日本語対訳文のペアにおいて、共通部分とそれに隣接する部分を局所部分として、そこからn：mの対訳語を抽出する。さらに、対訳文ペア方式では、テンプレート方式で使用するコロケーションテンプレートを自動獲得する。また、本論文では、以後、抽出された対訳語とコロケーションテンプレートに対し、アイヌ語文から抽出された部分をアイヌ語部、日本語文から抽出された部分を日本語部と呼ぶ。以下に対訳文ペア方式

[§] http : //chasen.aist-nara.ac.jp/

によるアイヌ語－日本語間の名詞および動詞対訳語とコロケーションテンプレートの抽出処理の詳細を述べる。

(1)アイヌ語文間、日本語文間共に共通部分の存在する対訳文ペアを選択する。

(2)選択された対訳文ペアに対し、以下の処理のいずれかを行う。

(a) アイヌ語文間と日本語文間において、共通部分が２つ存在する場合、共通部分に挟まれた部分をアイヌ語文と日本語文から抽出する。そして、抽出された日本語部分においては末尾の語が名詞、動詞、助動詞、形容詞かどうかを調べる。日本語文から抽出された部分が１語の場合には、名詞、動詞、形容詞かどうかを調べる。

(b) アイヌ語文間もしくは日本語文間において、共通部分が１つ存在し、その共通部分の左側に隣接する語が文頭の１語の場合にはその１語を抽出する。そして、対応するアイヌ語文もしくは日本語文からは共通部分の左側に隣接する語から文頭の語までを抽出する。抽出された日本語部分においては末尾の語が名詞、動詞、助動詞、形容詞かどうかを調べる。日本語文から抽出された部分が１語の場合には、名詞、動詞、形容詞かどうかを調べる。

(3)日本語文から抽出された部分が処理(2)の品詞に対する条件を満たす場合、アイヌ語文から抽出された部分と組み合わせ、名詞および動詞対訳語とする。なお、動詞対訳語の日本語部の動詞に対しては、基本形を用いる。

(4)対訳文中の抽出部分を変数“@”に置き換える。

(5)アイヌ語文と日本語文のそれぞれにおいて、個々の共通部分と変数の組を抽出する。

(6)抽出された変数と共通部分の組において、アイヌ語文から抽出されたものと日本語文から抽出されたものを組み合わせ、コロケーションテンプレートを得る。

対訳文ペア方式によるアイヌ語－日本語の対訳語とコロケーションテンプレートの抽出処理の具体例を図５に示す。対訳文１と対訳文２において、アイヌ語文では“ani”と“kar”、日本語

(1)対訳語の抽出

対訳文1

(ku= kor　totto　poro su ani **sayo** kar.

単語訳:[私・～の　おかあさん　大きな　鍋　で　お粥　～を作る]

;母/が/大鍋/で/お/粥/を/作り/ます.)

対訳文2

(k= onaha anakne sipe kap ani **ker** kar.

単語訳:[私・～の父　は　サケ　皮　で　靴　～を作る]

;父/は/サケ/の/皮/で/靴/を/作り/まし/た.)

⬇

名詞対訳語: (ker;靴) (sayo;お/粥)

(2)コロケーションテンプレートの獲得
対訳文1
(ku= kor totto poro su ani @ kar.
;母/が/大鍋/で/@/を/作り/ます.)
対訳文2
(k= onaha anakne sipe kap ani @ kar.
;父/は/サケ/の/皮/で/@/を/作り/ました.)
⬇
コロケーションテンプレート:
(ani @;で/@)　(ani @;@/を/作り)
(@ kar;@/を/作り)　(@ kar;で/@)

図5　対訳文ペア方式によるアイヌ語—日本語対訳語とコロケーションテンプレートの抽出例

文では“で”と“を/作り”が共通部分である。したがって、共通部分に挟まれた部分として、アイヌ語文からは“sayo”と“ker”、日本語文からは“お/粥”と“靴”が抽出される。日本語文から抽出された2つの部分において、末尾の語の品詞を調べる。この場合、“粥”、“靴”共に名詞であるため、名詞対訳語として（sayo；お/粥）と（ker；靴）が抽出される。さらに、コロケーションテンプレートを獲得する。対訳文1と対訳文2に対し、抽出された（sayo；お/粥）と（ker；靴）の部分を変数“@”に置き換える。そして、変数と左右それぞれの共通部分との組を抽出する。図5ではアイヌ語文からは“ani @”と“@ kar”が、日本語文からは“で/@”と“@/を/作り”が抽出される。したがって、これらを組み合わせることにより、コロケーションテンプレートとして（ani @；で/@）、（@ kar；@/を/作り）、（ani @；@/を/作り）、（@ kar；で/@）が得られる。コロケーションテンプレートの獲得処理では、（ani @；@/を/作り）と（@ kar；で/@）のように、アイヌ語部と日本語部の対応関係が誤ったコロケーションテンプレートも獲得されることになるが、このようなコロケーションテンプレートが適用される可能性は低いと考えられる。

5.1.2　テンプレート方式によるアイヌ語ー日本語対訳語の抽出

テンプレート方式では、コロケーションテンプレートが適用可能なアイヌ語ー日本語対訳文において、コロケーションテンプレートとの間の共通部分とそれに隣接する部分を局所部分として、そこから1：1の対訳語を抽出する。さらに、日本語品詞情報を使用することで、日本語文からは、名詞句と動詞句の抽出も行う。すなわち、1：nの対訳語を抽出する。以下にテンプレート方式によるアイヌ語ー日本語間の名詞および動詞対訳語の抽出処理の詳細を述べる。

(1)コロケーションテンプレートのアイヌ語部と日本語部共に、変数以外の全ての部分を共通部分に持つ対訳文を選択する。

(2)選択された対訳文のアイヌ語文と日本語文のそれぞれに対し、以下の処理のいずれかを行

う。

(a) コロケーションテンプレートのアイヌ語部もしくは日本語部において、左端に変数が存在する場合、アイヌ語文と日本語文に対して、共通部分の左側に隣接する1語を抽出する。さらに、日本語文に対しては、抽出された1語を起点として、動詞、形容詞、名詞で構成される名詞句、または名詞、動詞、助動詞で構成される動詞句を1部分として抽出する。

(b) コロケーションテンプレートのアイヌ語部もしくは日本語部において、右端に変数が存在する場合、アイヌ語文と日本語文に対して、共通部分の右側に隣接する1語を抽出する。さらに、日本語文に対しては、抽出された1語を起点として、動詞、形容詞、名詞で構成される名詞句、または名詞、動詞、助動詞で構成される動詞句を1部分として抽出する。

(3)アイヌ語文から抽出された部分と日本語文から抽出された部分を組み合わせ名詞および動詞対訳語を得る。なお、動詞対訳語の日本語部の動詞に対しては、基本形を用いる。

図6にテンプレート方式による対訳語の抽出処理の具体例を示す。図5で獲得されたコロケーションテンプレートを対象に、対訳文に対して適用可能なものを選択する。この場合、“ani”と“で”の両方が対訳文に出現する（ani @；で/@）が選ばれる。コロケーションテンプレート（ani @；で/@）はアイヌ語部、日本語部共に右端に変数が存在するため、対訳文のアイヌ語文からは共通部分“ani”の右側に隣接する1語“nuyanuya”を抽出する。日本語文からは共通部分“で”の右側に隣接する1語“揉み”が動詞であるため、これを抽出する。さらに、1部分としては動詞と助動詞が連続して出現する“揉みました”を抽出する。1語の抽出で得られた“揉み”と1部分の抽出で得られた“揉みました”のいずれも、動詞の基本形としては“揉む”のみが得られる。したがって、動詞対訳語として（nuyanuya；揉む）が抽出される。

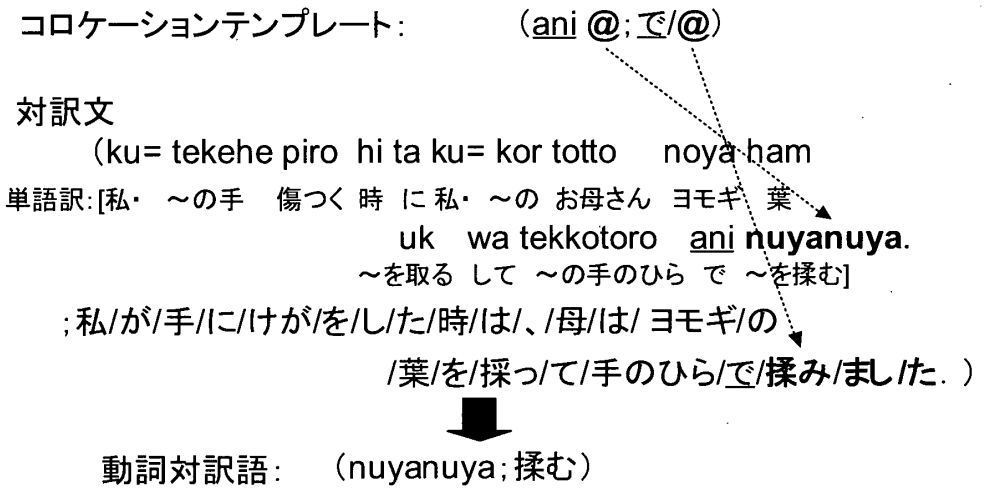

図6　テンプレート方式によるアイヌ語―日本語対訳語の抽出例

5.2 対訳知識評価部

対訳知識評価部では、対訳知識抽出部で抽出された対訳語とコロケーションテンプレートに対し、アイヌ語部と日本語部の対応関係の正誤を評価する。処理の詳細を以下に記す。

(1)抽出もとの対訳文以外の全対訳文を対象に、抽出された対訳語のアイヌ語部と日本語部の両方が同時に出現する頻度を求め、その値を正度数CF（Correct Frequency）とする。

(2)抽出もとの対訳文以外の全対訳文を対象に、抽出された対訳語のアイヌ語部、日本語部のいずれか一方のみが出現する頻度を求め、その値を誤度数EF（Erroneous Frequency）とする。

(3)正度数CFと誤度数EFに基づき、以下の式より評価値EV（Evaluation Value）を求める。

$$EV=CF-EF \tag{1}$$

この評価方法により、例えば、正しい名詞対訳語として（coka；私/たち）と誤った名詞対訳語として（coka；たち）が存在する場合、（coka；私/たち）のEVが（coka；たち）のEVに比べ大きくなる。正しい名詞対訳語（coka；私/たち）では、“coka”と“私/たち”が対訳文のアイヌ語文と日本語文に同時に出現する可能性は非常に高いため、CFは増加する。それに対し、誤った名詞対訳語（coka；たち）では、日本語部“たち”は“私”以外にも“子供/たち”、“妹/たち”など様々な名詞句に含まれる。したがって、対訳文の日本語文に“たち”が存在するのに対し、対応するアイヌ語文に“coka”が存在する可能性は低く、（coka；たち）のEFは増加する。その結果、（coka；私/たち）のEVは（coka；たち）のEVよりも大きくなる。

また、コロケーションテンプレートにおいてもEVを求める。その場合には、変数を除いた部分に対して上述した評価処理を行う。これは、同じEVを持つ対訳語が競合した場合に、EVの高いコロケーションテンプレートを用いて抽出された対訳語、もしくは、EVの高いコロケーションテンプレートと同時に抽出された対訳語を優先的に選択するためである。

6 性能評価実験

6.1 実験方法

実験データには、２つの文献［25］［26］に記載されているアイヌ語－日本語対訳文288文を用いた。これらは全10話からなる物語の対訳文である。アイヌ語文の平均語数は11.8である。また、対訳文288文中には546種類の名詞および動詞対訳語が存在する。その内訳は名詞が278個、動詞が268個である。また、これらの対訳語の出現頻度の平均は1.96である。実験は、４章と５章で述べたシステム構成と処理過程に基づき構築したシステムを用いて行った。なお、辞書の初期状態は空とした。これは、LFLが静的な言語知識に強く依存することなく、学習能力に基づき対訳語を自動抽出可能なことを確認するためである。

6.2　評価方法

評価はアイヌ語の名詞、動詞を１種類ずつ指定し、それと同じアイヌ語部を持つ対訳語を選択する。そして、選択された対訳語の日本語部を訳語として評価した。同一のアイヌ語部を持つ対訳語が複数競合した場合には、5.2の対訳知識評価部で述べたように、EVが最も高い対訳語を選択する。さらに、同じEVを持つ対訳語が複数競合した場合には、EVの高いコロケーションテンプレートから抽出された対訳語またはEVの高いコロケーションテンプレートと同時に抽出された対訳語を選択する。

6.3　実験結果

546種類の名詞、動詞対訳語に対する再現率と適合率を求めた。再現率は546種類の対訳語に対する、正しい対訳語の割合である。適合率は546種類の対訳語において対訳語が得られたものに対する、正しい対訳語の割合である。実験の結果、再現率として54.0%、適合率として60.8%が得られた。表１に本手法より抽出された正しい対訳語の具体例を示す。表１の既存辞書の訳語とは、アイヌ語－日本語辞書［24］、もしくは文献［25,26］の単語索引に記載されている訳語である。それに対し、本手法による訳語は、既存辞書［24,25,26］に記載されていない訳語である。また、図７には、対訳コーパスのサイズを50文ずつ増加した場合の再現率、適合率の推移を示す。

表１　抽出された正しい対訳語の例

名詞対訳語の具体例		
アイヌ語	既存辞書の訳語	本手法による訳語
aep	食べ物	料理
surku kina	トリカブト	毒草
ican	サケ，マスの産卵穴	産卵場所
uepeker	散文説話	昔話
動詞対訳語の具体例		
アイヌ語	既存辞書の訳語	本手法による訳語
popte	煮立てる	煎じる
isam	存在しない	消える
kisam	つかむ	持つ
sesekka	熱くする	あぶる

6.4　考察

6.4.1　本手法の特徴

本手法の大きな利点は、出現頻度の低い対訳語が数多く存在する小規模な対訳コーパスか

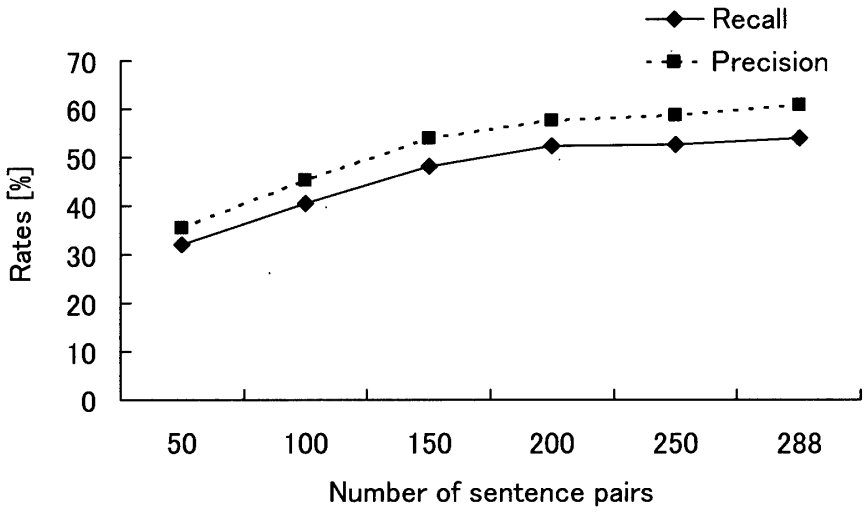

図7 対訳コーパスのサイズと精度の関係

ら、効率よく対訳語を自動抽出できることにある。今回使用したアイヌ語－日本語対訳コーパスでは、抽出対象となった546種類の対訳語の出現頻度ごとの割合は、出現頻度1の対訳語が64.3%、出現頻度2の対訳語が17.6%、出現頻度3以上の対訳語が18.1%であり、多くが出現頻度1の対訳語であった。図8に、本手法による低頻度の対訳語の抽出例を示す。

処理1:

対訳文1

(**onne kur** kor cise hankeno ・・・

単語訳:[年取った 人 ～の 家 近くに ・・・]

;老人/の/家/が/近く/に/・・・)

対訳文2

(ku= kor huci anakne yuptek wa ・・・

単語訳:[私・～の おばあさん は 働き者である して ・・・]

;私/の/祖母/は/働き者/で/・・・)

⬇

名詞対訳語: (onne kur;老人)

処理2:

コロケーションテンプレート (@ hawe;@/声)

対訳文3

(anun ne yakun **iruska** hawe ani emikmik.

単語訳:[よそもの ～である ならば 怒る 声 で ～に吠える]

;見知らぬ/人/に/は/怒っ/た/声/で/吠え/立て/ます.)

⬇

動詞対訳語: (iruska;怒る)

図8 低頻度の対訳語の抽出例

図８の処理１は、出現頻度が２の名詞対訳語（onne kur；老人）の抽出処理を示している。この抽出処理は、5.1.1の(2)の（b）に基づいている。対訳文１の日本語文において、共通部分“の”の左側に隣接する１語の名詞“老人”は文頭に位置するため、対訳文１のアイヌ語文に対して共通部分“kor”の左側に隣接する語から文頭の語までの“onne kur”を抽出する。その結果、“onne kur”と“老人”の組み合わせにより名詞対訳語（onne kur；老人）が得られる。また、図８の処理２は、出現頻度１の動詞対訳語（iruska；怒る）の抽出処理である。この抽出処理は5.1.2の(2)の（a）に基づいている。コロケーションテンプレート（@ hawe；@/声）の“hawe”と“声”の両方が対訳文３に出現し、変数“@”がアイヌ語部、日本語部共に左端に存在するため、対訳文３のアイヌ語文からは共通部分“hawe”の左側の１語“iruska”を抽出する。日本語文においては、共通部分“声”の左側の１語“た”は助動詞であるため抽出しない。日本語文からの１部分の抽出処理では、共通部分“声”を起点として、逐次左側の語の品詞を調べる。この場合、助動詞“た”の左側の語“怒っ”は動詞、その左側の語“は”は助詞であるため、助詞“は”が出現するまでの“怒っ/た”を１部分として抽出する。この１部分から、動詞の基本形として“怒る”が得られるため、“iruska”との組み合わせにより動詞対訳語（iruska；怒る）が抽出される。

このように、本手法では、出現頻度が１や２である低頻度の対訳語の抽出が可能である。本手法は、他のアプローチとは異なり、抽出対象の対訳語そのものの出現頻度が低い場合でも、抽出対象の対訳語と共通部分で構成される局所部分を対象として、言語間のコロケーションを有効利用することで対訳語を抽出できる。今回の実験では、アイヌ語－日本語対訳コーパスを対象に対訳語の抽出を行ったが、英語のような国際性の高い言語を用いた対訳コーパスにおいても低頻度語が多くを占める状況は存在する［14］。本手法は、特定の言語に強く依存することなく、かつ出現頻度の低い対訳語を抽出することができるため、様々な言語間の対訳コーパスに対して有効であると考えられる。

また、今回の実験で抽出された正しい対訳語295個において、対訳文ペア方式より抽出された対訳語は50個、テンプレート方式より抽出された対訳語は245個であった。コロケーションテンプレートは適用容易性が高く、かつ、１：１の対訳語を抽出するためのルールである。したがって、今回使用した対訳コーパスのように１：１の対訳語が全体の83.5％を占めている場合、テンプレート方式は有効である。

6.4.2　抽出された対訳語の種類について

本手法では、言語間のコロケーションを利用することにより、様々な構成単語数の対訳語を抽出することができる。表２に構成単語数ごとの対訳語の再現率を示す。

表２より、様々な対訳語を抽出できたことがわかる。しかし、１：nの対訳語の再現率が非

表2 語数別の再現率

対訳語の構成（アイヌ語部：日本語部）	再現率	抽出対象の対訳語の数
1：1	59.9%	456
1：n	23.2%	82
n：1	50.0%	4
n：m	33.3%	3

常に低い。1：nの対訳語について、その詳細を調査すると、抽出対象の名詞対訳語28個に対する再現率は53.6%であるのに対し、動詞対訳語54個に対する再現率は7.4%と非常に低かった。動詞対訳語は、アイヌ語、日本語共に文末に位置することが多く、その場合、2つの共通部分で挟まれる可能性は低い。したがって、対訳文ペア方式では抽出が困難になる。一方、テンプレート方式では、1：nの動詞対訳語の約6割は（wentarap；夢/を/見る）のように日本語部に助詞が含まれる動詞句である。5.1.2の日本語文からの1部分の抽出処理では、助詞が区切りとなるため“夢”もしくは“見る”のみが抽出されてしまう。このような問題を解決するためには、既に抽出されたEVの高い対訳語から、段階的に対応関係を決定していくことが有効と考えられる。

6.4.3 新たな訳語の抽出

本手法では、表1の正しい対訳語の具体例で示したように既存辞書の訳語に対し、より自然な訳語もしくは新たな表現の訳語を抽出することができた。今回使用した対訳コーパスに存在する546種類の名詞、動詞対訳語において、76個の日本語訳は、アイヌ語－日本語辞書 [24] および文献 [25][26] の単語索引のいずれにも記載されていない、対訳コーパスのみに存在する日本語訳であった。この76個の訳語に対する再現率は52.6%、適合率は62.5%である。この値は十分とはいえないが、出現頻度の低い対訳語の抽出を前提としていることや人手でこれらの対訳語を抽出するという作業を削減できることを考えると本システムは有意義なものといえる。現在のアイヌ語－日本語辞書は、様々な表現の訳語の網羅という点では十分とはいえず、かつ大規模な対訳コーパスを容易に入手することが困難である。そのような状況下で、小規模な対訳コーパスから対訳語を自動抽出可能な本手法は有効である。

6.4.4 多義語について

本システムでは複数の訳語が抽出された場合、5.2の処理に基づき決定されたEVが最も高い訳語のみを一意に選択する。そのため、多義語については実験結果に反映されていないが、複数の正しい訳語が得られた場合、他の正しい訳語をシステムがどのように評価したのかについ

て調査した。今回の実験では、正しい対訳語295個に対し、18個が多義語として抽出されていた。すなわち、これらの正しい対訳語18個においては、EVに基づき順位付けした結果、上位1位だけでなく2位以下にも正しい訳語が存在していた。そして、この18個の多義語に対して、2位以下に存在していた正しい訳語の6割が誤った訳語を挟まずに連続して出現していた。しかし、現システムでは、訳語の正誤を自動的には判定できないため、多義語の判定もできない。今後、多義語を抽出するための手法を取り入れたいと考えている。

6.4.5　本手法の問題点

6.4.2で、日本語部に助詞が含まれる動詞対訳語の抽出が困難であることは述べたが、他の問題点としては誤った対訳語の抽出が挙げられる。今回の実験では、対訳知識評価部の処理に基づき選択された対訳語だけではなく、抽出された全対訳語を対象とした場合の正しい対訳語の割合は17.6%と非常に低かった。誤った対訳語はその抽出過程の観点からいくつかに分類される。誤った対訳語が最も多く抽出されたのは、対応関係が成立していない共通部分に隣接する語を抽出した場合であった。その他には、共通部分の対応関係が成立していても、その共通部分が1つの対訳文中に複数存在したため多義性が生じ、誤った対訳語を抽出するケースである。誤った対訳語抽出の具体例を図9に示す。

図9では、対訳文に対して、コロケーションテンプレート（ani @；で/@）が適用される。しかし、アイヌ語文中に“ani”が、日本語文中に“で”がそれぞれ2箇所存在する。したがって、アイヌ語文からは“sipe”と“sapa”、日本語文からは“サケ”と“頭”が抽出される。その結果、正しい対訳語（sipe；サケ）と（sapa；頭）だけでなく、誤った対訳語（sipe；頭）と（sapa；サケ）も抽出される。このように本手法では、誤った対訳語の抽出も行われるが、5.2

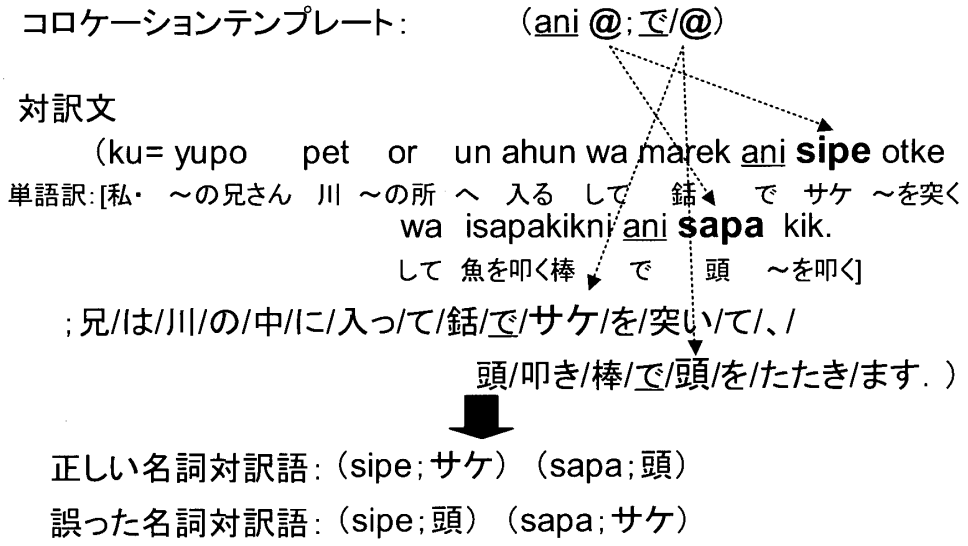

図9　誤ったの対訳語の抽出例

の対訳知識評価部の処理より、正しい対訳語を比較的高い精度で選択することが可能である。今回の実験では、正しい対訳語と誤った対訳語が競合した場合の正しい対訳語の選択精度は79.1%であった。さらに、ヒューリスティックスを導入することで、誤った対訳語の抽出を防ぐことが可能である［15］。

6.4.6 他手法との比較

アイヌ語－日本語間の対訳辞書を自動構築する場合、現時点では解析的なアプローチ[2,3]の利用は困難である。アイヌ語については、形態素解析ツールや構文解析ツールが存在しないため、アイヌ語文に対して解析的な知識を容易に付与することができない。また、アイヌ語－日本語間の電子化された対訳辞書をはじめ、アイヌ語とそれ以外の言語間の電子化された対訳辞書も容易には入手することができない。したがって、解析的なアプローチによるアイヌ語－日本語対訳語の自動抽出は現時点では困難と考えられる。また、今後、アイヌ語の対訳辞書や様々な解析ツールが開発された場合でも、静的な言語知識に強く依存する解析的なアプローチでは、多様な言語現象に対処することが困難となり、汎用性の低さが問題になると考えられる。それに対し、本手法は静的な言語知識に強く依存することなく、学習能力により対訳コーパスから対訳語を自動抽出するため、言語資源や解析ツールが整備された状況においても有効であると考えられる。

一方、統計的なアプローチ［4,5,6,7,8,9,10］には利用可能なものもある。そこで、代表的な統計に基づく手法であるBrownらの手法［4］を用い、本手法との比較実験を行った。Brownらは原言語と目的言語の単語間の相互情報量に基づき対応関係を決定する手法を提案している。以下に、その処理過程を示す。

(1)任意のアイヌ語の単語aと日本語の単語j_iの対応関係の強さは相互情報量Iとして以下の式(2)より定義される。そして、相互情報量Iの値を最大にするj_iをaの訳語とする。

$$I(j_i,a)=\log\frac{P(j_i \mid a)}{P(j_i)} \tag{2}$$

(2)$P(j_i \mid a)$は任意のアイヌ語単語aが日本語単語j_iに訳される確率であり、以下の式(3)で定義される。

$$P(j_i \mid a)=\frac{C(j_i,a)}{M(a)} \tag{3}$$

ここで、$M(a)$はアイヌ語単語aが全アイヌ語文に出現する頻度である。また、$C(j_i,a)$は以下の手順で求める。

(a)アイヌ語単語aと日本語の全ての単語j_iに対し、$C(j_i,a)$に0を設定する。

(b)アイヌ語単語aが出現するアイヌ語文に対応する日本語文がn個の単語からなる$J=j_{i1}$、

j_{i2}、・・・j_{in}である時、$C(j_{i1},a)$、$C(j_{i2},a)$、・・・$C(j_{in},a)$を$\frac{1}{n}$増やす。

(3)$P(j_i)$はランダムに選択されたアイヌ語単語aが日本語単語j_iに訳される確率であり、以下のように定義される。

$$P(j_i)=\sum_a \frac{P(j_i \mid a)}{P(a)}=\sum_a \frac{P(j_i \mid a)M(a)}{Ma} \tag{4}$$

ここで、Maは対訳コーパスのアイヌ語文の全単語数である。

また、LFLに基づくシステムと同様、日本語文については品詞情報を用いて、名詞、動詞、形容詞かを判定し、それ以外の品詞の語は抽出しない。任意のアイヌ語の単語に同じ相互情報量を持つ日本語訳が複数競合した場合には、対訳コーパスの日本語文において最も先に出現するものを選択する。

さらに、5.2で述べた対訳知識評価部の処理もまた統計的手法と位置づけることができるため、４章のシステム構成から対訳知識抽出部を除いた対訳知識評価部のみのシステムを構築し実験を行った。対訳知識評価部では、対訳語のアイヌ語部と日本語部の両方が同時に対訳文中に出現する場合に正の評価を与え、対訳文のアイヌ語文、日本語文のいずれかのみに出現する場合には負の評価を与える。その結果、誤った対訳語に対する評価が可能となる。このように、対訳知識評価部は、共起に基づき対訳語のアイヌ語部と日本語部の対応関係を評価することから、本論文では、この処理を共起に基づく手法と呼ぶこととする。共起に基づく手法では、Brownらの手法と同様に、対訳文中のアイヌ語単語と日本語単語の全ての組み合わせを対象として評価する。その結果、最も高いEVを持つ対訳語の日本語部を訳語として選択する。同じEVを持つ対訳語が複数競合した場合には、対訳語の日本語部が対訳コーパスの日本語文に対し最も先に出現するものを選択する。表３にBrownらの手法、共起に基づく手法、本手法のそれぞれの再現率を示す。

表３より、２つの統計的な手法の再現率が共に50.0%以下であるのに対し、本手法は50.0%を超える再現率が得られた。統計的手法では、対訳文中の全ての語の組み合わせを用いて対応関係を調べるため多義性が高くなる。さらに、出現頻度の低い対訳語が数多く存在する場合、それらを差別化するための情報が少ないため多義性を解消できずに決定不能状態［14］に陥ってしまう。

このような統計的な手法に対し、本手法は対訳文中の局所部分を対象とすることで探索範囲

表３　比較実験の結果

手法	再現率	抽出された正しい対訳語の数
Brownらの手法	26.4%	144
共起に基づく手法	43.6%	238
本手本	54.0%	295

を限定できる。さらに、対訳文のアイヌ語文と日本語文中の全ての語の組み合わせを対象に対応関係を決定するのではなく、言語間のコロケーションを利用することにより、組み合わせの多義性を抑えることができる。図10に3つの手法それぞれの出力例を示す。図10は、アイヌ語“nisatta”に対する日本語訳の出力例である。本手法では、訳語候補が少なく、その結果、容易に差別化が可能となり正しい訳語“明日”を抽出できた。それに対し、Brownらの手法では“祖母”、共起に基づく手法では“今日”といった誤った訳語が抽出された。

アイヌ語:nisatta

Brownらの手法		共起に基づく手法		本手法	
訳語	相互情報量	訳語	EV	訳語	EV
祖母	**5.562750**	**今日**	**0**	**明日**	**0**
今日	5.562750	明日	0	考え	-4
食べる	5.562750	何とか	0	祖母	-22
もの	5.562750	考え	-4	もの	-24
明日	5.562750	食べる	-6		
	・		・		
	・		・		
省略		省略			
	・		・		
	・		・		

図10 訳語の出力例

また、様々な解析ツールの使用を前提とした統計的な手法との比較について述べる。佐藤ら[10]はサポートベクタマシンを用いて対訳表現を抽出する手法を提案している。そこでは、日英の訓練コーパス4,000文、テストコーパス1,000文を用いて低頻度の句の対訳対を再現率77.6%、適合率80.8%の精度で抽出できたことが示されている。しかし、文献[10]中の対訳文の数と精度の関係において、訓練コーパス400文を用いた実験では、再現率が50%以下、適合率が60.0%以下となっている。それに対し、本手法では対訳文288文から再現率54.0%、適合率60.8%が得られた。対象言語や抽出単位が異なるため単純に比較することはできないが、本手法では小規模な対訳コーパスから非常に効果よく対訳語を抽出できていることがわかる。また、本手法は、統計的なモデルを学習する手法に対し、訓練コーパスとテストコーパスを区別することなく、与えられた全ての対訳コーパスを対象として対訳語を自動抽出することが可能である。

さらに、類似研究として言語間のコロケーション情報を持つテンプレートを獲得する手法[16,17,18,19]、共起情報に基づく手法[20]などが挙げられるが、これらの手法は大規模な対訳コーパスもしくは表層レベルで類似した対訳文が数多く存在する対訳コーパスが不可欠で

ある。また、獲得されるルールは対訳文の局所部分を対象としたものではないため、対訳語の自動抽出を目的に適用した場合、その適用は限定されると考えられる。

一方、対応付けされていないコーパスから対訳語を抽出する研究 [21,22,23] も提案されているが、それらは共起情報や静的な言語知識に強く依存している。したがって、問題点として、大規模な対訳コーパスが必要となることや多様な言語現象に追随することが困難となることが考えられる。ただし、大規模な対訳コーパスが得られない言語を対象とした場合、対応付けされていないコーパスを利用することは有効である。したがって、今回抽出した対訳語の利用を出発点として、今後、対応付けされていないコーパスからの対訳語の自動抽出についても取り組んでいきたいと考えている。

7　おわりに

本論文では、出現頻度の低い対訳語が多くを占める、小規模な対訳コーパスから対訳語を自動抽出可能な手法として、局所着目型学習を用いた対訳語の抽出手法を提案した。性能評価実験では、大規模な対訳コーパスの収集が困難であるアイヌ語－日本語対訳コーパスを対象に名詞および動詞対訳語の自動抽出を行った。その結果、平均出現頻度1.96である288文の対訳コーパスに対して、再現率54.0%、適合率60.8%が得られた。この結果は、統計的手法の再現率に比べ10%以上高く、本手法の有効性を示すものである。

また、本論文では、アイヌ語文に対しては解析ツールが存在しないため言語知識を使用していない。たとえ、今後、アイヌ語の電子化された対訳辞書や形態素解析ツール、構文解析ツールなどが開発された場合でも、自然な訳語や新たな表現の訳語が対訳コーパスに出現する限り、対訳コーパスの重要性は変わらない。したがって、対訳コーパスから出現頻度の低い対訳語を自動抽出可能な本手法の有効性もまた変わらないと考えられる。

今後は、更なる精度向上を目指すと共に、今回の実験で得られた名詞および動詞対訳語をもとに対応付けされていないコーパスからの対訳語の抽出を図る。そして、他の言語にも適用していくことで本手法の有効性をさらに検証する予定である。

参　考　文　献

[1] 関根聡：不意打ち言語試験!?、情報処理学会学会誌、Vol.44、No.11、pp.1157－1159 (2003).

[2] 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).

[3] 田中（石井）久美子，梅村恭司，岩崎英哉：第三言語を介した対訳辞書の作成，情報処理学会論文誌，Vol.39, No.6, pp.1915－1924 (1998).

[4] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer., R. and Roossin, P. ：A Statistical

Approach to Language Translation,Proc. 12th International Conference on Computational Linguistics, pp. 71－76 (1988).

[5] Kupiec, J. : An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, Proc. 31st Annual Meeting of the Association for Computational Linguistic, pp. 17－22 (1993).

[6] Haruno, M., Ikehara, S. and Yamazaki, T. : Learning Bilingual Collocations by Word-Level Sorting, Proc. 16th International Conference on Computational Linguistics, pp. 525－530 (1996).

[7] Smadja, F., McKeown, K. and Hatzivassiloglou, V. : Translating Collocations for Bilingual Lexicons : A Statistical Approach, Computational Linguistics, Vol. 22, No. 1, pp. 1－38 (1996).

[8] 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, pp. 727－736 (1997).

[9] 山本薫，松本裕治：統計的係り受け結果を用いた対訳表現抽出，情報処理学会論文誌，Vol. 42, No. 9, pp. 2239－2247 (2001).

[10] 佐藤健吾，斎藤博昭：サポートベクタマシンを用いた対訳表現の抽出，自然言語処理，Vol. 10, No. 4, pp. 109－124 (2003).

[11] 越前谷博，荒木健治，桃内佳雄，栃内香次：実例に基づく帰納的学習による機械翻訳手法における遺伝的アルゴリズムの適用とその有効性，情報処理学会論文誌，Vol. 37, No. 8, pp. 1565－1579 (1996).

[12] 越前谷博，荒木健治，桃内佳雄，栃内香次：翻訳例に基づく再帰チェーンリンク型学習による機械翻訳手法，電子情報通信学会論文誌D－II, Vol.J 85-D-II, No. 12, pp. 1840－1852 (2002).

[13] Echizen-ya, H., Araki, K., Momouchi, Y. and Tochinai, K. : Effectiveness of Automatic Extraction of Bilingual Collocations Using Recursive Chain-link-type Learning, Proc. 9 th Machine Translation Summit, pp. 102－109 (2003).

[14] 辻慶太，芳鐘冬樹，影浦峡：対訳コーパスにおける低頻度語の性質－訳語対自動抽出に向けた基礎研究－，電子情報通信学会信学技報，NLC2000－16, pp. 47－54 (2000).

[15] 越前谷博，荒木健治，桃内佳雄，栃内香次：局所着目方式によるアイヌ語－日本語名詞対訳語の抽出について，電子情報通信学会信学技報，NLC2003－49, pp. 93－98 (2003).

[16] Och, F. J. and Ney, H. : Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, Proc. 40th Annual Meeting of the Association for Computational Linguistics, pp. 295－302 (2002).

[17] Güvenir, H. A. and Cicekli, I. : Learning Translation Templates from Examples, Information Systems, Vol. 23, No. 6, pp. 353－363 (1998).

[18] Daelemans, W., Gillis, S. and Durieux, G. : Skousen's Analogical Modelling Algorithms : A Comparison with Lazy Learning, New Methods in Language Processing : Edited by Jones, D. and Somers, H.,UCL Press, pp. 3－15 (1997).

[19] McTait, K. : Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns, In Proc. Workshop on EBMT, 8 th Machine Translation Summit.

[20] Brown, R. D. : Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation, Proc. 7th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 111－118 (1997).

[21] Tanaka, K. and Iwasaki, H. : Extraction of Lexical Translations from Non-Aligned Corpora, Proc. 16th International Conference on Computational Linguistics, pp. 580－585 (1996).

[22] 梶博行，相薗敏子：共起語集合の類似度に基づく対訳コーパスからの対訳語抽出，情報処理学会論文誌，Vol. 42, No. 9, pp. 2248－2258 (2001).

[23] 田中貴秋，松尾義博：対訳関係のないコーパスからの複合名詞対訳表現の獲得，電子情報通信学会論文誌D－II, Vol.J 84-D-II, No. 12, pp. 2605－2614 (2001).

[24] 田村すず子：アイヌ語沙流方言辞典，草風館（1998）.
[25] 中本ムツ子，片山龍峯：アイヌの知恵・ウパシクマ１，片山言語文化研究所，新日本教育図書株式会社（1999）.
[26] 中本ムツ子，片山龍峯：アイヌの知恵・ウパシクマ２，片山言語文化研究所，新日本教育図書株式会社（2001）.