対訳テキストを用いた分野別翻訳規則の抽出

北村 美穂子*
沖電気工業株式会社

松本 裕治
奈良先端科学技術大学院大学

1 はじめに

機械翻訳システムの専門辞書を作成する方法の一つに既存対訳テキストを利用する方法がある．これはテキストに頻出する定型表現や専門用語に関する翻訳対を既存の対訳テキストから抽出し，辞書に登録する方法である．従来，この作業は人手で行なわれていたが，対訳テキストの統計情報や構文情報を用いて単語や依存構造の自動対応付けを行ない，その結果を辞書として利用する方法が提案されている(2)(6)．対応付けを自動化することにより，人手による作業が軽減されるだけでなく，一定した基準で翻訳対を抽出することができる．さらに分野別に分類された対訳テキストがあれば，分野別の辞書も容易に作成することができる．

専門性の高いテキストには専門用語だけでなく，複数単語からなる固有名詞や定型表現が多く含まれる．単語列や文字列の出現頻度を用いた共起表現や定型表現を抽出する試みは単言語テキストでは広く行なわれており(3)(4)，複数単語にまたがる翻訳対の抽出においても類似の方法が適用できると考える．

本稿は，文対応の付いた対訳テキストから複数単語列の二言語間での組合せを作成し，その出現頻度から対応度を求め，対応度の高いものから順に翻訳対を抽出する方法を提案する．さらに，本方法を用いて，分野の異なる2つのテキスト(ビジネス文書，計算機マニュアル)に関する専門用語，定型表現ならびにその訳語を抽出する実験を行なった．その結果についても報告する．

2 分野別翻訳対の抽出

分野別に分類されたテキストでは使用される単語や言い回しは限定され，その分野の専門用語やテキスト固有の定型表現が頻出する．さらに，対訳テキストの場合では，テキスト内で統一した訳語が使用される場合が多い．このため，原文に複数回出現する単語や言い回しは，常に同じ訳語が対応する．このような分野依存の対訳テキストの性質に基づいて翻訳対の抽出を行なう．

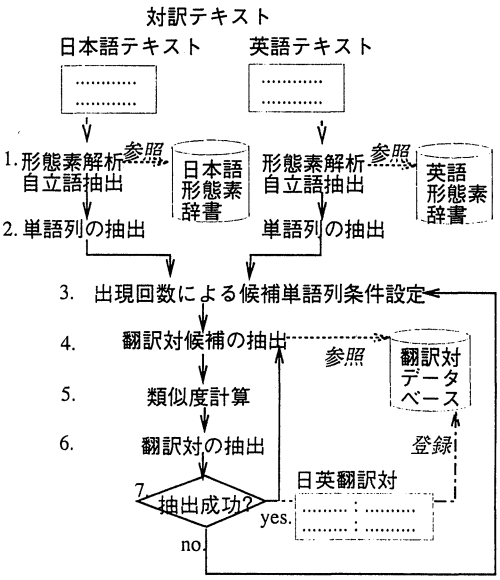

図 1: 翻訳対抽出のための処理のながれ

翻訳対は原言語と目的言語の単語または連続する複数単語の対により構成される．翻訳対の候補となる単語列の対が翻訳対として適切かどうかを判断するために，対訳テキスト内の原言語と目的言語の単語列の対応度を示す計算式(「2.1 翻訳対の抽出手順」の5.を参照)を定義する．これは(1)で定義された計算式に出現回数による重みを加えたもので，出現回数が多いほど対応付けの信頼性が高くなるように設定されている．

この設定下において，出現回数が多いものから段階的に抽出を行なう．翻訳対を構成する各言語の単語列

*kita@kansai.oki.co.jp

の対の出現回数が多いものから少ないものへと抽出対象を広げていくことにより，対応度の高い翻訳対から順に抽出することができる．また，一単語だけでなく，複数単語からなる単語列を対象としているため，候補となる各言語の候補単語列数およびその組合せ数は非常に大きくなるが，抽出候補を最初は出現回数の多いものに限定し，そこで抽出された翻訳対を候補の対象から省きながら出現回数の制限を緩めていくという方法により計算時間を短縮することができる．さらに，抽出した翻訳対を取り除いて計算を繰り返すことにより，ある単語列の第一候補以外の複数の訳語も抽出することができる．

2.1 翻訳対の抽出手順

図1に処理のながれを示し，翻訳対抽出のための処理方法を以下に説明する．使用する対訳テキストは日本語と英語の対訳テキストとし，文単位の対応付けは既に行なわれているものとする．

1. 日本語テキストと英語テキストを形態素解析し，自立語単語のみを抽出する．

2. 各テキスト中の自立語単語の出現回数を数え，2回以上出現した単語のみを抽出する．次にその単語を先頭とした2連続単語を作成し，その出現回数を数える．さらに2回以上出現した2連続単語を先頭とする3連続単語に対しても同様に処理し，以降連続単語の数を拡張しながら各単語列の出現回数を数える．あらかじめ連続単語の拡張数を設定しておき，その値に達すれば処理を終了する．

3. 日本語，英語の単語列の出現回数の下限(f_{min})を設定する．これは，以下4.から7.の処理を経た後，再度繰り返されることになるが，2回目以降，f_{min}は前回より低い値に設定される．

4. 3.の出現回数の下限条件を満たす日本語，英語の単語列(一単語も含む)において，各単語列の出現回数および，一対訳文内に出現する日本語単語列と英語単語列の組合せ回数を数える．ただし，両者の対応が対応が既に翻訳対データベースに登録されている場合は，数え上げの対象から省く．

5. 4.における対訳テキスト内の日本語単語列(w_J)，英語単語列(w_E)の対応度を計算する．対応度は，w_J, w_Eが出現する回数とその対応が出現する割合を利用するだけでなく，その対応の出現回数も考慮した以下の計算式(1)で定義し，(2)式を満たす対応だけを以下の処理の対象とする．

$$sim(\langle w_J, w_E\rangle) = (log_2 f_{je})\frac{2f_{je}}{f_j + f_e} \quad (1)$$

$$sim(\langle w_J, w_E\rangle) \geq log_2 f_{min} \quad (2)$$

w_J: 日本語単語列　w_E: 英語単語列
f_J: w_Jの出現回数　f_E: w_Eの出現回数
f_{JE}: w_J, w_E対応の出現回数
f_{min}: 出現回数の下限条件

なお，出現回数f_{min}未満の対応類似度の最高値は$log_2 f_{min}$を越えることはないため，出現回数にかかわりなく，この値以上の対応度を持つ対はすべて考慮の対象になっていることが保証されている．

6. 計算されたw_J, w_Eの対応類似度計算結果から翻訳対として最適な対応を以下の方法で求める．

   (a) w_Eに対応する全ての$WJ = \{w_{J1}, w_{J2}, ..., w_{Jn}\}$を取り出す．$w_{Jl}, w_{Jm} \in WJ$において$w_{Jm} \subset w_{Jl}$[1]かつ，$sim(\langle w_{Jm}, w_E\rangle) < sim(\langle w_{Jl}, w_E\rangle)$であれば，$\langle w_{Jm}, w_E\rangle$を翻訳対の候補から外す．$w_J$と，対応する$WE = \{w_{E1}, w_{E2}, ..., w_{En}\}$についても同様に処理する．

   (b) w_Eに対応する全ての$WJ = \{w_{J1}, w_{J2}, ..., w_{Jn}\}$の中で最大の対応類似度をもつ$\langle w_{Jt}, w_E\rangle (w_{Jt} \in WJ)$を取り出す．$w_J$も同様に，最大の対応類似度をもつ対応$\langle w_J, w_{Es}\rangle (w_{Es} \in WE)$を取り出す．その結果，$\langle w_{Jt} w_E\rangle$と$\langle w_J, w_{Es}\rangle$において，$w_{Jt} = w_J$かつ$w_E = w_{Es}$を満たす$\langle w_J, w_E\rangle$を翻訳対とする．

[1] w_{Jl}の構成語がw_{Jm}の構成語を含むを意味する

7. 6. の翻訳対を翻訳対データベースに登録する．抽出された翻訳対の数がある値以上であるならば，4. の処理に戻り，その出現回数条件における翻訳対抽出処理を繰り返す．ある値以下であれば，3. の処理に戻り，出現回数の下限(f_{min})を現在より低い値に再設定した後，4. から6. の処理を繰り返す．もし，出現回数が翻訳対抽出の最低条件に達したならば処理全体を終了する．

3 翻訳対の抽出実験

3.1 実験方法

実験には，「取引条件表現辞典例文」[5](以下ビジネス文書と呼ぶ)，「計算機マニュアル」の2種類の対訳テキストを用いた．まず，各対訳テキストの各文を形態素解析し，自立語のみを抽出した．日本語，英語の形態素解析には，機械翻訳システム PENSÉE[2]の形態素解析部を使用した．対訳テキストの文数は，計算機マニュアルは18,249文，ビジネス文書は18,754文で重複する文を一文にまとめた結果，各テキストはそれぞれ11,477文，10,016文になった．

実験における各種の設定は以下のとおりである．図1.2における翻訳対を構成する単語列の拡張数は10連続を最高とする．抽出対象とする翻訳対の出現回数の最低条件は計算機マニュアルでは2回，ビジネス文では3回とした．図1.3. で設定する段階別の出現回数条件は表2の出現回数の欄に示すとおりである．また，各段階において抽出された翻訳対の数が30以下になった時に繰り返し処理を停止することにした．

3.2 実験結果および考察

表1に各対訳テキストの候補単語の数および候補単語列(一単語も含む)の数を示す．候補となる単語および単語列は，計算機マニュアルでは2回以上，ビジネス文書では3回以上出現したものである．これは，()内に示す全出現単語数と比較すると，全単語数の約7割であるが，出現単語延べ数にすると全体の99%に相当する．専門用語，定型表現の抽出を目的とした場合では出現回数で候補を絞ることは問題ないと思われる．

[2]PENSÉEは，沖電気工業株式会社，大阪ガス株式会社および株式会社オージス総研の登録商標

	候補単語(全単語)		候補単語列	
	英語	日本語	英語	日本語
計算機	3,479(4,927)	4,801(7,372)	22,642	30,561
ビジネス	1,801(2,219)	2,627(3,571)	16,708	20,535

表 1: 日英候補単語数

計算機マニュアル				ビジネス文書			
出現回数	繰数	翻訳対数	精度	出現回数	繰数	翻訳対数	精度
14	2	270	99	16	2	213	96
7	3	909	96	8	2	282	96
5	2	166	96	5	3	400	91
4	2	260	92	4	3	319	85
3	3	539	91	3	3	755	84
2	4	1,355	87	—	—	—	—
計	16	3,499	91	計	13	1,969	87

表 2: 翻訳対抽出総数

次に，翻訳対の抽出数を出現回数段階別に表した結果を表2に示す．「出現回数」は各段階の抽出条件となる出現回数を，「繰数」は抽出処理が何回繰り返されたかを示す．「精度」は100翻訳対を任意に抽出し，人手によってその正否を判断した結果の正解数である．表2にみるように，最初の段階で抽出される翻訳対ほど精度が高く，後になるほど精度が低くなるという予想通りの結果が得られた．また，各テキストともに，最終段階でも80%以上の精度で翻訳対を抽出することができた．

抽出された翻訳対を語の構成別にみると，計算機マニュアルでは日本語と英語一単語対応における精度は100%であった．一方，複数単語列の対応は，複合名詞のようなパターンは高精度で抽出されているが，名詞，動詞が混在したパターンの場合の間違いが目立つ．これは，現仕様では「データ グラム(を) 送る(が) ソケット(が) すでに: send (a) datagram specify and (the) socket (is) already」のように句のまとまりを無

1. 計算機文書	
インタネット	internet
インタネット アドレス	internet address
インタネット プロトコル	internet protocol
インタネット プロトコル ＩＰ	internet protocol IP
p コマンド(および)	output (from the) p
ヌル コマンド(からの) 出力	(and) null command
ネーム トゥ アドレス マッピング	name (to) address map
倍精度	double precision
倍精度 浮動 小数点	double precision float point
発見的 手法	heuristics

2. ビジネス文書	
営業秘密	trade secret
営業時間	business hour
営業所	office
確認 取消 不能 信用状	irrevocable(and)confirmed letter(of)credit
技術 情報 ノウハウ	technical information know-how
技術 製造 ノウハウ	technique manufacture know-how
国際 商事 仲裁 協会	Japan commercial arbitration association

表 3: 抽出された翻訳対の例

視した単語列も翻訳対の抽出候補となり得るためである．候補単語列を絞るために，句読点など句の切れ目を考慮する，形態素解析結果の品詞情報を参照し，その構成品詞によって単語列を限定するなどの方法が考えられる．本結果から，特に連続名詞など名詞句に絞った翻訳対の抽出が有効であると思われる．

抽出された翻訳対で特徴的な例を表3に示す．両テキスト共，翻訳対の中には，専門性の高い翻訳対が数多くみられた．計算機マニュアルでは，日本語が英文字，カタカナ文字からなる翻訳対が多く，ビジネス文書では，熟語からなる翻訳対が多かった．計算機マニュアルにみられる「インターネット」を含む４つの翻訳対の例にみるように，ある一単語に関係する複数の翻訳対を抽出することができた．また，一つの対訳テキスト内においても後ろに接続する語によって訳語が異なる，「営業秘密:trade secret」，「営業時間:business hour」のような例も抽出された．

4 おわりに

本稿は，対訳テキストに出現する複数単語列の出現回数およびそれらの両言語間での対応度から，単語だけでなく連続する複数単語における翻訳対を抽出する方法を提案した．また，その方法を用いて分野の異なる２つの対訳テキストの翻訳対抽出実験を行なった結果，80%以上の精度で専門性の高い翻訳対を抽出できることが確かめられた．

今回の実験では専門用語や分野固有の翻訳対抽出を目的としたため市販の翻訳辞書など既存の翻訳知識は用いなかったが，対訳テキストと同分野の専門辞書が既にあれば，それらを翻訳対データベースにあらかじめ組み込むことも可能である．さらに，本方法により抽出した翻訳対をあらかじめ翻訳対データベースに組み込んで，それとは別の同分野の対訳テキストを用いて翻訳対抽出を行なうことにより，翻訳対データベースを増強していくことも可能である．

本方法では抽出できる翻訳対は連続する単語に限定される．今後は，ここで提案したテキストの表層的な解析のみを用いる方法と，構造照合のような依存構造までを考慮した方法[7]を融合させることにより，翻訳処理に有益な翻訳規則を獲得することが課題となる．

参考文献

(1) M. Kay and M. Röscheisen. Text-Translation Alignment. *Computational Linguistics*, 19(1):121–142, 1993.

(2) A. Kumano and H. Hirakawa. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. *Coling-94*, 1:76–81, 1994.

(3) F. A. Smadja. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

(4) 下畑 さより, 杉尾 俊之, 永田 淳次. 隣接文字の分散値を用いた定型表現の自動抽出. 情報処理学会研究会 自然言語処理, 110-11, pages 71–78, 1995.

(5) 石上進. 取引条件表現法辞典 電子ブック版 第*1*巻 物品取引. 国際事業開発株式会社, 1992.

(6) 北村美穂子, 松本裕治. 対訳テキストを利用した翻訳辞書の自動作成. 「自然言語処理における学習」シンポジウム論文集, pages 150–157, 1994.

(7) 北村美穂子, 松本 裕治. 二言語対訳コーパスからの翻訳知識の自動獲得. 人工知能学会全国大会 *(第8回)* 論文集, pages 645–648, 1994.