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

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ている。その第一歩として、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に対し、訳語候補とその前後数単語から構成される英単語列（以降、検索クエリと呼ぶ）のコーパス上での検索ヒット数から最も適当と思われる単語を選択（訳語選択）するシステムを開発した[1]。検索クエリの構成手法として従来手法などのベースラインに加え、可変型とルールベース型を提案した。訳語選択は、提案手法で構成した検索クエリを用い、高品質英語コーパス、Web、さらにその両方での検索ヒット数を調べることによって行った。高品質コーパスと Web の両方を利用する統合手法は、高品質コーパスでヒット数不足のときに、Web データを用いることにより、ヒット数不足を補うものであった。

本研究では、訳語選択のさらなる精度向上を図るために、[1]に導入した高品質コーパスでヒット数不足のときに Web 検索を使用する、という２手法間に限定した統合手法に留まらず、高品質コーパスの利用と Web 検索と各種検索クエリの構成手法の種々の組み合わせによる統合手法の性能を調べた。計算機実験の結果、4～5 手法の組み合わせによる統合手法は[1]に用いた 2 手法間の最適な組み合わせよりも概ね 5%以上精度が高いことがわかった。また、むやみに手法を増やしても効果があまり得られないこともわかった。以上より、多少の取捨選択が必要であるが、手法間、コーパス間の多手法による統合アプローチは訳語選択の精度向上に有効であることがわかった。

２．システムの概要

本システムではまず、「The no-nuke 運動 is as active as ever before」のような日英混在入力文に対し、日本語部分「運動」を辞書引きし、訳語候補（motion, exercise, sport, campaign, movement など)を取得する。そして、個々の訳語候補に対し、訳語候補の前後にある英単語を用いて検索クエリを構成し、高品質コーパス・Web 検索を行い、検索ヒット数を取得する。検索ヒット数が最も多い訳語候補(今の例の場合は movement)をシステムの回答として出力する。

Web 検索は直接 Google を利用する。一方、高品質コーパスへの検索は独自に構築した検索システムを用いる。そのシステムは品詞やワイルドカードなどを用いた検索も可能なため、検索クエリは非常に柔軟に構成することができる。

３．検索クエリの構成手法

3.1 先行研究[1]に用いた手法

(1)訳語候補のみ

訳語候補のみで構成するものである。

(2)単語列

訳語候補とその前後 0～3 の単語列で構成するものである。

(3)品詞列

(2)の単語列手法の訳語候補以外の単語を、品詞に置き換えて検索クエリを構成するものである。

(4)内容語

検索クエリを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構成するものである。

(5)ルール

人手で検索クエリを作成する場合に見られる、いくつかの傾向をルール化し、そのルールを元に検索クエリを構成するものである。

(6)長さ可変型

初めに訳語候補の前後に多めの単語（今回は前後それぞれ 3 つ）を結合したものを検索クエリとし、そこから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、クエリの長さを短縮する手法である。

3.2 N-gram に基づく手法

N-gram に基づく手法は訳語選択など機械翻訳

分野によく用いられる手法である。そこで、この伝統的な手法と比較するために、本稿では一種の簡易的なスムージングを加えたN-gramに基づく手法を新たに追加した。これは以下のようなものである。まず、訳語候補とその前の2つの単語を検索クエリとし（trigram）、検索ヒット数が0件であれば、訳語候補とその前の1つの単語を検索クエリ（bigram）、さらに訳語のみを検索クエリとする（unigram）というものである。ただし、この手法は単純に検索クエリの長さを変化させているだけのため、3.1.1 に述べた長さ可変型より性能的に劣ると予想される。

４．訳語選択の統合手法

本研究では、統合手法に使用する訳語選択手法の数を、先行研究[1]のように2つに限定するのではなく、高品質コーパスと Web 検索と 3 節に述べた各種検索クエリの構成手法を組み合わせ、さまざまな統合手法を試みた。統合手法の基本的な考え方は先行研究に用いた統合手法のそれと同様、「まずはすべての問題に対し信頼度の高い訳語選択手法を優先的に適用する。そして、検索ヒットしない問題に対し、その次に信頼度の高い訳語選択手法を優先的に適用していく。…」ということである。したがって、統合手法の作成は、「高品質コーパスまたは Web データに 3 節で述べた各手法で構成した検索クエリを適用する」といった単一手法の実験結果を参考に行った（単一手法の実験結果も統合手法のそれと合わせて次節に示す）。以降の統合手法の説明においては、手法名の後ろに「Web」とついているものが、Web 検索を使用するもの、手法名の後ろに何もついていないものが、高品質コーパスを使用するものとする。また、検索クエリ作成法の単語列及び、品詞列の前にl語、後にr語の単語を結合するときは、単語列(l, r)、品詞列(l, r)と表記することとする。

4.1　統合手法①

高品質コーパスを使用した訳語選択手法の中で最も正解率（ヒット数不足含まない）が高かった内容語と、Web 検索を使用した訳語選択の中で最も正解率（ヒット数不足含む）が高かった単語列(0, 2)を用いた。すなわち、「高品質コーパス＋内容語」と「Web 検索＋単語列(0,2)」を順に実行する手法である。これは２手法間の最適な統合手法と思われる。

4.2　統合手法②

まず、各訳語選択手法を正解率（ヒット数不足含まない）の降順（正解率が等しいものは、ヒット不足率が大きいものが上位）に並べる。次に、降順に並べたものから、「正解率が 1 つ上の手法よりもヒット数不足率が多い手法を取り除く」ということを繰り返す。このようにして得られる、訳語選択手法の順番は、正解率の高いものから、徐々にヒット数不足率が低くなっていくようになっており、高い正解率が得られることが予想される。このようにして得られた訳語選択手法の実行順番は「**単語列 (3, 3)Web＋内容語＋単語列(2, 3)Web＋単語列(1, 2)＋単語列(0, 3)＋単語列(2, 2)Web＋単語列(2, 0)＋ルール＋N グラム＋単語列(0, 2)Web**」である。

4.3　統合手法③、統合手法④

統合手法②は使用する手法が 10 個と多く、実行するのに非常に時間がかかる。そこで使用する訳語選択手法を減らした「**単語列(3, 3)Web＋内容語＋単語列(2, 3)Web＋ルール＋単語列(0, 2)Web**」（統合手法③）も実験に使用する。手法の取捨選択は、「正解率及びヒット数不足率が 1 つ上の手法との違いが小さいものを取り除く」という基準で行った。

また、統合手法③から使用する手法を更に減らした「**単語列(3, 3)Web＋内容語＋ルール＋単語列(0, 2)Web**」（統合手法④）も実験に使用する。

4.4　統合手法⑤、統合手法⑥

内容語と単語列(2, 3)Web は正解率が等しく、ヒット数不足が内容語の方が高いという違いだけであるため、統合手法③の内容語と単語列(2, 3)Web の実行順番を入れ替えた「**単語列(3, 3)Web＋単語列(2, 3)Web ＋内容語＋ルール＋単語列(0, 2)Web**」（統合手法⑤）も実験に使用する。

統合手法⑤から使用する手法を減らした「**単語列(3, 3)Web＋単語列(2, 3)Web＋ルール＋単語列(0, 2)Web**」（統合手法⑥）も実験に使用する。

4.5　統合手法⑦、統合手法⑧

統合手法④から使用する訳語選択手法を更に減らした「**内容語＋ルール＋単語列(0, 2)Web**」（統合手法⑦）も実験に使用する。

統合手法⑦から使用する訳語選択手法を更に減らし、高品質コーパスのみを使用した「**内容語＋ルール**」（統合手法⑧）も実験に使用する。

4.6　統合手法⑨～統合手法⑭

統合手法①～統合手法⑦は、統合手法の最後に

全て、単語列(0, 2)Web を用いてきた。しかし、単語列(0, 3)Web は、ヒット数不足を含まない正解率が、単語列(0, 2)Web よりも高いが、ヒット数不足率も高くなっているため、統合手法に使用していない。しかし、この 2 つのヒット数不足を含まない正解率は等しいという結果が得られているので、**統合手法①～統合手法⑦（②を除く）の最後に、単語列(0, 2)Web を用いるのではなく、単語列(0, 3)Web を用いたもの**を統合手法⑨～統合手法⑭とする。

5．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 Dairy Yomiuri（約 25 万文）、自動対応付けされた日英コーパス JENNAD[2]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[3]を用いた。英単語の品詞情報の取得には SS Tagger[4]、名詞句などの特定には SS parser[5]を用いた。テスト問題は NICT 日英対訳コーパス[6]から無作為に 150 の英文を取り出し、それらの各文に対して無作為に 1 個の単語（ただしその正解訳語候補の品詞がそれぞれ名詞 60 個、動詞 60 個、形容詞 30 個になるように）を選び、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。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。

実験ではヒット数不足（本実験ではヒット数が 0 件の場合のみをヒット数不足とした）を含む正解率と、ヒット数不足を含まない正解率という 2 つの正解率を用いて各手法を評価した。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の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、または複数の訳語選択手法を実行することで、ヒット数不足の問題はほぼ解決されると考え、ヒット数不足要素を排除した手法間の優劣を測るためである。

表 1 と表 2 に高品質コーパスと Web データを用いた各手法の正解率を示す。単語列及び品詞列を使用する各手法は、統合手法に使用する手法（すなわち各検索方法のそれぞれの正解率が最も高い手法）の正解率のみを示している。単語列及び品詞列を使用する各手法のヒット数不足を含む正解率は、訳語候補に結合する単語数が、単語列は 2 個以下、品詞列は 4 個以下、単語列 Web は 3 個以下のときに、それぞれ高い正解率が得られており、訳語候補に結合する単語列が、これら以外の結合数のときは、正解率が急に低くなっている。単語列及び品詞列を使用する各手法の、ヒット数不足を含まない正解率は、訳語候補に結合する単語数に関わらず、単語列は 47%～60%、品詞列は 41%～52%となっている。単語列 Web のヒット数不足を含まない正解率は、訳語候補に結合する単語数が増えるにつれて高くなっている。

表 1 高品質コーパスを用いた各手法の正解率

手法名	ヒット数不足含む	ヒット数不足含まない
訳語候補のみ	40.56%	40.56%
単語列(0, 3)	30.07%	56.58%
単語列(1, 0)	48.95%	50.36%
単語列(1, 2)	28.67%	58.57%
単語列(1, 3)	18.88%	60.00%
単語列(2, 0)	44.06%	55.26%
品詞列(2, 1)	48.25%	50.74%
内容語	22.38%	61.54%
ルール	49.65%	55.04%
N グラム	53.15%	53.15%
長さ可変型	52.45%	52.45%

表 2　Web データを利用した各手法の正解率

手法名	ヒット数不足含む	ヒット数不足含まない
訳語候補のみ	40.56%	40.56%
単語列(0, 2)	49.65%	49.65%
単語列(0, 3)	49.65%	52.99%
単語列(2, 2)	37.76%	56.25%
単語列(2, 3)	27.97%	61.54%
単語列(3, 3)	18.18%	70.27%
品詞列		
内容語	39.16%	54.37%
ルール	37.06%	37.86%
N グラム	40.29%	41.26%
長さ可変型		

高品質コーパスを使用する手法において、今回追加した N-gram 手法が予想に反し、先行研究[1]で提案した長さ可変手法より精度が若干、高かった。しかし、この 2 つの手法は正解の問題数が、

1 問しか変わらないため、訳語選択の精度はさほど変わらないといえるが、訳語選択に必要な時間は、長さ可変型手法の方が長くなっている。

表 3 統合手法を使用した各手法の正解率

手法名	手法数	ヒット数不足含む	ヒット数不足含まない
統合手法①	2	57.34%	57.34%
統合手法②	10	62.94%	62.94%
統合手法③	5	62.94%	62.94%
統合手法④	4	60.84%	61.54%
統合手法⑤	5	63.64%	63.64%
統合手法⑥	4	61.54%	61.54%
統合手法⑦	3	58.04%	58.04%
統合手法⑧	2	53.15%	57.58%
統合手法⑨	2	58.04%	61.03%
統合手法⑩	5	64.34%	64.34%
統合手法⑪	4	60.84%	60.84%
統合手法⑫	5	65.03%	65.03%
統合手法⑬	4	62.94%	62.94%
統合手法⑭	3	59.44%	59.44%

表 3 に各統合手法の正解率を示す。各統合手法の結果を見ると、予想に反して、訳語選択手法を 10 個使用した統合手法②ではなく、訳語選択手法を 5 個使用した統合手法⑫の結果が最も良いものとなった。この結果から、多くの訳語選択手法を使用した方が、良い結果が得られる傾向にあるが、訳語選択手法を多く用いれば用いるほど、良い結果が得られるわけではないということがわかる。2 番目に結果が良かったものは、統合手法⑩であった。これは、最も結果が良かった統合手法⑫と、「単語列(2, 3)Web と内容語」の実行順序が逆になっているだけであり、4.4 節で述べた通り、単語列(2, 3)Web と内容語の結果には、あまり違いがないので、この統合手法⑩と統合手法⑫の組み合わせが、良い組み合わせであるといえる。

全体的に見て、統合手法の最後に、単語列(0, 2)Web を用いるのではなく、単語列(0, 3)Web を用いたものの結果が良い結果になっている。4.6 で述べた通り、単語列(0, 2)Web と単語列(0, 3)はヒット数不足を含む正解率が等しく、ヒット数不足を含まない正解率が、単語列(0, 3)Web の方が高くなっている。この 2 つの手法のうち単語列(0, 3)Web を用いた統合手法の結果が良くなっているのは、多くの訳語選択手法を実行した後に、単語列(0, 3)Web を実行しているために、単語列(0, 3)Web 単独で実行したときに、ヒット数不足になっていた問題が、単語列(0, 3)Web 以前の手法で、結果が出されていたためと考えられる。つまり、統合手法では、訳語選択手法を単独で実行する時とは違い、単語列(0, 3)Web を用いた方が、良い結果が得られるということである。

6．終りに

本研究では、われわれが開発した訳語選択システム[1]の性能向上を図るために、高品質コーパスの利用と Web 検索と提案した各種検索クエリの構成手法の種々の組み合わせによる統合手法の性能を調べた。計算機実験の結果、4～5 手法の組み合わせによる統合手法は[1]に用いた 2 手法間の最適な組み合わせよりも概ね 5%以上精度が高いことがわかった。また、むやみに手法を増やしても処理時間が増大するだけでなく効果もあまりないことがわかった。以上より、多少の取捨選択が必要であるが、手法間、コーパス間の多手法による統合アプローチは訳語選択の精度向上に有効であることがわかった。今後はまず新しい問題セットを用いた確認実験を行った後、訳語選択の基本手法とした最適な統合手法を確立したい。次に日英辞書を改良するなど他のアプローチによる訳語選択の性能向上を図りたい。さらに英作文支援の対象を単語から複合語、句、節などの表現へ拡張し、システムの実用化を目指したい。

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 (2006)

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

[3]英辞郎：http://www.eijiro.jp/

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

[5]Tsuruoka and Tsujii: Chunk Parsing Revisited, IWPT2005, pp. 133-140 (2005)

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