辞書情報と規則を用いた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
‡鳥取大学大学院工学研究科情報エレクトロニクス専攻

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]．しかし，フレーズレベルでの英作文支援では，支援できるフレーズの範囲が大きく限定されてしまうという問題と，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という問題が存在した．これら問題の解決方法の一つとして，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を考え，そこでわれわれは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てきた[2].

本研究では，対訳表現抽出の更なる改良を行い抽出数の向上と精度の向上に取り組んだ．まず，より多くの対訳表現が抽出できるように日英それぞれの単語n-gramの抽出方法に改良を加えた．次に，単語列の先頭に「いる」，「こと」，「みたい」などの不適切な単語がつく表現を，n-gram作成時に取り除けるように人手で作成した規則を導入した．さらに，抽出した対訳表現をできるだけ正しいものに絞り込めるように対訳辞書情報を適用した．実験の結果，計28万文対の日英対訳コーパスに対し，対訳表現約12万5千個を精度0.96で抽出することができた．その結果から，改良手法が従来手法に比べ抽出数と精度の両方において飛躍的に向上したことがわかった．

2. 対訳表現の抽出

これまで対訳表現の抽出は(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った[2]. より具体的には，手順(1)はまず，日英対訳コーパスに対して形態素解析を行い，単語単位に分割する．次に，n-gram以下の任意長の単語列を作成し，コーパス内に日本語文と英文のそれぞれに複数回出現した単語列を取り出す．ただし，断片的な単語列を取り除くために，コーパス全体に対して単語列が他の単語列と重なっている表現を取り除く抑制処理を加えた．手順(2)においては，Dice係数を類似度計算式に用いた．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，その値が閾値以上のものを日英の対訳表現とした．Dice係数の計算式を以下に示す．

$$sim\ (x_j, x_e) = \frac{2\ f(x_j, x_e)}{f(x_j) + f(x_e)}$$

ただし，x_jは日本語の単語n-gram，x_eは英語の単語n-gram，$f(x_j)$と$f(x_e)$はx_jとx_eが独立に出現する回数，$f(x_j,x_e)$は対訳文に同時に出現する回数である．

本研究では，上記で説明した対訳表現の抽出手法をベースにいくつかの改良を行った．

2.1 単語n-gram抽出方法の改良

これまで対訳表現の抽出は単語n-gramの上限を5-gramまでと10-gramまでの二通りで行った．実験結果を分析すると，5-gramまでては抽出することのできた【連立与党内の支援⇔support from coalition partners in】という表現が10-gramまでの場合では抽出することができなかった．抽出できなかった理由として考えられる点以下に示す．これまでの対訳表現抽出手法では，それぞれのコーパス内で複数回出現した単語列に対して，抑制処理を加えて断片的な単語列を取り除く処理を行っていたが，その処理の際に5-gramまでては対訳表現として抽出することのできた単語列においても，その単語列を含む単語列が5-gramより大きいn-gramにおいて抽出された場合，その単語列は取り除かれてしまう．また，その単語列を含んだ単語列が類似度において閾値を満たさなかった場合，対訳表現としても抽出されないことになる．上記の【連立与党内の支援⇔support from coalition partners in】という例の場合，「連立与党内の支援」に対して「連立与党内の支援に対して」という単語列が6-gramで抽出されたために，「連立与党内の支援」という表現は取り除かれてしまった．また，「連立与党内の支援に対して」と「support from coalition partners in」のペアの類似度が閾値以下のため対訳表現としても抽出することができなかった．このような問題点に対処するために，本研究では5-gramまでと10-gramまでだけでなく2-gramまで，3-gramまで，4-gramまで，5-gramまで，6-gramまで，といったようにそれぞれ対訳表現を抽出し1つにまとめ，重複している対訳表現に関しては取り除く処理を行った．

この改良方法による対訳表現の抽出手順は以下の

通りである．(1)日英それぞれのコーパスに対して形態素解析を行い，単語単位に分割する．次に，n-gramを作成し，コーパス内に日本語文と英文のそれぞれに複数回出現した単語列を取り出す．(2)始めに2-gramまでの連続単語列を抽出し，抑制処理を行った後，類似度を計算し閾値以上のペアを対訳表現として抽出する．(3)2-gramまでから一つ増やして3-gramまでにして，手順(2)と同様の処理を行う．この処理をあらかじめ設定しておいた上限Nまで値を1つずつ繰り上げながら，繰り返し行う．(4)各n-gramまででそれぞれ抽出された対訳表現を1つにまとめ，重複している表現のみを取り除く．

2.2 人手規則の導入

これまでの実験結果を分析してみると，日本語表現において，「いる」や「こと」のような文字列から始まる単語列が対訳表現として抽出されていた．しかし，このような単語が先頭に来ることは文法上不適切であると考えることができる．また，「この」や「いろいろな」などの単語が最後に来ている場合も，文法上不適切であると考えることができる．英語表現においても同様に，冠詞や前置詞などが単語列の最後に来ることや所有格が先頭に来ることは文法上不適切であると考えることができる．

そこで，人手で品詞情報を用いた規則を作成し，単語N-gramを作成する際にその規則を参照することにより，文法上成立しない単語列の作成を抑えることにした．

規則1

下記のリストにある単語が対訳表現の末尾にあればこの対訳表現を除外する

リスト

日本語：連体詞，接頭詞，接続詞，副詞，名詞-接続詞的

英語：冠詞，前置詞，助動詞，to，what，where，who，which，how

規則2

下記リストにある単語が対訳表現の先頭にあればこの対訳表現を除外する

リスト

日本語：動詞-非自立，動詞-接尾，形容詞-接尾，形容詞-非自立，名詞-非自立，名詞-動詞非自立，名詞-接尾，助詞，助動詞

英語：所有格

2.3 辞書情報の利用

さらに精度向上を図るために，2.2節までの手法で抽出した対訳表現に対して対訳辞書情報を用いて絞込みを行う．その処理手順は以下の通りである．(1)対訳表現の日本語から自立語を抽出する．取り出した自立語に対して辞書から英語訳語を取得する．ただし，サ変動詞に対しては形態素解析により分割された単語を再結合しておく．次に，これら自立語と付属語を含む日本語表現に対して連続単語列も単語とみなして単語n-gramを作成し，それに対しても辞書から英語訳語を取得しておく．(2)取得した英語訳語と対訳表現(の英語部分)の各単語と照合を行い，一致すればその元の日本語と英語のそれぞれの単語を「一致単語リスト」に分類し，一致する単語が一つもなければ元の日本語単語を「不一致単語リスト」に分類する．ただし，辞書から訳語候補を取得できない日本語単語に関してはどちらにも分類しないものとする．(3)一致単語数と不一致単語数を比較し，一致数の方が大きければ対訳表現の候補として取り出す．その中で一致数の最も多いペアを日英対訳表現として取り出す．

上記手順を【湾岸戦争の勝利⇔victory in the gulf war】を例に用いて詳しく説明する．手順(1)はまず，日本語表現「湾岸戦争の勝利」から自立語のみを取り出す．ただし，たとえば【湾岸戦争⇔gulf war】や【外務省⇔foreign ministry】のような対訳表現においては，2つ以上の単語がそろって初めてその意味を成すため，個々の日本語単語「湾岸」や「外務」からそれらの訳語候補を取得しても英語表現と一致させることができない．そのため，完全に対訳関係にあるペアであっても辞書情報を用いると対訳関係ではないと判定されてしまう可能性がある．よって，自立語を取り出すときに連続単語列については単語n-gramも作成しておく（ここではn-gramの上限は設けずにすべての単語n-gramを作成しておく）．しかし，上記で述べたn-gramを用いては辞書検索に多くの時間を費やすことになってしまう．そこで，2.2節で用いた規則を利用することにより，「の勝利」のようなn-gramを取り除く．上述方法によって得られた単語・単語n-gramは，自立語：「湾岸」，「戦争」，「勝利」の3個，単語n-gram：「湾岸戦争」，「戦争の」，「湾岸戦争の」，「戦争の勝利」，「湾岸戦争の勝利」の5個の合計8個で，それらが辞書検索の対象となる．次に，取り出した単語・単語n-gramに対して辞書検索を行い，訳語を取得する．手順(2)は，手順(1)によって取得された訳語と対訳表現の英語部分との照合・分類処理を行う．始めに自立語の訳語結果と英語表現の各単語との照合を行う．照合が一致すれば日英それぞれの単語を「一致単語リス

ト」に，一致する単語が一つもなければ日本語単語を「不一致単語リスト」に分類する．その結果，「湾岸」が「不一致単語リスト」に分類される．次に単語n-gramの訳語結果と英語表現との照合を行う．もし一致するものがあれば，不一致単語リスト内にその単語n-gramと部分一致する単語を一致単語リストに移動する．このような処理を行うことにより，自立語のみでは不一致単語リストに登録されてしまった「湾岸」や「gulf」などの単語を一致単語リストに戻すことができる．手順(3)は，一致単語リスト内の単語数と不一致単語リスト内の単語数とを比較し，一致数の方が多ければ，対訳表現候補であると判断し取り出し，対訳表現候補の中で一致数の最も多いものを日英対訳表現として取り出す．

3. 実験結果と考察

3.1 データ

NICTコーパス(4万文対)[3]とJENNADコーパス[4](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．対訳辞書は英辞郎を使用した．日本語と英語の形態素解析にはそれぞれChaSenとTreeTaggerを用いた．

3.2 対訳表現の抽出条件

文献[2]の実験結果を分析してみると，単語数7個以上の単語列とのペアのほとんどが不正解の対訳表現であったため，抽出する単語n-gramの上限を6-gramまでとした．また，日英の単語間の差が3個以上離れているペアにおいても，不正解の対訳表現ばかりであったため，そのようなペアを抽出しないこととする．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．また，英文に対しては複数形の表現を単数形に，現在形で3人称単数の動詞は原型に戻し，be動詞の現在形はisに過去形はwasに統一した．

3.3 対訳表現の評価方法

対訳表現の抽出数の精度は，抽出したすべての対訳表現から等間隔に500個取り出し，対訳関係にあるか否かを人手で評価した．また，文献[2]では「こと」などノイズから始まる対訳表現も人手で修正すれば正解としていたが，本研究では人手による修正をできる限り少なくするために，不正解とした．

3.4 対訳表現の抽出結果

対訳表現の抽出結果を表1にまとめる．表中の手法名は2.1節で述べた抽出手法に関連している．

表1：対訳表現の抽出結果

手法名	抽出数	精度	再現数
従来手法 1	208,856	0.10	21,721
従来手法 2	109,546	0.19	20,814
n-gram 改良	808,556	0.18	145,540
規則	481,396	0.35	168,489
辞書 1	125,165	0.96	120,518
辞書 2	170,000	0.90	153,000

従来手法1とはn-gramの上限を5-gramまでとした場合，従来手法2とはn-gramの上限を10-gramまでとした場合を表す．辞書1と辞書2に関する説明は後ほど行う．また，再現数は適切な対訳表現がどれだけ抽出できたかを見るための指標で抽出数と精度を掛けたものとした．

n-gram改良手法の抽出結果と従来手法のそれらと比べてみると，精度はそのままに抽出数のみを飛躍的に向上させることができた．このことにより，各n-gramまでで対訳表現を抽出し，重複しているものを取り除くことは，抽出数を向上させる上で有効であると考えることができる．ここで，従来手法では抽出することができず，n-gram改良手法によって抽出することができた対訳表現の例を図1に示す．

J1: 暗殺されたラビン首相
E1: prime minister rabin who was assassinated
J2: 不均衡を縮小するため
E2: to reduce imbalance
J3: 拉致疑惑問題
E3: the kidnapping issue
J4: 揶揄された
E4: have been ridiculed

図1：n-gram改良手法により抽出された例

n-gram改良手法では従来の手法に比べ，図1に示したような「VされたN」や「NをVするため」のような過去分詞句や不定詞句などのような様々なパターンの対訳表現を抽出することができた．したがって，本研究の大きなテーマである様々なパターンに対応できる英作文支援の開発においても大きな成果を挙げることができたと考えている．このような対訳パターンが従来手法では得られていなかった理由は，抑制処理によって，取り除かれたためである．

規則を用いた手法に関しては，抽出数はn-gram改良手法に比べると減少しているが，精度と再現数が

向上した．したがって，n-gram作成時に規則を用いて，文法的におかしい表現を削除することが，精度と再現数の向上に寄与することが確認できた．また，規則を用いた手法では，従来手法では抽出することのできなかった【寄与する必要がある⇔need to contribute】を抽出することができた．理由を検証していくと以下のようなことが分かった．【寄与する必要がある⇔need to contribute】は日本語表現の単語数が5個，英語側の単語数が3個と，日本語と英語で単語数の異なる対訳表現である．このような対訳表現を抽出するには，5-gramまでの連続単語列を取り出し，類似度を求める必要がある．しかし，従来手法とn-gram改良手法では，「need to contribute to」という表現が4-gramに存在したために，抑制処理により「need to contribute」という表現が取り除かれたために抽出することができなかった．しかし，規則を用いることによって最後に「to」がきている「need to contribute to」という表現は取り除かれたため，規則を用いた手法では抽出することができた．規則を用いる手法は精度と再現数の向上のみならず，上記のように日本語と英語で単語数が異なりなおかつ先頭または最後に不要語が付与されてしまったがために抽出できなかった表現にも対応することができることがわかった．

表中の辞書1は2.3節に述べた手法である．すなわち，規則を用いた手法によって抽出することのできた対訳表現に辞書情報を適用した手法である．実験の結果から精度が飛躍的に向上したことがわかる．この精度であれば，人手による修正を行わなくても，このまま対訳表現データベースとして英作文システムに組み込むことができる．しかし，表1からも分かるように再現数が規則を用いた手法に比べて減少していることが分かる．その原因は，辞書に掲載されている単語とは違う単語を対訳表現が多く使用している場合，人手で確認すると対訳表現であると判断することのできる表現も辞書情報では対訳表現ではないと判定される結果となったためであると考えられる．

辞書2は辞書1と同様，規則と辞書を用いた手法であるが，辞書1の手法と違い，日英それぞれの単語n-gram表現の抽出に抑制を加えず，かつ対訳表現の類似度計算をせずに抽出したすべての日英n-gramの組み合わせに対し規則と辞書を適用した．このような実験を行った理由は，例えば【アクセスを改善するために⇔to improve access】のような対訳表現ペアを取り出そうと考えてもimproveには「改善する」や「向上する」といった異なる対訳表現がコーパスに存在するため類似度計算を用いれば閾値を満たさず対訳表現として抽出することのできなかった対訳表現が多く存在したからである．また，本手法に規則を用いたのは，抽出の精度向上という本来の目的に加え，膨大な対訳表現からあらかじめ文法的におかしいものを取り除き，辞書検索や辞書検索結果と対訳表現間の照合処理の時間を短縮するという目的もある．実験の結果から，辞書1に比べ，処理時間が長く精度も多少落ちるが，より多くの対訳表現が抽出できることがわかった．

4. 終わりに

本稿では，対訳表現の抽出数と精度の向上を目的にこれまでの抽出手法の改良を試みた．単語n-gramを作成する際に規則を用いて文法上おかしい表現を取り除くことにより精度の向上を図り，また各単語n-gramまでで対訳表現の抽出をそれぞれ行い，抽出結果を1つにまとめ重複している表現を取り除くことにより対訳表現の抽出数を飛躍的に向上させることができた．更に，辞書情報を用いることにより精度を飛躍的に向上させることができた．実験の結果，n-gram改良手法の抽出数が約80万個で精度が0.18であった．それに規則を加えた場合，抽出数が約48万個で精度が0.35であった．さらに辞書情報を加えた場合，抽出数が約12万5千個で精度が0.96であった．また，対訳表現の抽出条件を緩和した状況で規則と辞書を適用した場合，対訳表現17万個を精度0.90で抽出できた．

今後は，本研究により抽出された対訳表現を利用して用例ベース翻訳的なアプローチにより英作文支援システムの開発に取り組んでいく予定である．

参考文献

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

[2] 坂上，馬，村田：英作文支援のための大規模な日英対訳表現の抽出，言語処理学会第１６回年次大会，pp.660-663 (2010)

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

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