E-56

対訳コーパスからの辞書未登録語抽出における再帰チェーンリンク型学習の有効性について

Effectiveness of Recursive Chain-link-type Learning in Extraction of Unknown Words from Bilingual Text Corpora

越前谷博† 荒木健治‡ 桃内佳雄† 栃内香次†

Hiroshi Echizen-ya Kenji Araki Yoshio Momouchi Koji Tochinai

1 はじめに

機械翻訳システムの問題の一つに，辞書未登録語の問題がある．翻訳結果中に不適切な訳語が含まれている状態は，ユーザに違和感を与えることになる．辞書未登録語の問題に対し，対訳コーパスから対訳語を自動的に獲得する研究はこれまで数多く行われている．その代表的なものとしては，統計処理に基づき大量のコーパスから対訳語を獲得する手法 [1] や基本となる言語情報を用意し，それに基づき対訳語を決定していく手法 [2][3] が提案されている．しかし，統計情報を用いる場合，大量の対訳コーパスが必要になることが問題となる．言語情報を利用する場合は，初期状態として与えられる言語知識によって獲得精度は変化してしまうことが問題となる．近年では，対応関係にある対訳コーパスを収集することの困難さから，対訳関係のないコーパスから対訳語を獲得する手法 [4] も提案されている．しかし，そのような研究の多くは，構成語間の対応関係を利用することが可能な複合語を対象としている．

そこで，我々は，文対応での対訳テキストを前提とするが，静的な言語知識を必要とせずに，かつ少規模な対訳コーパスからでも効率よく対訳語を自動獲得することが可能な，再帰チェーンリンク型学習による辞書未登録語の抽出手法を提案する．本手法では，対訳テキストから自動獲得された対訳ペア自体が，他の対訳テキストからの対訳ペアの抽出に必要な情報を保持していると捉えることで，連鎖的な対訳ペアの獲得を実現する．その際，利用する情報は表層情報のみであり静的な言語情報を必要としない．更に，本手法での抽出対象は，対訳テキスト中の部分であるため，1語の対訳語から複数の語で構成される対訳語まで語数に依存せずに統一的な枠組みで対訳語を獲得できる．本稿では，この再帰チェーンリンク型学習と本手法の有効性を確認するために行った性能評価実験の結果について述べる．

2 再帰チェーンリンク型学習による対訳語の獲得

本稿では，原言語文から得られたテキストの部分と目的言語文から得られたテキストの部分をそれぞれ原言語部，目的言語部と呼ぶ．そして，原言語部と目的言語部の組を対訳ペアとする．再帰チェーンリンク型学習では，様々な対訳テキストから連鎖的に対訳ペアを獲得する．再帰チェーンリンク型学習の処理の概略図を Fig.1 に示す．Fig.1 の対訳ペア A（"Z"；"ζ"）において，それがある対訳テキストから抽出された対応関係の正しい部分であるなら，対訳ペア A は原文に "Z" が，また，訳文に "ζ" が含まれている他の対訳テキストからも（"Z"；"ζ"）を抽出可能であるという情報を持っていると捉えることができる．また，その結果得られた対訳ペア B は，変数に相当する部分，すなわち，英文中の "E" の右側から "H" の左側まで，訳文中の "θ" の右側から "η" の左側までを抽出可能な範囲であるという情報を持っていると捉えることができる．その結果，システムは対訳ペア B の持つ情報に基づき，原文に "E" と "H" がこの順で存在し，かつ，訳文に "θ" と "η" がこの順で存在する，他の対訳テキストからそれらに挟まれた部分を抽出することができる．このように，再帰チェーンリンク型学習では，獲得された全ての対訳ペアは，他の対訳テキストから新たな対訳ペアを獲得するために必要な情報を有していると捉えることで，Fig.1 に示すように，様々な対訳テキストからの連鎖的な対訳ペアの獲得を実現している．

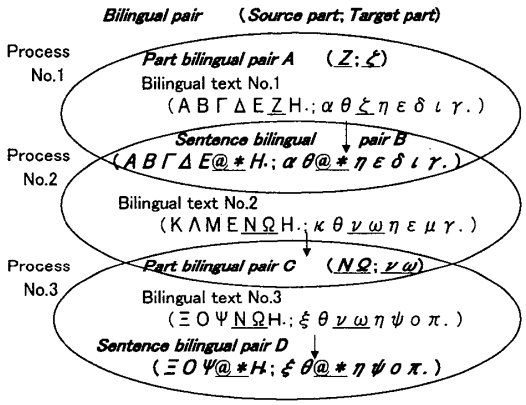

Fig. 1 再帰チェーンリンク型学習の概略図

本稿では，対訳ペア A，C のように対訳テキストの一部分を表現している対訳ペアを部分対訳ペア，また，対訳ペア B，D のように対訳テキストの全体を表現している対訳ペアを文対訳ペアと呼ぶ．対訳語は部分対訳ペアに属する．また，部分対訳ペア A のように，再帰チェーンリンク型学習を稼動させる際の起点となる対訳ペアは，我々が既に提案している学習型機械翻訳手法である，遺伝的アルゴリズムを適用した帰納的学習による機械翻訳手法（GA-ILMT）[5] を用いて獲得する．

3 再帰チェーンリンク型学習の処理過程

最初に，Fig.1 の処理 1 のような，部分対訳ペアを用いた，文対訳ペアの獲得の処理過程を述べる．

(1) 対訳ペアの原言語部と目的言語部が共に対訳テキストの原文と訳文に対し，包含関係にある部分対訳ペアを選択する．

(2) 対訳テキストの原文と訳文のそれぞれにおいて，重複する部分を変数に置き換える．そして，それらを一つのペアとすることで文対訳ペアを獲得する．

次いで，Fig.1 の処理 2 のような，文対訳ペアを用いた，部分対訳ペアの獲得の処理過程を述べる．

† 北海学園大学，札幌市
‡ 北海道大学，札幌市

(1) 対訳テキストとの間で，共通部分の存在する文対訳ペアを選択する．
(2) 共通部分が文対訳ペア中の変数と隣接する場合，以下のいずれかの処理を行う．
- 文対訳ペアの変数が共通部分に挟まれている場合，対訳テキストの原文と訳文中の共通部分で挟まれている部分を抽出する．
- 文対訳ペアにおいて，変数の右側のみに共通部分が存在する場合，対訳テキストの原文と訳文中の共通部分の左側から文の先頭までを抽出する．
- 文対訳ペアにおいて，変数の左側のみに共通部分が存在する場合，対訳テキストの原文と訳文中の共通部分の右側から文の末尾までを抽出する．

(3) 対訳テキストの原文と訳文のそれぞれから抽出された部分を組み合わせることで，部分対訳ペアを獲得する．

4 性能評価実験

4.1 実験方法

初めに中学1，2年生用の英語テキスト [6][7][8][9][10] に掲載されている対訳テキスト 2,856 組を再帰チェーンリンク型学習を備えたシステムに与え，対訳ペアを自動獲得させた．その際の辞書の初期状態は空から始めた．これは，本システムが静的な言語知識を必要とせずに，対訳語を自動獲得可能なことを確認するためである．また，同様の対訳テキスト 2,856 組の英文を A 社より市販されている商用機械翻訳システムに翻訳させた．その結果，商用機械翻訳システムにおいて，辞書に登録されていないため，翻訳結果に英単語の状態で出力された語を抽出することで辞書未登録語を得た．

4.2 実験結果と考察

商用の機械翻訳システムを用いて得られた辞書未登録語の数は 37 語であった．これらは全て “Shinkansen” や “Mt.Asama” などの固有名詞であった．また，未登録語 37 個の構成単語数は 1 単語の未登録語が 30 個，2 単語の未登録語が 7 個であった．実験の結果，再帰チェーンリンク型学習を備えたシステムにより，未登録語 37 個中の 29 個に対して正しい訳語が獲得された．したがって，78.4%の未登録語が解決されたことになる．また，再帰チェーンリンク型学習を用いない，GA-ILMT のみを備えたシステムでは，獲得された対訳語は 23 個となり，精度としては 62.2%であった．この結果は，再帰チェーンリンク型学習により，対訳語の獲得能力が向上したことを示している．

また，本システムの学習能力の高さを確認するために，未登録語を含む対訳テキストがどれだけ与えられた結果，対訳語を獲得できたのかについて調査を行った．その結果，獲得された 29 個の対訳語のうち 19 個が，1 度だけ対訳テキストに出現した段階で獲得されていた．また，2 度対訳テキストに出現した後に獲得可能となった場合が，5 個存在した．3 度以上対訳テキストに出現した後に獲得された場合は，5 個だけであった．したがって，本システムにより獲得された対訳語 29 個の 82.8%が 1 度もしくは 2 度だけの少数の出現回数にもかかわらず，効率よく獲得されていることが確認された．Fig.2 に，対訳語の獲得の具体例を示す．Fig.2 では，再帰チェーンリンク型学習により，対訳語として（Rika；りか）が獲得された．

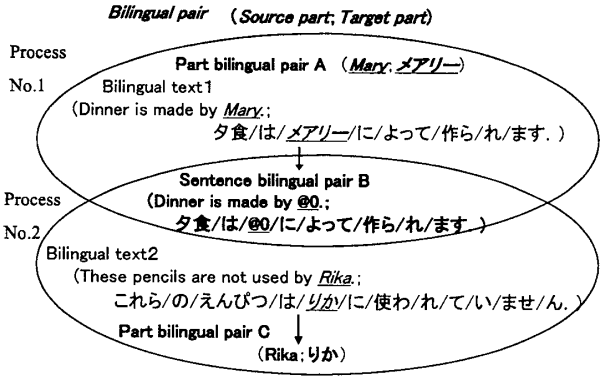

Fig. 2 対訳語の獲得の具体例

5 おわりに

本稿では，対訳コーパスから対訳語を学習能力に基づき獲得する，再帰チェーンリンク型学習による対訳語の自動抽出手法を提案した．再帰チェーンリンク型学習を備えたシステムにより，商用の機械翻訳システムにおける辞書未登録語の 78.4% が解決された．本手法の利点は，表層情報のみを用いて対訳ペアを獲得できるため，静的な言語知識を必要としない点，獲得対象が対訳テキスト中の部分となるため，獲得する対訳語の構成単語数に対する制約を持たない点，そして，少規模の対訳コーパスから効率よく対訳語を獲得できる点である．更に，一般的に手がかりとなる情報が乏しく，獲得が困難であると考えられる固有名詞に対し，本手法が有効であることを評価実験に基づき確認した．

謝辞

本研究の一部は，北海学園大学ハイテク・リサーチ・センター研究費によって行なわれている．

参考文献

[1] 北村美穂子，松本祐治：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(1997).
[2] 山本由紀雄，坂本仁：対訳コーパスを用いた専門用語対訳辞書の作成，情報処理学会研究報告，NL94-12(1993).
[3] 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(1994).
[4] 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.J84，No.12(2001).
[5] 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(1996).
[6] 教科書ガイド教育出版版ワンワールド 1，日本教材，東京 (2001).
[7] 教科書ガイド教育出版版ワンワールド 2，日本教材，東京 (2001).
[8] 教科書システム問題集 2，朋友出版，東京 (2001).
[9] 教科書ワーク 2，文理，東京 (2001).
[10] 教科書トレーニング 2，新興出版社，大阪 (2001).