高精度で実用的な英作文支援システムの開発を目指して

盛 竜太† 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

われわれは日英単語が混在する文（以降，混在文と呼ぶ）から，英文を自動生成する英作文支援システムの開発を目指している．その第一歩として，入力された混在文中の日本語を辞書引きし，得られた訳語候補の中から，高品質コーパス[1]及びWeb上での出現回数を調べ，最適と思われる訳語を選択するシステムを開発した[1]．また，訳語選択を行う上で高品質コーパス及びWeb上での出現回数を調べるための文脈情報（以降，クエリと呼ぶ）の各種構成手法の組み合わせ，更に高品質コーパスとWebの両方を使用することにより精度の向上を図れることを示した[2][3].

これらの研究[1]-[3]では，システムの精度を，「システムの出力した訳語が原文の単語と同じであるか」という基準で判定していた．つまり，“The no-nuke 運動 is as active as ever before.”という入力の“運動”に対する訳語として“movement”，“campaign”，“motion”，“move”などが考えられるが，この文の原文には“movement”が使用されていたため，これまでは“movement”のみを正解としていた．しかし他の訳語を見ると，“motion”と“move”は訳語として適切ではないが，“campaign”は訳語として適切であると考えられる．このように，日本語の訳語として，適切な訳語は 1 つとは限らず，1 つの日本語に対して適切といえる訳語は複数あると考えられる．そこで，プロの日英翻訳者による，訳語として「非常に適切である」「適切である」「文脈として適切ではないが，選択可能である」「選択することはできない」という４分類の評価データの作成と，その評価データを用いたシステムへの再評価実験を行った．その結果，80%の正解率が得られた.

また，これらの研究[1]-[3]では，英作文の支援対象を単語単位に限定していた．本研究では，その対象をフレーズ単位に拡張し，名詞句[2]と動詞句[3]も扱える英作文支援システムを構築した．これまで，対訳関係にないコーパスから複合名詞の対訳表現を獲得する方法[4]，対訳関係にあるコーパスから対訳パターンを獲得する方法[5]がそれぞれ提案されている．しかしこれらは複数の日英対訳表現を一括で「オフライン的に」獲得することを目的としており，大規模な対訳表現の獲得には大規模な２言語コーパスを必要とし，その対訳表現抽出にも莫大な処理コストを要することから，これらの手法は本研究のような，入力された混在文中の日本語部分の訳語表現を即時で「オンライン的に」出力することを目的とするシステムには適していない.

本研究で構築したフレーズ単位の英作文支援システムでは，与えられた混在文に対し，日本語部分を抽出し，それを茶筌[6]で単語分割する．次に分割したそれぞれの日本語単語を辞書引きし，得られた訳語候補を組み合わせる．最後に最適と思われる組み合わせを選択し出力とする．このシステムは，２言語コーパスを必要とせず単言語英語コーパスのみを用いて訳語選択を行うことができる．人手で作成した評価データを用いた評価実験の結果，名詞句は 71%，動詞句は 67%の正解率が得られた.

2. システム

2.1 入力

システムへの入力はそれぞれ，単語単位では“The no-nuke 運動 is as active as ever before.”，フレーズ単位では“宗教評論 has not been established as a discipline in this country.”のような混在文である.

2.2 日本語の分割

入力された混在文中の日本語を特定し，茶筌を用いて分割する．例えば，“宗教評論”という入力のときは“宗教”と“評論”に分割している.

一方，単語単位のときは，分割の必要はないが，2.3 に述べる辞書引きを行うために，単語の原形を求める必要がある.

[1] BNC コーパスなどのネイティブが作成した英文を集めたもの.

[2] ここでは，名詞及び形容詞が連続して出現するものを名詞句とする.

[3] ここでは，名詞と動詞，または，動詞と名詞が連続して出現するものを動詞句とする.

2.3 辞書

本システムでは和英辞書として，見出し語約 176 万語の英辞郎[7]を用いている．しかしこの和英辞書には，英語の品詞情報が載っていない．品詞情報は 2.4 に述べる訳語候補の限定処理に必要であるため，品詞情報を英辞郎の英和辞書を元に付与した．また，この和英辞書は例文が多く載っているという特徴があり，それらの例文は 2.4 に述べる訳語候補の限定に利用している．しかし，それらは訳語候補の取得には不要なものである．そこで，訳語候補の検索時間の短縮のために，例文の削除を行った．

2.4 訳語候補の取得

2.3 で述べたように整理した辞書を用いて，分割した単語の原形を辞書引きしその訳語候補を取得する．ただし，単語単位のときはこのように取得した訳語候補をそのまま用いるが，フレーズ単位のときは以下のように訳語候補の追加・限定を行っている．

名詞句に対する英作文支援では，入力された混在文中の日本語部分を分割したものに加え，分割した前の単語に“の”を付けたものでも辞書引きを行い，訳語候補に加えた．例えば，入力が“宗教評論”のときは，“宗教”と“宗教の”を辞書引きしたものを前の単語の訳語候補とし，“評論”を辞書引きしたものを後の単語の訳語候補としている．

動詞句に対する英作文支援では，入力された混在文中の日本語部分を分割したものに加え，分割した単語の動詞部分が複数の単語から構成されているときは，動詞部分の「1 つ目の単語」と，「1 つ目の単語と 2 つ目の単語を結合したもの」でも辞書引きを行い，訳語候補に加えた．例えば，入力が“国民に応答する”のときは，名詞部分として“国民”を辞書引きしたものを訳語候補とし，動詞部分として“応答”と“応答する”を辞書引きしたものを訳語候補としている．

このようにして取得した訳語候補に対し，さらに品詞情報を用い，限定を行っている．具体的には，名詞句と動詞句の名詞部分の訳語候補は品詞が形容詞と名詞の単語に限定し，動詞句の動詞部分の訳語候補は品詞が動詞の単語に限定している．このように品詞によって訳語候補を限定しているのは，名詞句と動詞句の名詞部分には基本的に形容詞と名詞しか使用されず，動詞句の動詞部分には基本的に動詞しか使用されないと考えられるためである．また，「高品質コーパス」「Web」「英辞郎の例文」での各訳語候補の出現回数により訳語候補の限定を行っている．このように出現回数により訳語候補を限定しているのは，訳語候補の数が多いと検索に時間がかかるためである．

2.5 クエリの構成

単語単位のクエリの構成には，[1]-[3]で提案された以下の 8 つの手法を用いている．①訳語候補のみ：訳語候補のみでクエリを構成する．②単語列：訳語候補にその前後 0～3 単語を結合してクエリを構成する．③品詞列：訳語候補にその前後 0～3 単語を品詞に置き換えたものを結合してクエリを構成する．④内容語：訳語候補の前後に存在する内容語で挟まれた最小の単語列でクエリを構成する．⑤ルール:人手で検索クエリを作成するときに見られる，いくつかの傾向をルール化したものを元にクエリを構成する．⑥長さ可変型：訳語候補の前後に単語を結合し，ヒット状況に従い徐々に単語を品詞に置き換えたり，単語を削除したりすることで，クエリを短縮する構成法である．⑦N グラム：まず訳語候補とその前 2 単語を結合したものでクエリを構成し，ヒット状況に応じて結合する単語を減らしていく構成法である．⑧統合手法：まずクエリ構成法①～⑦のいずれかでクエリを構成し，ヒット状況に応じて使用するクエリ構成法を変更させるものである．

フレーズ単位のクエリは，文脈情報を用いず，2.4 で述べた方法で取得した訳語候補を組み合わせて構成する．具体的には，名詞句のクエリ構成には，各訳語候補の全通りの組み合わせ」「単語の間に“of”を入れた全通りの組み合わせ」「前後の単語を逆にしてその間に“of”を入れた全通りの組み合わせ」の 3 つのルールを用いている．動詞句のクエリの構成には，「各訳語候補の全通りの組み合わせ」「前後の単語を逆にした全通りの組み合わせ」の 2 つのルールを用いている．また，名詞句及び動詞句の名詞部分が 3 単語以上に分割されたときは，分割された単語 2 つずつをペアとし，それぞれ別の問題として考える．例えば，入力が“経済構造調整機構”のときは，“経済”と“構造”と“調整”と“機構”に分割し，“経済構造”と“構造調整”と“調整機構”の 3 つの別々の問題として扱う．

2.6 クエリの検索

単語単位のときは，高品質コーパス及び Web でクエリ検索を行いその出現回数を調べる．フレーズ単位のときは，高品質コーパスのみでクエリ検索を行いその出現回数を調べる．

2.7 出力

調べた出現回数を元に，回答を出力する．単語単位のときは，出現回数 1 位の訳語候補を出力し，フ

レーズ単位のときは，出現回数 5 位までの訳語候補の組み合わせを出力している．5 位までを出力しているのは，フレーズ単位の英作文支援では，訳語候補の組み合わせによって回答を得ているが，訳語候補の組み合わせ数は多く，適切と思われる回答が 1 つとは限らない場合も多いと考えられるためである．また，ユーザ支援という視点からも，回答が 1 つしか存在していなくても上位 5 つの回答を示し，そのうち 1 つでも正解していれば役立つであろうとも考えられる．

3. 実験結果と考察

実験に用いた高品質コーパスは，BNC コーパス（約 605 万文），Wikipedia アブストラクト（約 200 万文），NICT コーパス[8]及び JENNAD[9]の英語データ（約 50 万文），The Dairy Yomiuri（約 25 万文）の計 900 万文であった．テスト問題は NICT コーパスから無作為に英文を取り出し，それらの各文に対して無作為に 1 単語を選び，日本語訳に置き換えて作成した．ただし，テスト問題作成に使用したデータは，高品質コーパスから除外した．このようにして，単語単位の問題として，原文の訳語候補の品詞が名詞 60 問，動詞 60 問，形容詞 30 問となるように，150 問の問題を作成した． 1 単語あたりの平均訳語候補数は 15 個であった．ただし，辞書引き不能な問題があったため，実際のテスト問題の数は 143 であった．フレーズ単位の問題として，名詞句用問題を 15 問，動詞句用問題を 15 問作成した．ただし，置き換えた日本語が 3 つ以上に分割されるときは別の問題と考えるため，実際の名詞句用問題は 21 問であった．

表 1 と表 2 は単語単位支援の再評価実験の結果を示す．表中の手法名は 2.5 で述べたクエリの構成方法であり，単語列及び品詞列の正解率は訳語候補の前後の単語数を 0 から 3 まで変化させたときのそれぞれの平均値である．また，統合手法には高品質コーパスと Web データ両方を使用しており，その末尾に付いている番号は文献[2]と同様，2 番目の統合手法を意味している．評価データはプロの翻訳者により 4 分類に作成されているが，ここでは「非常に適切である」と「適切である」の 2 分類にされている単語に「原文の単語」を加えたものを正解としている．正解率①は「正解問題数／全問題数」で算出したものであり，正解率②は「正解問題数／（全問題数－ヒット不足問題数）」で算出したものである．

これらの結果より，統合手法が 80.42%と最も正解率が高かったことがわかる．また，文献[2]の実験結果と比べると，各手法間の優劣関係が大きく変化することなく，全体的に正解率が 15%程度以上向上していることもわかる．つまり，プロの翻訳者により作成された，より合理的な評価データを用いた評価により，先行研究[1]-[3]の提案手法の有効性がさらに実証された．

表 1　高品質コーパスを用いた正解率（%）

手法名	正解率①	正解率②
訳語候補のみ	62.24	62.24
単語列	38.55	72.94
品詞列	62.83	70.06
内容語	25.87	71.15
ルール	65.03	72.09
N グラム	72.03	72.03
長さ可変型	69.93	69.93

表 2　Web データを利用した正解率（%）

手法名	正解率①	正解率②
訳語候補のみ	58.74	58.74
単語列	51.38	66.55
内容語	52.45	72.82
ルール	55.24	56.43
N グラム	64.34	64.34
統合手法②	80.42	80.42

次に表 3 に名詞句の英作文支援の正解率を示し，表 4 に動詞句の英作文支援の正解率を示す．これらフレーズ単位の英作文支援では，2.4 で述べた通り，フレーズを構成する各単語の訳語候補に対し，それらの出現回数による限定を行っている．そのため，手法名は限定方法で表し，具体的には「訳語候補の限定に使用したコーパス_訳語候補をそのコーパス上の出現回数の高い順から何位までに限定する」というようになっている．たとえば，英辞郎_2,3 は，訳語候補の限定に「英辞郎の例文」を使用し，訳語候補をそれらの「英辞郎の例文」上の出現回数の高い順から 2 位または 3 位までのものに限定するという 2 つの限定方法を意味している．ただし，数字 0 は限定処理なし（つまりすべての訳語候補を使用する）という意味で用いている．訳語候補の限定を行わなければ，名詞句問題の最大訳語候補数は 34 個，構成した最大クエリ数は 1564 個，動詞句問題の最大訳語候補数は 44 個，構成した最大クエリ数は 1140

個に上る. 正解率は, システムの出力の上位 5 位に, 人手で作成した正解データのフレーズがあれば正解として算出した.

表 3　名詞句の正解率（%）

手法名	正解率	手法名	正解率
英辞郎_0,1	52	高品質_2,5	57
英辞郎_2	67	高品質_3,4	52
英辞郎_3,4,5	71	Web_0,2,3	52
高品質_0	67	Web_1	38
高品質_1	43	Web_5	67

表 4　動詞句の正解率（%）

手法名	正解率	手法名	正解率
英辞郎_0,2,4,5	53	高品質_4,5	53
英辞郎_1	40	Web_0,4	60
英辞郎_3	47	Web_1	33
高品質_0	60	Web_2	40
高品質_1	20	Web_3	47
高品質_2,3	40	Web_5	67

これらの結果より，名詞句の英作文支援は「英辞郎_3,4,5」が 71%と最も正解率が高く，動詞句の英作文支援は「Web_5」が 67%と最も正解率が高かったことがわかる．また，訳語候補の限定を上位 5 個で行うときは，すべてのケースにおいて，最高の正解率が得られていることもわかる．実際，表には示していないが, 訳語候補の限定を上位 5-10 個にしても、すべてのケースにおいて，正解率のそれ以上の向上が見られなかった．したがって，訳語候補を限定するということは，クエリの数が減り，システムの実行時間の短縮につながり，正解率・時間の両方にとって有効であるということが言える．

4. 終りに

本研究ではまず，これまでわれわれが開発した単語単位の英作文支援システムに対し，プロの翻訳者が作成したより合理的な評価データを用いた再評価実験を行った．その結果，統合手法においては 80%の正解率が得られた．また，各手法間の優劣関係が従来と大きく変化することなく，全般的に正解率が 15%程度以上向上していることもわかった．これらより，われわれが先行研究[1]-[3]で提案した手法の有効性が実証された．本研究ではさらに，英作文支援の対象を単語からフレーズに拡張し，名詞句と動詞句も取り扱える英作文支援システムを構築した．計算機実験の結果，名詞句の英作文支援が 71%，動詞句の英作文支援が 67%という正解率が得られた．また，フレーズ単位の英作文支援を行うときに，その構成要素である単語の訳語候補を限定する方法は, 正解率の向上と処理時間の短縮の両方に有効であることがわかった．

今後は，単語と句だけでなく，支援できる日本語表現のパターンを増やしてしていくこと，英訳の処理時間を短縮させること，さらに，単語単位の支援と同様，各種日本語表現の英訳処理にも文脈を導入することや高品質コーパスと Web を融合利用することなど，英作文支援システムを性能と精度の両面から改良していく予定である．

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] 盛，馬，村田：高品質コーパスと Web データの統合的アプローチによる日英訳語選択，言語処理学会第 14 回年次大会，pp.281-284 (2008)

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

[4] 田中，松尾：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，電子情報学会論文誌，pp.2605-2614 (2001)

[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] 松本，今一，山下，北内，今村：日本語形態素解析システム 茶筌 version 2.2.9 使用説明書，奈良先端科学技術大学院大学 松本研究室(2001)

[7] 英辞郎：http://www.eijiro.jp/

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

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