E-49

再帰チェーンリンク型学習を用いたアイヌ語-日本語間の対訳辞書の自動構築について

Automatic Construction of the Bilingual Words Dictionary for Ainu-to-Japanese Using Recursive Chain-link-type Learning

加藤大樹† 越前谷博† 荒木健治‡ 桃内佳雄† 栃内香次†
Daiki Kato Hiroshi Echizen-ya Kenji Araki Yoshio Momouchi Koji Tochinai

1 はじめに

機械翻訳システムにおいて，対訳辞書は必要不可欠である．しかし，この対訳辞書を人手で作成するには膨大な時間と労力を必要とする．また，原言語と目的言語のどちらか一方の解析的な知識が乏しい場合，辞書作成は非常に難しい．辞書未登録語に対して，対訳コーパスもしくは対応付けされていないコーパスから対訳語を獲得する研究は盛んに行われている．それらは，言語知識を利用する手法[1][2]と統計的な手法[3][4]に大きく分けられる．言語知識を利用する手法では，基本的な対訳辞書を始めとして解析的な知識の使用を前提としている．一方，統計的な手法では，膨大な量のコーパスを必要とする．アイヌ語と日本語間の機械翻訳システムの構築を目的として対訳語を獲得する際，日本語においては解析ツールを用いて豊富な解析的な知識を得ることができる．しかし，現在のところ，アイヌ語においては同様の方法で解析的な知識を得ることは困難である．また，アイヌ語-日本語の対訳コーパスを大量に得ることも困難であるため，統計的な手法を用いて対訳語を獲得することは難しい．我々は本稿において，学習機能に基づき少量の対訳コーパスからシステムが対訳語を効率良く獲得することができる再帰チェーンリンク型学習を用いたアイヌ語-日本語間の対訳辞書の自動構築を行う手法を提案する．

2 概要

再帰チェーンリンク型学習を用いることにより，少量の与えられた対訳語をもとに，アイヌ語文と日本語文の対による対訳コーパスからアイヌ語の解析的知識を必要とすることなく対訳語を連鎖的に獲得する．今回，獲得の対象となる対訳語とは，名詞または，形容詞，助詞，名詞から構成される名詞句である．名詞または名詞句を獲得の対象としたのは，語形変化が少ない名詞または名詞句の獲得が実験の第一段階として適していると判断したためである．Fig.1に本手法による対訳語の獲得処理の具体例を示す．

Fig.1の処理を順に見ていく．対訳辞書に（ontaro ; { 樽 : 名詞-一般 }）という対訳語が存在するとき，対訳文1のアイヌ語文と日本語文のそれぞれから対訳語に対応した部分"ontaro"と"樽"を抽出し，変数に置き換える．そして，システムはアイヌ語と日本語のそれぞれにおいて，置き換えられた変数とその変数に隣接する語を切り離し情報として獲得する．このとき，日本語の切り出し部分には品詞情報が付与され，切り離しテンプレート（ta @ or ; { に @ の : 助詞-格助詞-一般 @ 助詞-連体化 }）が獲得される．この切り離しテンプレートは，アイヌ語の"ta"と"or"に隣接する語は日本語の「（助詞-格助詞-一般）の"に"」と「（助詞-連体化）の"の"」に隣接する語と対応関係を持つことを意味する．この獲得された切り離しテンプレートを次に対訳文2に適用する．対訳文2においてアイヌ語の"ta"と"or"に隣接する語は「tusir」，日本語の「（助詞-格助詞-一般）の"に"」と「（助詞-連体化）の"の"」に隣接する語は「墓」である．こうして新たな対訳語（tusir ; { 墓 : 名詞-一般 }）が獲得される．さらに，この対訳語を対訳コーパスに適用することにより，新たな切り離しテンプレートの獲得が期待できる．このように，対訳語と切り離しテンプレートが連鎖的に獲得されていく．

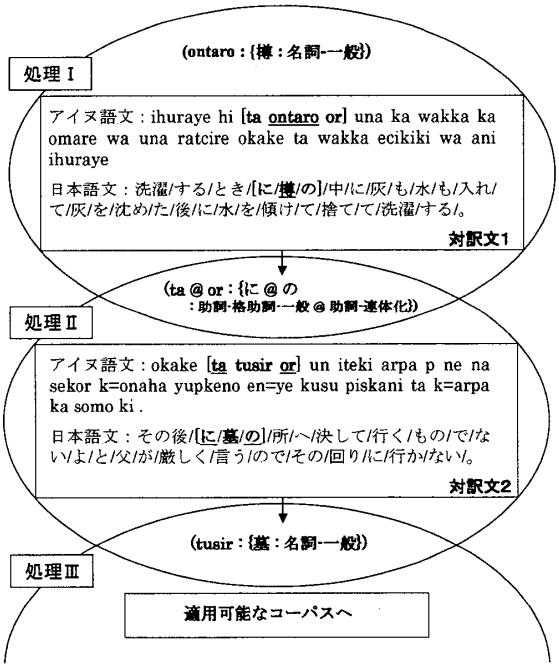

Fig. 1 対訳語獲得の具体例

3 処理過程

本システムでは対訳語と切り離しテンプレートが交互に獲得されることによって，連鎖的に対訳語を獲得していく．その際には，対訳語の獲得精度を向上させるため，日本語部分には形態素解析ツール茶筅[5]を用いて品詞情報を付与する．

まず，対訳語を用いた切り離しテンプレート獲得の処理過程について述べる．

(1) 対訳文のアイヌ語文と日本語文に対し，アイヌ語部分と日本語部分が共に含まれている対訳語を選択する．

(2) 対訳文のアイヌ語文と日本語文のそれぞれにおいて，対訳語が当てはまる部分を変数に置き換え，その変数に隣接する語を抽出する．隣接する語の存在パターンは前後に隣接する語がある場合，前の

† 北海学園大学，札幌市
‡ 北海道大学，札幌市

みに隣接する語がある場合，後ろのみに隣接する語がある場合の3通りがある．なお，日本語抽出部分には品詞情報を付与する．この品詞情報を用いて，同一字面で違う品詞の語を区別することにより，誤った対訳語の獲得を防ぐことができる．

(3) 抽出されたアイヌ語と日本語の変数に隣接する語を切り離しテンプレートとして獲得する．

次に，切り離しテンプレートを用いた対訳語獲得の処理過程について述べる．

(1) 対訳文のアイヌ語文と日本語文に対し，アイヌ語部分の変数に隣接する語と日本語部分の変数に隣接する語が共に含まれる切り離しテンプレートを選択する．このとき，日本語部分の変数に隣接する語は，字面と品詞の両方が一致する語を選択することにより，正しい対訳語を抽出できる．

(2) 対訳文のアイヌ語と日本語のそれぞれにおいて，変数に隣接する語で挟まれている部分を抽出する．このとき，抽出された日本語部分の品詞情報を調べる．その結果，名詞または名詞句でない場合は対訳語としての登録は行わない．

(3) アイヌ語と日本語の抽出された語を対訳語として獲得する．

4 性能評価実験

4.1 実験方法

対訳コーパスとして「アイヌの知恵・ウパシクマ1」[6]に記載されているアイヌ語文とその日本語訳文の対を116組用いた．この対訳コーパスは5つの章に分かれており，出現する名詞の数は172個である．初期辞書として第1章の名詞34個の対訳語を与え，それらを基に対訳語をどの程度獲得できるかということを調査するための実験を行った．

4.2 実験結果と考察

Table 1 実験結果

初期数	獲得数	増加率
34	19	55.9%

システムが新たに獲得できた名詞の数は19個であった．したがって，初期状態として与えた対訳語数34個に対する増加率は55.9%となる．システムが獲得した対訳語が対訳コーパス内で何回出現したかを調べたところ，1回のみ出現した語は5個，2回出現した語は6個，3回出現した語は4個，4回以上出現した語は4個であった．1回または2回しか出現していない対訳語の割合は，獲得した全対訳語の57.9%となった．これは，システムが対訳語の出現回数が少なくても対訳語を獲得できるという高い学習能力を持つことを意味する．アイヌ語-日本語の対訳コーパスは，英語-日本語の対訳コーパスほど多くのデータが得られることを期待できず，また，用意した対訳コーパスに必ずしも同じ単語が何度も出現するとも限らない．実際，本実験で用いた対訳コーパスにおいて，各単語の平均出現回数を調べたところ1.8回であった．しかし，このような同じ単語の出現が何度も望めない状態からでも，対訳語を獲得することができる本手法は有効性が高いといえる．また，初期辞書に与えなかった全名詞に対する対訳語の正対訳語取得率は13.8%であった．したがって，正対訳語取得率の観点から見た場合，十分な結果が得られたとは言えない．その理由は，システムが対訳語を獲得するには1回の出現で獲得できるが，さらなる対訳語を獲得するためには新たな切り離しテンプレートの獲得が必要であり，そのためには獲得した対訳語がもう一度他の対訳コーパス中に出現しなければならないためである．しかし，今回用いた対訳コーパスにおいては，各単語の平均出現回数は1.8回であったことから，そのような対訳語と切り離しテンプレートの獲得の連鎖が不十分であった．

一方，システムが獲得した対訳語に対する精度は，39.1%となった．これは，一つのアイヌ語の切り離し部分に対応する日本語の切り離し部分が複数存在した場合，全てのパターンを辞書に登録していたことが理由の一つに挙げられる．しかし，アイヌ語と日本語の語順は似ていることから，仮に，アイヌ語の切り離し部分の位置と日本語の切り離し部分の位置が同じ場合のみを対訳語とするヒューリスティックを適用すると，精度は44.6%に向上する．

今回は非常に厳しい学習環境での実験であるなか，55.9%の正対訳語の増加率がみられた．したがって，本手法は有効であると考えられる．

5 おわりに

本稿では，アイヌ語-日本語対訳コーパスから，アイヌ語に対し解析的な知識を得られない状況において，システムが自動的に対訳語を効率良く獲得できる再帰チェーンリンク型学習を用いた対訳辞書構築について述べた．再帰チェーンリンク型学習を備えたシステムにより，55.9%の正対訳語の増加率がみられた．本手法で獲得した対訳語のうち，1回もしくは2回しか出現していない単語が全獲得対訳語のうちの57.9%を占めた．これは本手法が高い学習能力を持つことを示し，対訳コーパスの量を多く望めないアイヌ語の対訳語の獲得において有効である．また，システムが獲得した対訳語に対する精度は39.1%であったが，対訳語獲得の出現位置に関するヒューリスティックスを適用することで44.6%に向上可能である．

今後は，再帰チェーンリンク型学習が有効に働くような学習環境（対訳コーパス）を用い，評価実験を継続する予定である．

参考文献

[1] 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-pp736(1997).

[2] 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，Vol.J84-D-ll, No.12, pp2605-pp2614(2001).

[3] Masahiko Haruno, Satoru Ikehara, Takefumi Yamazaki : Learning Bilingual Collocations by Word-Level Sorting, COLING-96, pp525-pp530 (1996).

[4] Reinhard Rapp : Identifying Word Translations in Non-Parallel Texts, 33rd Annual Meeting of ACL, pp320-pp322(1995).

[5] 日本語形態素解析システム「茶筅 (Chasen) for Windows」(2000).

[6] 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).