大規模コーパスに基づく文脈可変型日英訳語選択

～日英混在型入力を対象とした英作文支援システムの開発～

中尾 孔一[†]　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

本研究では，日英単語が混在する入力文から，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，その第一歩として訳語選択を行なうシステムを開発した．開発したシステムは入力された日英混在文にある日本語単語を辞書引きし，そこから得られた複数の訳語候補から最も適切と思われる単語を出力（訳語選択）することになっている．訳語選択は，訳語候補とその前後数単語から構成される部分英単語列（以降，文脈情報または検索クエリと呼ぶ）のコーパス上での検索ヒット数を用いて行う．

高性能な訳語選択システムを実現するためには，高品質なコーパスと状況に応じて最適に作成される文脈情報の使用が重要不可欠である．しかしながら，従来提案されてきたシステム[1][2]においてはコーパスとして信頼性の低い Web データを用いていた．また，文脈情報もその構成要素の種類（語彙か品詞かなど）が固定であり，長さ（前後の単語の数）も固定または利用者によって決定されるようになっていた．そのため，訳語選択の精度が低い，または回答結果が利用者の英語習熟度に依存してしまうという問題がある．一方，提案システムでは，Web データに加えて英字新聞や英字図書などの信頼性の高い英語コーパス（約 900 万文，以降，高品質コーパスと呼ぶ）を収集し，利用することにした．高品質コーパスを利用することにより高性能な訳語選択の実現を可能とする一方，Web データを利用することによって高品質コーパスでのヒット数不足問題の解消が期待できる．さらに，訳語選択に用いる文脈情報の種類と長さを共に可変とし，コーパス上の検索ヒット状況に基づき自動的に決定するようにしており，さらなる精度向上を図っている．

計算機実験の結果，一単語あたりに平均 15 個の訳語候補を持つ 150 の訳語選択問題において，高品質コーパスを用いた場合，提案手法は文脈固定の従来手法よりも約 10%高い 51.39%の訳語選択精度を得られた．また，高品質コーパスと Web データの両方を利用する統合手法ではさらに諸提案手法の中でもっとも高い 52.08%の精度を得ることができた．

２．システムの概要

本システムの処理概要を図１に示す．

2.1 入力

本システムへの入力は日英混在の文であり，図１の例では「He didn't know 結果 of the meeting.」となっている．

2.2 辞書引き

ここでは入力された文から日本語単語を取り出し，その単語に対する訳語候補単語を和英辞書から取得する．

2.3 検索クエリの構成

このステップでは，個々の訳語候補に対し，その訳語候補とその前後にある英単語を用いて検索クエリを構成する．どのような検索クエリを用いるかは本研究のもっとも重要な部分であり，検索クエリの構成手法については次章で詳しく述べる．

2.4 検索

2.3 で得られた個々の検索クエリで大規模コーパスへの検索を行ない，それぞれの検索ヒット数を取得する．

2.5 回答

2.4 で取得したヒット数を元に，一番ヒット数が多い訳語候補をシステムの回答として出力する．例では一番ヒット数が多い"outcome"を回答としている．

3. 検索クエリの構成手法

従来の訳語選択では検索クエリは訳語候補と利用者（または開発者）が指定した範囲内の単語列を用いて構成するようにしていた．そのため，利用者の英語力によって結果に違いが生じる．また，使用する範囲を固定するような方法では不必要な単語が検索クエリに混じり，良い結果が得られない可能性が高くなってしまう．本節ではまず，検索クエリの従来の基本構成を踏まえながらいくつかの改善を施した構成手法をベースライン

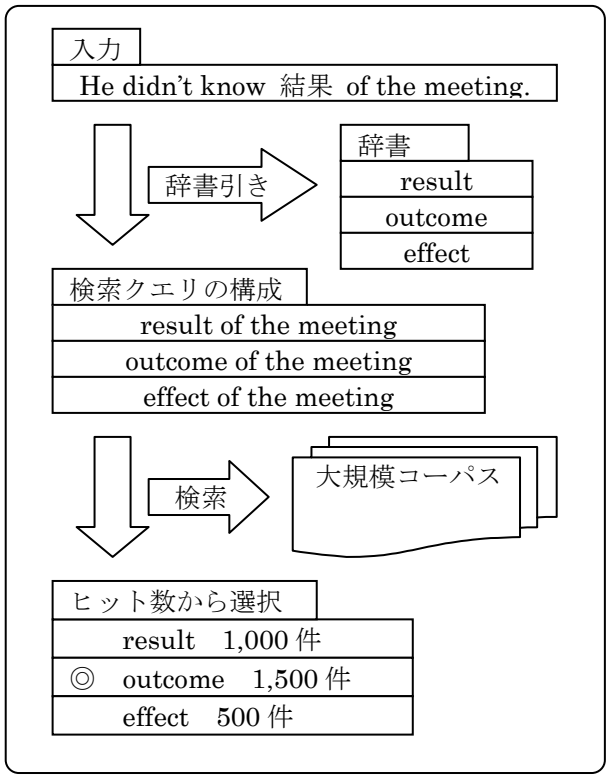

図 1 システムの処理概要

として述べる．それらは提案手法との比較に必要なだけでなく高品質コーパスと Web データの両方を利用する統合手法にも用いられている．次に，検索クエリの範囲などを固定とせず，コーパス上の検索ヒット状況などに基づき自動的に決定する提案手法について述べる．

3.1 ベースライン手法

3.1.1 訳語候補のみを用いる

この方法では検索クエリは訳語候補のみで構成される．すなわち，コーパスにおいて最も出現回数の多い訳語候補を回答とする．

3.1.2 固定長の単語列を用いる

検索クエリは訳語候補とその前後の指定した範囲内の単語列で構成される．しかし単語数を多くすると，検索ヒット数が大幅に減少し，結果が悪くなってしまうことが予想される．そこで以下のような 2 つのルールを加えることにした．

ルール 1

指定した範囲内の単語であってもそれが記号であり、それが訳語候補の左側（右側）にあるなら，その記号，及びそれより左（右）にある単語は検索クエリに含めない．

ルール 2

訳語候補から 2 つ以上離れた単語は品詞に置き換える．

3.1.3 固定長の品詞列を用いる

検索クエリの構成にすべて品詞を用いる以外は 3.1.2 と同じである．

3.1.4 内容語で挟まれた部分を用いる

これは[3]で提案された手法である．訳語候補を中心とし，前後に存在する内容語（名詞，動詞，形容詞）で挟まれた最小の単語列を検索クエリとして用いる．この手法は検索クエリが長くなる傾向があり十分なヒット数が得られない可能性が高いため，「内容語以外を品詞に置き換える」という修正を加えた．

3.2 提案手法

3.2.1 ルールを用いる

人手で検索クエリを作成する場合，いくつかの傾向が見られる．たとえば，訳語候補が名詞の場合，周囲に動詞があればそれを使用し，動詞と名詞をセットにしたクエリを作るなどである．本提案手法では，それらの傾向をルール化し，それにしたがって検索クエリを構成する．

訳語候補が動詞以外の場合のルール

1.訳語候補を含む句が名詞句であれば，その名詞句に含まれる単語を全て用いて検索クエリを構成する．ただし，訳語候補以外の単語が冠詞のみであった場合（ここでは「例外」と呼ぶ）これを含めず，2 のルールを適用する．
2.訳語候補を含む句が名詞句でなければ（または上記「例外」の場合），訳語候補の前後の動詞の位置を調べ，近い方の動詞と訳語候補の間にワイルドカードを挿入し検索クエリを構成する．

訳語候補が動詞の場合のルール

1.訳語候補の後方を調べ，名詞があれば，訳語候補と名詞の間にワイルドカードを挿入し検索クエリを構成する．
2.後方に名詞がなければ，訳語候補の前方を調べ，名詞があれば名詞と訳語候補の間にワイルドカードを挿入し検索クエリを構成する．

3.2.2 検索クエリの長さ可変型

通常，検索クエリに含まれる単語数が多いほど手がかりが増え，より信頼性の高い結果が得られる．しかし単語数が多くなると検索ヒット数が激減し，結果的に精度の低下を招いてしまう．そこで検索クエリを構成する範囲を最初に大きく設定しておき，そこから徐々に範囲を短縮していく手法が考えられる．範囲を短縮する場合，訳語候補の前後どちらを優先して短くしていくかを選

1. 訳語候補とその前の Nf 個の単語と後ろの Nb 個の単語列で検索クエリを構成し，検索を行う
2-a（ヒットした場合）ヒット件数が一番多い訳語候補を回答とし，処理を終了
2-b（ヒットしなかった場合）ステップ 3 に進む

3. 検索クエリ中の（訳語単語以外の）単語の中から（Nf + Nb）/ 2 個の単語を品詞に置き換えて検索を行なう．ただし，置き換える単語はコーパス内の出現回数が多いものを選ぶ
4-a（ヒットした場合）ヒット件数が一番多い候補単語を回答とし，処理を終了
4-b（ヒットしなかった場合）ステップ 5 に進む

5. Nf か Nb のいずれかを選んで 1 減らす．ただし，
左端の単語と訳語候補とその間の単語の品詞で新たに構成された検索クエリと，
訳語候補と右端の単語とその間の単語の品詞で新たに構成された検索クエリと
のヒット数を比べ，少ない方を選ぶ

6. ステップ 1 に戻る

図 2 検索クエリの長さ可変型手法のアルゴリズム

択する必要がある．ここではその決定を，左端(または右端)の単語と訳語候補とその間の単語の品詞で構成した検索クエリでの検索ヒット数を用いて行う．具体的にはヒット数の少ない方，すなわち訳語候補と関係性が低い方が優先して短縮するようにしている．具体的な処理のアルゴリズムを図 2 に示す．

3.2.3 統合手法

まず，上記すべての手法の中で高品質コーパスを利用した場合でもっとも有効な手法を用いて高品質コーパスに対し検索を行いヒット数不足でなければそのヒット数を用いて訳語候補を決定する．次に，高品質コーパス利用ではヒット数不足になった問題に対しては，上記手法で Web データを利用した場合でもっとも有効な手法を用いて訳語候補を決定する．

4. 実験結果と考察

実験に用いた高品質コーパスは，英字読売新聞 The Daily Yomiuri（約 25 万文），NICT 日英対訳コーパス[4]の英語データ(約 50 万文),Wikipedia アブストラクト（約 200 万文），BNC 英語コーパス（約 605 万文）の計 900 万文であった．一方，Web データはあらかじめ収集して使用するのではなく直接 Google 検索を行いそのヒット数を用いた．また,和英辞書は見出し語約 176 万語の英辞郎[5]を用いた．英単語の品詞情報の取得には SS Tagger[6]，名詞句などの特定には SS parser[7]を用いた．テスト問題は NICT 日英対訳コーパスから無作為に 150 の英文を取り出し，それらの各文に対し無作為に 1 個の単語（ただしその正解訳語候補の品詞がそれぞれ 60 個の名詞，60 個の動詞，30 個の形容詞になるように）を選んで正解日本語訳に置き換えて作成した．作成されたテスト問題において 1 単語あたりの平均訳語候補数は 15 個であった．

表 1 は高品質コーパスを用いた各手法の正解率を示す．高品質コーパスを用いた場合，コーパス規模が相対的に小さいためヒット数不足（本実験ではヒット数がゼロの場合をヒット数不足とした）の問題が生じる．そのため正解率はヒット数不足のテスト問題を含む場合と含まない場合の二通りの方法で算出した．ヒット数不足を含まない正解率は正解数を，全問題数からヒット数不足問題数を引いたもので割った値である。これを求めた理由は，Web データを利用することでヒット数不足の問題はほぼ解消されるだろうと考え，ヒット数不足要素を排除した手法間の優劣を計るためである．また，3.1.2 と 3.1.3 の手法は検索クエリを構成する訳語候補の前後の範囲を指定する必要があるので，その前後の単語の数を 0～3 に変化させ，全 16 通りの実験を行ないそれらの結果の中で最も正解率が高かった値をその手法の正解率とした．また，検索クエリの長さ可変型手法の初期値は Nf = 3, Nb = 3 とした．

表 1 高品質コーパスを用いた各手法の正解率

	ヒット数不足含む	ヒット数不足含まない
訳語候補のみ	34.72%	34.72%
固定長単語列	42.36%	**57.78%**
固定長品詞列	41.76%	45.67%
内容語	26.39%	55.78%
ルール	45.14%	47.79%
長さ可変型	**51.39%**	51.39%

「ヒット数不足を含む」結果を見ると，提案手法の「検索クエリの長さ可変型手法」が最も良い精度を出していることが分かる．また，同じく提案手法である「ルールを用いる手法」も他のベースラインに比べて良い結果が得られている．しかし「検索クエリの長さ可変型手法」は精度の高い手法ではあるが，「ルールを用いる手法」に比べ実行時間が長くなる傾向がある．これは長さ可変を行なう処理ループに大半の時間を使っているからである．処理スピードを重視するならば，ルールを用いる手法を採用するべきである．

次に「ヒット数不足を含まない」結果を見ると，「固定長単語列を用いる手法(前1語,後ろ3語)」の正解率が最も高かった．ヒット数不足の問題はWeb データを用いることで解決できることを考えれば，この手法を Web データに適用すれば良い結果が得られることが期待できる．それを確認するため各手法に対し Web データを利用した場合の正解率（ヒット数不足を含む）を求め，表 2 にまとめた．ただし，Web データに対しては品詞検索が行なえないため，固定長品詞列の手法は実験していない。また，長さ可変型では品詞置き換えをせず、そのままの単語で検索している。

表 2　Web データを利用した各手法の正解率

	ヒット数不足含む
訳語候補のみ	29.17%
固定長単語列	37.50%
固定長品詞列	――――――
内容語	22.22%
ルール	**47.92%**
長さ可変型	24.31%

前述の予想に反し，Web データを利用した場合は（「固定長単語列を用いる手法」ではなく）「ルールを用いた手法」が最も正解率が高かった．

そこで，高品質コーパスに対しヒット数不足を含まない場合のもっともよい手法の「固定長単語列を用いた手法」と Web データに対しもっともよい手法の「ルールを用いた手法」との「統合手法」を用いた実験を行った．結果を表 3 に示す．ここでは比較のため，高品質コーパスと Web データの両方ともに同じ「固定長単語列を用いた手法」を適用した手法を統合手法のベースラインとした（「固定長単語列を用いた手法」は高品質コーパスに対し，ヒット数不足を含まない場合にもっとも高い精度を出した手法である）．高品質コーパスを用いた場合，ヒット数不足の問題は 54 問あり，それらについて Web データを利用した．表 3 から「統合手法」が予想通りすべての手法の中で最も高い精度を出したことが確認できた．

表 3　統合手法の正解率

	ヒット数不足含む
ベースライン	45.14%
統合手法	**52.08%**

5. おわりに

本研究では日英単語が混在する入力文から，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，その第一歩として新しい訳語選択手法を提案し，GUI ベースの日英訳語選択システムを実装した．計算機実験の結果，一単語あたりに平均 15 個の訳語候補を持つ 150 の訳語選択問題において，高品質コーパスを用いた場合，提案手法は文脈固定の従来手法よりも約 10%高い 51.39%の訳語選択精度を得られた．また，高品質コーパスと Web データの両方を利用する統合手法ではさらに諸提案手法の中でもっとも高い 52.08%の精度を得ることができた．今後はより大規模なテスト問題を作成しオープンテストを行うことにより手法の有効性を検証するとともに，訳語選択性能のさらなる向上を図る．さらに，英作文支援の対象を単語からフレーズや節などの英語表現に拡張し，システムの実用化を目指す．

参考文献

[1] 大鹿，佐藤，安藤，山名：Google を活用した英作文支援システムの構築，電子情報通信学会データ工学ワークショップ，4B-i8 (2005).

[2] 佐藤，安藤，山名：検索エンジンを利用した英作文支援システムの構築，言語処理学会第 12 回年次大会，pp.664-667 (2006).

[3] 隅田，菅谷，山本：英語能力測定のための空所補充問題の自動生成手法，電子情報通信学会信学技報，TL2004-22 (2004).

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

[5] 英辞郎：http://www.eijiro.jp/

[6] Tsuruoka，Tsujii : Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data，HLT/EMNLP，pp.467-474 (2005).

[7] Tsuruoka，Tsujii : Chunk Parsing Revisited，Proceedings of the 9th International Workshop on Parsing Technologies ，pp.133-140 (2005).