英作文支援のための大規模な日英対訳表現の抽出

坂上 信也† 馬 青† 村田 真樹‡
†龍谷大学大学院理工学研究科
‡情報通信研究機構

１．はじめに

われわれは単語レベルとフレーズレベルでの英作文支援システムを開発してきた[1][2][3]．フレーズレベルでの英作文支援では，与えられた日英混在の文に対し，まず日本語のフレーズ部分を抽出し，それを単語に分割する．次に分割したそれぞれの日本語単語を辞書引きし，英語の訳語候補を得る．最後にそれらの訳語候補の組み合わせから最適なものを選択し出力する．しかし，このような方法では 2 つの大きな問題点が存在する．まず，英語フレーズは訳語候補の組み合わせのみから生成されるため，支援できるフレーズの範囲が大きく限定されてしまう．また，単語の訳語候補の数が多いためその組み合わせの数が膨大となり処理時間がかかってしまう．このような問題を解決するためには，日英対訳パターンに基づくアプローチ，すなわち，日英対訳パターン辞書を構築し，その辞書を介在して英作文支援を行う手法を導入することが考えられる．そこでわれわれはまず，パターン辞書を作成するにあたり必要となる大規模な日英対訳表現を大規模な日英対訳コーパスから抽出することを試みた．実験の結果，計 28 万文対の日英対訳コーパスに対し，約 23 万個の単語 n-gram ベースの対訳表現と約 3 万個の文節 n-gram ベースの対訳表現を抽出することができ，それぞれの精度が 23%と 40%であった．

２．英作文支援システム

英作文支援システムへの入力は日本語の単語またはフレーズ（名詞句と動詞句）を含む日英混在文である．入力された混在文に対し，その日本語部分を特定し，茶筌を用いて分割する（単語レベルの場合には単語の原型を求める）．分割した単語を辞書引きしその訳語候補を取得する．ただし，単語レベルのときは取得した訳語候補をそのまま用いるが，フレーズレベルのときは，訳語候補の追加･限定を行う．次に，単語レベルでは訳語と文脈，フレーズレベルでは訳語候補の組み合わせでクエリを構成し，クエリの高品質コーパスまたは Web でのヒット回数を調べる．ヒット回数の多いものを英訳として出力する（詳細は文献[3]を参照）．

３．英作文支援システムの問題点

本システムには特にフレーズレベルでの支援について(1)支援できるフレーズの範囲が非常に限定的であることと，(2)訳語候補の組み合わせ数が膨大で処理速度が低下することが 2 つの大きな問題点である．より具体的に述べると，問題点(1)については，名詞句において支援できるフレーズは複合名詞，形容詞＋名詞のみであった．同じく動詞句において支援できるフレーズは名詞＋動詞であった．したがって，たとえば「走っている犬」のような動名詞句を支援することができない．また，名詞句などを不定詞句に英訳するような，異なるパターンへの英訳も支援できない．問題点(2)については，本システムでは単語の訳語候補の組み合わせをフレーズの英訳候補（クエリ）としているため，単語の訳語候補が多いとその組み合わせ数が膨大となり処理速度が遅くなってしまう．

４．対訳表現の抽出

前節に述べた問題の 1 つの解決方法として，日英対訳パターンに基づくアプローチ，すなわち，日英対訳パターン辞書を構築し，その辞書を介在して英作文支援を行う手法が考えられる．そこでわれわれはパターン辞書を作る上で必要である大規模な日英対訳表現対を大規模な日英対訳コーパスから抽出することを試みた．

4.1 関連研究

関連研究としてはたとえば文献[4][5][6]を挙げることができる．文献[4]は名詞句のみを対象とし 15 万文対の日英対訳コーパスから 4.5 万個の対訳表現を抽出している．文献[5]は複数の対訳辞書から抽出した例文からなる小規模(8,500 文対)な対訳コーパスを対象とし，わずか（強抑制では）74 個，または（弱抑制では）161 個の対訳表現を抽出している．また，抽出した対訳表現は単語 n-gram のものだけであった．一方，文献[6]では異なる三つの分野のコーパスを対象とし，それぞれのコーパスは 1 万文対程度の小規模なものであった．また，それぞれのコーパスから抽出した対訳表現も単語 n-gram のものだけでその数も一番多いもので約 3,300 個程度であった．さらに，対訳コーパスからの，対訳表現の候補となる日英それぞれの単語 n-gram の抽出とそれ

らの類似度計算による対訳表現の抽出は，抽出の中間結果をフィードバックしながら繰り返しで行われているため，処理時間がかかる．これらに対し，われわれは実用レベルでのさまざまなフレーズに対する英作文支援の立場から，文献[4][5][6]に用いられた手法を必要に応じて用い，さらに文節 n-gram（つまり複数文節）ベースの対訳表現を抽出する方法も加え，先行研究より規模が大きく分野も広い新聞データのコーパスから，単語 n-gram ベースのみならず文節 n-gram ベースの大規模な対訳表現の抽出を試みている．文節 n-gram ベースの抽出は，「抽出した対訳表現に対訳関係にない単語（不要語）が存在する」という単語 n-gram ベースの抽出の問題点をクリアできると同時に，不要語のない長い対訳表現の抽出も可能となる．また，本研究では上述した文献[6]のような複雑な処理手法を取っていないため，抽出時間が短縮でき，より大規模なコーパスからの対訳表現の抽出が可能である．

4.2 対訳表現の抽出方法

4.2.1 単語 n-gram ベースの対訳表現の抽出

単語 n-gram ベースの対訳表現の抽出は(1)対訳コーパスからの日英それぞれの単語 n-gram の抽出と(2)日英の単語 n-gram の類似度計算による対訳表現の抽出という手順で行う．より具体的には，手順(1)は文献[4][5]と同様，日英の n-gram 以下の任意長の単語列（以下これを「単語 n-gram」と呼ぶ）についてコーパス内に日本語の文と英文のそれぞれに複数回出現したものを取り出す．ただし，抽出されたものの中には断片的な単語列が多く含まれておりそのようなものを取り除くため，文献[5]と同様，コーパス全体に対し単語列が他の単語列と重なっている場合抽出しない「強抑制型」と単語列が他の単語列と重なっていてもほかの文に独立して出現している場合抽出しない「弱抑制型」という２つの抑制方法を用いている．手順(2)においては，文献[5]の文番号の一致率と文献[6]の Dice 係数を類似度計算に用いた．すなわち，手順(1)で取り出した日英単語 n-gram について，類似度を計算し，その値が閾値を超えたものを日英の対訳表現とした．文番号の一致率と Dice 係数の計算式を以下に示す．

文番号の一致率：

$$sim(X_j, X_e) = \frac{X_j と X_e の文番号の一致数}{X_j の出現回数}$$

Dice 係数：

$$sim(X_j, X_e) = \frac{2 f_{je}}{f_j + f_e}$$

ただし，X_jは日本語の単語 n-gram，X_eは英語の単語 n-gram，f_jと f_eは X_jと X_eが独立に出現する回数，f_{je}は対訳文に同時に出現する回数である．なお，文番号の一致率の計算式において，分母は「X_jの出現回数」ではなく「X_jの出現する文の数」の方がより合理的と思われるが，実質的にはそれほど変わるものではないためここでは先行研究に合わせた．類似度の計算に文番号の一致率のほかに Dice 係数を用いた．その理由は文番号の一致率は日本語の出現回数のみで正規化しているため，例えば「東京大学⇔University of」のような，英語の方がその出現回数が高い「一般的な表現」の場合であっても抽出されてしまう可能性が高い．一方，Dice 係数では両言語表現の出現回数で正規化しているため，上記のような片方が「一般的な表現」の場合，類似度は小さくなり抽出されにくくなる．

しかし，このような単語n-gramベースによる対訳表現の抽出は，例えば「ベルルスコーニ首相と⇔with prime minister Berlusconi on the」のように，全体的には対訳関係にあるが部分的に対訳関係にない不要語”on the”が含まれているというような問題点がある．この問題を解決するために，本研究では文節n-gramベースでの対訳表現の抽出を試みた．

4.2.2 文節 n-gram ベースの対訳表現の抽出

文節 n-gram の対訳表現の抽出は(1)日本語文と英文の文節（句）単位への分割，(2)日英対訳コーパスからの日英それぞれの文節 n-gram の抽出，そして(3)日英の文節 n-gram の類似度計算による対訳表現の抽出という手順で行う．より具体的には，日本語文の文節単位への分割には CaboCha を，英文の文節（句）単位への分割には Charniak パーザを用いた．手順(2)と(3)は基本的に 4.2.1 節の手順(1)と(2)と同様に，まず日英の n-gram 以下の任意長の文節 n-gram を作成し複数回出現したものを抽出する．次に Dice 係数で類似度を計算する．

５．抽出実験

5.1 対訳コーパス

NICT コーパス(4 万文対)[7]と JENNAD コーパス[8](18 万文対)とロイター日英記事対応付けコーパス（7 万文対）を合わせた約 29 万文対から重複している文を取り除いた 28 万文対を用いた．

5.2 抽出条件

抽出する単語 n-gram の上限を 5-gram と

10-gram とした．文節 n-gram の上限を 5-gram とした．日英のそれぞれの n-gram の抽出条件はコーパス中の日本語文と英文にそれぞれ２回以上出現することとした．対訳表現の抽出条件は日英同時出現回数が２回以上で類似度が 0.5 以上であるとした．

5.3 単語 n-gram ベースの対訳表現の抽出

表１と表２に抽出結果（強抑制型）を示す．ただし，ここでの精度は抽出されたすべての対訳表現から無作為に抽出した 100 個に対し，対訳関係にあるか否かを人手で評価した結果であった．なお，全体的には対訳関係にあるが部分的に対訳関係にない不要語が含まれている場合も正解とした．

表１： 5-gram の場合：抽出数と精度

類似度計算	抽出数	精度
文番号の一致率	227,453	0.23
Dice 係数	208,856	0.23

表２： 10-gram の場合：抽出数と精度

類似度計算	抽出数	精度
文番号の一致率	114,342	0.28
Dice 係数	109,546	0.28

抽出条件を 5-gram から 10-gram に変更した場合，精度は 5%しか向上せず抽出数は半減するという結果となった（減少する理由は抑制をかけていることにある）．そこで，実際の抽出結果を見てみると上限を 10-gram に設定した場合では「労働市場の構造改革⇔the structural reform of the labor market」のように長い対訳表現も多く抽出されていることがわかった．図１と図２にそれぞれの場合でのみ抽出することのできた対訳表現の例を示す．

1J：朱鎔基副首相は
1E：vice-premier zhu rongji
2J：連立与党内の支援
2E：support from coalition partners in

図１ 5-gram の場合のみ抽出された例

1J：労働市場の構造改革
1E：the structural reform of the labor market
2J：和歌山地検は
2E：the Wakayama district public prosecutors office

図２ 10-gram の場合のみ抽出された例

図２の対訳表現はたとえば(2J, 2E)は英語表現の単語数が５個を超えたため，5-gram の場合では抽出できなかった．

表１と表２の結果から類似度計算に文番号の一致率とDice係数のいずれを用いてもその精度が同じであった．抽出数で見ると逆にDice係数の方がわずかに減っている．そこで，それぞれの類似度計算で抽出された対訳表現を比較してみることにした．ここで，文番号の一致率でのみ抽出された正しくない対訳表現例①「日本工業規格（ＪＩＳ）⇔ the international trade and industry」とDice係数でのみ抽出された，全体的には対訳関係にあるが不要語を含む対訳表現対例②「政治宣言⇔ political declaration that」を用いてそれぞれの類似度値・同時出現回数・日英それぞれの出現回数を用いて比較した．表３に例①の場合，表４に例②の場合を示す．

表３：例①の場合

類似度計算	類似度	文番号の一致数	日本語表現	英語表現
文番号の一致率	1	2	2	
類似度計算	類似度	同時出現回数	日本語表現	英語表現
Dice 係数	0.0164	2	2	254

表４：例②の場合

類似度計算	類似度	文番号の一致数	日本語表現	英語表現
文番号の一致率	0.4	2	5	
類似度計算	類似度	同時出現回数	日本語表現	英語表現
Dice 係数	0.571	2	5	2

表３から日本語表現の出現回数と英語表現の出現回数の差が非常に大きいことがわかる．このような場合，確率的に正しい対訳表現になっている可能性が低い．しかし，文番号の一致率の類似度計算では日本語表現の出現回数のみで正規化しているため，対訳表現としては間違っているにもかかわらず計算された類似度が高く抽出されてしまった．一方，Dice 係数の類似度計算では日本語と英語の両方の表現の回数で正規化しているため類似度が非常に低くなり抽出されなかった．次に，例②のように日英両表現の片方（英語）に一部不要な単語が含まれると，両方の表現の同時出現回数が大きく減る．しかし日本語表現の出現回数が減らないため，文番号の一致率で計算される類似度は低くなり抽出できなくなる可能性が高い．一方，Dice 係数の類似度計算では正規化に英語表現の出現回数も用いており，その出現回数も，不要語を含めているため大きく減る．そのため，Dice 係数で計算された類似度はそれほど下がら

なくて済む可能性が高く，上記のような対訳表現を抽出できる可能性が高い．

以上のことから，両言語表現の出現回数を考慮に入れた Dice 係数を用いる類似度計算は，片方が一般的な表現のものを抽出することを防ぐことと対訳関係にはあるが一部不要な単語を含んでいるような対訳表現をより多く抽出することができるということが言える．

次に抑制手法を弱抑制型に変更して類似度計算に Dice 係数を用いて実験を行った．その結果，抽出数は強抑制型で約 21 万個であったのに対して弱抑制型では約 23 万個で，抽出数はわずかしか増加しなかった．それは，長い単語列に含まれる短い単語列が，その長い単語列が出現する文以外に，単独に出現したケースが極めて少ないことを意味する．

5.4 文節 n-gram ベースの対訳表現の抽出

表５に抽出結果を示す．表５からもわかるように文節単位に分割することにより（抑制がある場合）精度が向上したが，抽出数が大きく減少してしまった．一方，無抑制の場合，抽出数は大幅に増えたが，精度が大きく低下した．抽出結果を分析してみると単語n-gramベースでは抽出された「ベルルスコーニ首相と⇔with prime minister Berlusconi on the」は「ベルルスコーニ首相と⇔with prime minister Berlusconi」となり，不要語を含まず正確に抽出できた．また，単語n-gramベースでは抽出することのできなかった「国民の生命身体財産を守る⇔to protect the lives persons and property of the people」のような不要語を含まず対訳表現の長いものを抽出することができた．

表５　抽出数及び精度

抑制方法	抽出数	精度
強抑制型	32,837	0.40
弱抑制型	33,332	0.40
無抑制	458,394	0.15

6. 終わりに

本稿ではまずフレーズレベルでの英作文支援システムの問題点を挙げその問題点を解決するために日英対訳パターンに基づくアプローチ，すなわち，日英対訳パターン辞書を構築し，その辞書を介在して英作文支援を行う手法を提案した．パターン辞書を作成するにあたり必要となる対訳表現を大規模な日英対訳コーパスから抽出することを試みた．抽出する対訳表現は単語 n-gram だけでなく，不要な単語が取り除ける文節 n-gram も対象とした．実験の結果，計 28 万対の日英対訳コーパスに対し，約 23 万個の単語 n-gram ベースの対訳表現と 3 万個の文節 n-gram の対訳表現を抽出することができ，それぞれの精度が 23%と 40%であった．文節 n-gram ベースで抽出した対訳表現の数が単語ベースでのそれに比べはるかに少なかったが精度の面では向上した．また，対訳表現の類似度計算に文番号一致率だけでなく Dice 係数も用いた．その両者については，抽出結果を分析することによりそれぞれが持つ特徴を明らかにすることができた．しかし，長い対訳文をできるだけ多く抽出するために n-gram の上限を上げると，n-gram の抽出に抑制を加えているために対訳表現の抽出数が大幅に減少してしまった．

今後は長い対訳表現の抽出を保証しながら対訳表現の抽出数の増加と精度の向上を図りたい．また，対訳関係が離散型の対訳表現も含め，より多くの対訳表現の抽出に取り組んでいきたい．最終的には，抽出された対訳表現から対訳パターンの辞書を作成し，英作文支援システムに組み込む予定である．

参考文献

[1] 中尾，馬，村田：大規模コーパスに基づく文脈可変型日英訳語選択，言語処理学第 13 回年次大会，pp. 195-198 (2007)

[2]Ma, Nakao, Murata, Isahara : Selection of Japanese-English Equivalents by Integrating High-quality Corpora and Huge Amounts of Web Data, LREC2008 (2008)

[3]Ma, Mori, Murata: Development of English-Writing Support Systems, Pacling2009, pp. 171-176 (2009)

[4] 神野，徳久，村上，池原:文型パターンによる日英翻訳のための名詞句パターン辞書の構築，言語処理学会第 11 回年次大会, pp.376-379 (2005)

[5] 道祖尾，村上，徳久，池原：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] 北村，松本，: 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)

[7] Uchimoto, Zhang, Sudo, Murata, Sekine, and Isahara: Mutilingual Aligned Parallel Treebank Corpus Reflecting Contexual Information and Its Applications, MLR2004, pp. 63-70 (2004)

[8] Utiyama and Isahara: Reliable Measures for Aligning Japanese-English News Articles and Sentences, ACL-2003, pp. 72-79 (2003)