Vol.2009-NL-192 No.7
2009/7/22

日英単言語 Web コーパスからの対訳 treebank 自動獲得

後藤 功雄[†1]　内元 清貴[†1]
河原 大輔[†1]　鳥澤 健太郎[†1]

大規模な日本語と英語の単言語 Web コーパスから，動詞を含む頻出部分構文木を探索して対訳辞書を用いて対応付けし，対訳の部分構文木（対訳 treebank）を幅広く獲得する手法を提案する．提案手法を用いた実験で，日本語 36 億文，英語 15 億文の単言語 Web コーパスから動詞と目的語を含む対訳の頻出部分構文木を幅広く獲得した．獲得した対訳を利用した機械翻訳の実験により，NTT 機械翻訳機能試験文の翻訳において自動獲得した対訳の有効性が確認された．

Automatic acquisition of a bilingual treebank from monolingual web corpora in Japanese and English

Isao Goto,[†1] Kiyotaka Uchimoto,[†1] Daisuke Kawahara[†1]
and Kentaro Torisawa [†1]

This paper proposes a method to acquire bilingual partial syntactic trees (bilingual treebank) by searching and aligning frequent partial syntactic trees including verbs using bilingual dictionaries from large-scale monolingual web corpora in Japanese and English. In our experiment using the proposed method, we widely acquired bilingual frequently partial syntactic trees including verbs and objects from a monolingual Web corpus consisting of 3.6 billion sentences in Japanese and a monolingual Web corpus consisting of 1.5 billion sentences in English. Our experiments of machine translation using the acquired bilingual treebank show the effectiveness of the bilingual treebank for translation of NTT MT test set.

†1 情報通信研究機構　知識創成コミュニケーション研究センター
NICT Knowledge Creating Communication Research Center

1. はじめに

近年，統計翻訳や用例翻訳など対訳コーパスに基づく機械翻訳の研究が盛んに行なわれ，着実に成果を上げている[12),13)]．この対訳コーパスに基づくアプローチでは，一般に，どんな言語ペアやドメインでも，対訳コーパスと各言語の解析器さえあれば機械翻訳システムを構築できるという利点がある．しかし，十分な性能を得るためには各言語ペア，ドメインごとに大量の対訳コーパスが必要であり，あるドメインの対訳コーパスで学習したシステムを新しい別のドメインに適用しようとすると，そこで使われる語彙や表現が異なるために，翻訳性能の低下を引き起こすことが多い．特に構文構造の中心となる動詞とその格要素の組の出現傾向がドメインにより異なる点が，性能低下の主因の一つである．例えば，日常的に我々がよく耳にする表現で，NTT 機械翻訳機能試験文にも現れる「ボタンを付ける」,「薬を塗った」,「茶を入れた」などは，読売新聞の対訳コーパス 25 万文[10)] には現れない．そのため，読売新聞の対訳コーパスで機械翻訳システムを学習してもこれらの表現はうまく訳出できない．したがって，動詞とその格要素の組を含む対訳をいかに集めるかが，翻訳性能低下を軽減するための鍵になると考える．本稿では，動詞とその格要素の組を含む対訳を自動獲得するための新しい手法を提案する．

提案手法の着想に至った経緯は次の通りである．一般に，対訳コーパスを人手により構築するには膨大なコストがかかる．そのため，対訳コーパスを自動構築する方法[2),6),7),10)] が多く提案されてきた．しかし，これらの方法ではコンパラブルテキストを必要とするため，動詞とその格要素の組を網羅するのは困難である．一方，動詞とヲ格の格要素の組のうち，上記の読売新聞コーパスや NTT 機械翻訳機能試験文に現れるものに着目すると，そのほとんど（予備調査では少なくとも 95%以上）は World Wide Web（Web）に出現している．さらに，そのうち半数程度は Web コーパス（クローリングして集めた Web のテキスト）に 100 回以上も現れている．

例えば，“(((中心的な) 役割を) 果たす)” という部分木（丸括弧は構文構造を表す）は日本語 Web コーパス中で 8072 回出現し，“(plays(a(central)role))” という部分木は英語 Web コーパス中で 1902 回出現している．それぞれは高頻度であることから，各言語において意味のある表現と考えられる．また，これらは，対訳辞書を用いれば内容語が対訳の関係にあることも容易に分かり，構文構造の対応も容易につくため，互いに対訳の関係にある可能性が高いと推測できる．さらに，Web は急速に拡大していることから，このような部分木の組の割合も今後さらに増えると予想される．したがって，両言語において頻出する部分木に着目

Vol.2009-NL-192 No.7
2009/7/22

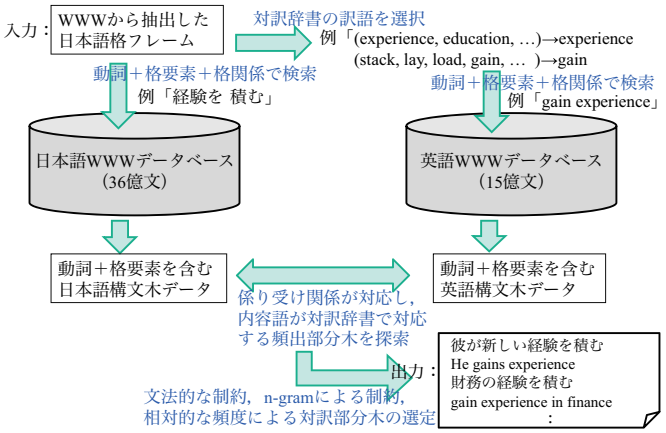

図 1 提案手法の概要

すれば，動詞とその格要素の組を含む対訳を網羅的に抽出できるかもしれない．提案手法はこのようなアイデアに基づいている．

以下，2 章で提案手法について説明し，3 章で実験について述べ，4 章で関連研究について述べ，5 章でまとめる．

2. 対訳 treebank 自動獲得

まず，提案手法の概要について説明する．提案手法の概要を図 1 に示す．提案手法は以下の 3 つのプロセスからなる．

(1) **種の対訳対の獲得**

Web 中の動詞と格要素の組を網羅するために，Web から構築した日本語格フレームから，動詞と格要素の組を抽出し，そのうち，頻度が 100 以上のものを種とする．そして，この種の動詞と格要素を対訳辞書で英訳する．この際に，対訳辞書にはさまざまな英訳が登録されているため，訳語選択を行う．

(2) **対訳部分木候補の抽出**

日本語の種を含む構文解析済の日本語 Web を検索する．また，英語の種を含む構文解析済の英語 Web を検索する．そして，日英それぞれ種の動詞を根とする構文構造の部分木の中で，係り受け関係が対応し，内容語が対訳辞書で対応する日英頻出部分木を網羅的に探索し，対訳候補とする．

(3) **対訳部分木の選定**

文法的な制約や n-gram による制約を満たさない対訳候補を除き，対訳の相対的な頻度が上位のものを選択することで，対訳候補を選択し，対訳表現として出力する．

以下，上記手順のうち，特に重要な要素技術である，1 の訳語選択，2 の探索アルゴリズム，3 の対訳部分木の選定について説明する．

2.1 訳 語 選 択

本節では，日本語格フレーム中の動詞と格要素の組に対応する英訳語を選択する手法について述べる．

対訳辞書には様々な訳語が登録されている．例えば，我々が用いている EDR[16] とクロスランゲージの日英対訳辞書をマージした辞書には，動詞「積む」の訳語として，「heap, load, ship, stack, save, acquire, gain, ...」が登録されている．何を積むのかによって，適切な語義となる訳語を選択する必要がある．

訳語選択の重要性は動詞だけではない．格要素の名詞についても様々な訳語が登録されている．上記辞書には，「門」の訳語として，「home, way, exam, house, door, school, gate, ...」が登録されている．ここで，格要素の主辞として適切な訳語を選択する必要がある．単純に共起頻度の高いものを選択する手法ではうまくいかないことが多い．例えば，「門をくぐる」の「門」と「くぐる」について，前記対訳辞書の英訳語が 15 億文の英語 Web コーパス中で動詞と目的語の関係で共起した頻度は多いものから，get home (107,304), get way (33,099), pass exam (22,643), go home (15,009), ... であった．丸括弧内は共起頻度を示す．ここで，共起頻度が最も高いものを訳語とすると，get と home が訳語として選ばれてしまう．

このように動詞および格要素の名詞の両方について訳語を適切に選択することが重要である．そこで，以下では，動詞の訳語選択手法，格要素の訳語選択手法，選択した訳語の検証方法について述べる．

動詞の訳語選択

文脈が近い語の語義が近いという分布仮説[4] にもとづいて訳語選択する．動詞に係る格要素を文脈として利用する．例えば，「経験を積む」の「積む」の訳語候補 “load” と “gain” から訳語としてどちらかを選ぶ場合を考える．load が Web 上でとる目的語の集合は {baggage, box, program, ...} などで，gain が Web 上でとる目的語の集合は {information, experience, weight, ...} などとなるとする．これらの集合のうち，「積む」の目的語である「経験」の対訳辞書の英訳語の集合 {experience, experiencing, undergo, ...} との近さで順位付けし，最も近い訳語候

補を選択する．

ここで，近さの定義として，次式のスコア s_1 が大きいものが近いとする．スコア s_1 は，ヲ格の日本語格要素の英訳語が，英語 Web コーパスで動詞訳語の目的語として多く出現した場合に大きくなる．ただし，ヲ格の日本語格要素の英訳語が，英語 Web コーパスで多くの種類の動詞の目的語になる場合はスコアは小さくなる．

$$s_1 = \sum_{n_j \in D} \frac{P(N = n_j | V = v_i, R = \text{“object”})}{H(V | N = n_j, R = \text{“object”}) + 1} \quad (1)$$

ここで N は名詞の変数，V は動詞の変数，R は動詞と名詞の関係の変数を表す．また n_j は英語名詞，v_i は英語動詞，D は格要素の英訳語の集合，P は確率，H はエントロピーを表す．

格要素の訳語選択

我々は格要素の主辞となる訳語は最も高い頻度で使われる語義とその訳語であると仮定し，まず最頻の語義とその訳語に訳語候補をしぼることにした．我々が用いている日英対訳辞書には，人が参照する辞書には通常記載されている語義の順位が登録されていなかったため，語義の順位は使えなかった．そこで，最頻の語義とその訳語は多くの辞書に登録されていると仮定し，複数の対訳辞書において，登録辞書数の多い訳語を最頻の語義とその訳語であるとみなすことにした．

格要素の訳語選択は次のようにする．まず，訳語候補を登録辞書数が最も多いもののみとする．この段階で訳語候補が複数ある場合は，動詞の訳語選択と同様に文脈（係り先の動詞）の近さにもとづいて順位付けし，最も近い訳語候補を選択する．ここで，近さの定義として，次式のスコア s_2 が大きいものが近いとする．

$$s_2 = \sum_{v_i \in E} \frac{P(V = v_i | N = n_j, R = \text{“object”})}{H(N | V = v_i, R = \text{“object”}) + 1} \quad (2)$$

ここで E は動詞の英訳語の集合を表す．

選択した訳語の検証

選択した動詞・格要素の訳語の組が英訳として適切かどうかを Web コーパス中の頻度を利用して検証する．選択した動詞・格要素の訳語と格関係（e.g. gain, experience, object）の三つ組の Web コーパスにおける頻度を調べ，その頻度がしきい値以上の場合に適切と認定する．格関係には，英語構文解析器が出力する関係ラベルを利用する．

2.2 探索アルゴリズム

本節では，日英それぞれ種の動詞を根とする構文構造の部分木の中で，係り受け関係が対応し，内容語が対訳辞書で対応する日英頻出部分木を網羅的に探索するアルゴリズムについて述べる．英語の構文木については，あらかじめ機能語を内容語のノードにまとめておく．

探索は以下の手順で行う．

(1) 種の英語動詞を根とする頻出部分木を，頻出部分木列挙アルゴリズム FREQT[1),11)] で列挙する．そして，部分木間の親子関係をグラフ構造で表現する（図 2 上の英語部分木と部分木間のエッジ）．部分木間の親子関係は，ある部分木にノードを 1 つ追加して新しい部分木ができた場合，元の部分木を親，ノードを追加した部分木を子の関係とする．このとき，親の部分木と子の部分木の差分のノードの情報も，グラフのエッジに記録しておく．

(2) 日本語部分木と対応する英語部分木を英語のグラフ構造の開始点（動詞のみの部分木）から探索する（図 2 上）．
ここで，次の条件を満たす場合に，日英の部分木が対応するものとする．
- 日英の各ノードが過不足なく対応する．このとき，1 対多の対応を認める．
- 対応するノードの日英の内容語が対訳辞書に対訳として登録されている．
- 構文木の係り受け関係及び係り受け関係の種類（ヲ格，object など）が一致する．

係り受けの種類の日英間の対応関係ルール（e.g. ガ格と subject，ヲ格と object，デ格と “in” など）は人手で作成し，このルールと一致する場合に，係り受け関係の種類が一致するものとする．

(3) FREQT により日本語の部分木に 1 ノード追加し，それに対応する英語部分木があるかどうかを調べる（図 2 中）．対応する英語部分木があれば，さらに FREQT により日本語の部分木に 1 ノードを追加し，それに対応する英語部分木があるかどうかを調べる処理を繰り返す（図 2 下）．日本語側でノードを追加した部分木と対応する英語部分木がない場合は，その日本語の部分木のさらなる拡張はしない．

上記探索手順の (3) では，対応しているかどうかは日本語側の追加されたノードとその係り関係，英語側も親の部分木と比べて追加されたノードとその係り関係を調べればよい．

2.3 対訳部分木の選定

本節では，前記探索アルゴリズムにより取得した対訳候補から質の高い候補を選択する手法について述べる．はじめに文法的な制約と n-gram の制約を満たさないものをルールにより除く．今回は英語表現が表 1 の条件にあてはまる対訳候補を除くというルールを用いる．次に相対的に頻度が高い対訳候補を選択する．対訳候補において，日本語または英語の表現に対応する対訳表現が多数ある場合，対訳表現の頻度が相対的に高いものを選択する．相対的に頻度が高い対訳候補の例を図 3 に示す．対訳側が同じ部分木である対訳候補の集合にお

Vol.2009-NL-192 No.7
2009/7/22

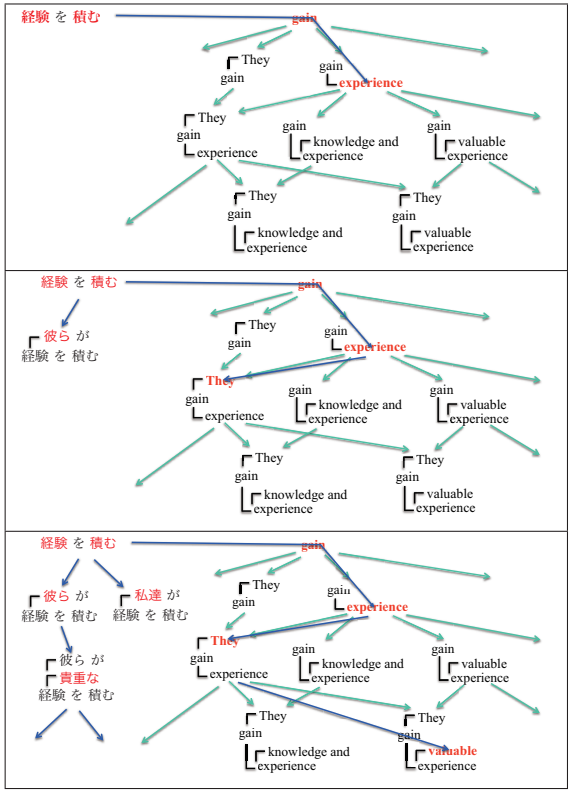

図 2 探索アルゴリズム

いて，日本語の部分木の頻度でソートした場合の上位と英語の部分木の頻度でソートした場合の上位をまず選択し，これらの選択結果のどちらにも含まれる対訳候補だけを選択する．この相対的な頻度による選択は 2 段階で行う．具体的には，まず種の対訳対が同じものの集合毎に選択を行う．次に，この選択結果全てを 1 つの集合として選択を行う．

3. 評 価 実 験

提案手法の有効性を確認するため，評価実験を行った．以下，提案手法による対訳 treebank 自動獲得実験，獲得した対訳 treebank の有効性を検証するための翻訳実験，そして考察について述べる．

表 1 英語表現除外の条件

冠詞が an（a）でなければならないが a（an）である．
root の動詞のより後の子ノードに not または品詞が MD（will など）を含む．
目的語の子ノードに any があり，動詞の子ノードに not がない．
英語の to 不定詞に対応する日本語表現が to 不定詞の用法に一致しない．
動名詞が親ノードより後に位置する．
名詞を修飾するノードが親ノードより後に位置する．
英語 Web コーパス中で頻度 0 のバイグラムを含む．

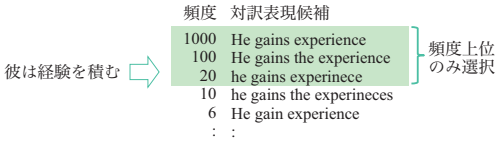

図 3 相対的に頻度が高い対訳候補の例

3.1 対訳 treebank 自動獲得実験

3.1.1 実 験 設 定

この実験では，対訳を収集する格要素の対象はヲ格のみとした．1 つの格フレームでの頻度 100 以上のヲ格の格要素とその格フレームの動詞との組を種とした．ただし，格要素のうち，全ての格フレーム中での格毎の格要素の総頻度において，ガ格と未格（KNP の出力結果，主に格助詞が「は」の文節）以外の格で，最も頻度が高い格がヲ格でない格要素は対象外とした．日本語格フレームには，京都大学格フレーム Ver 1.0 を利用した．

対訳辞書には，EDR 日英対訳辞書 V3.0，クロスランゲージ日英対訳辞書（対訳登録数 1,559,903），JMDict（EDR，クロスランゲージに登録されていない見出し語は使わない）の 3 つを利用した．これらの辞書をマージした辞書の対訳登録数は 1,940,211 であった．

日本語 Web コーパスには，TSUBAKI（情報爆発プロジェクト検索エンジン基盤）2007 データセットのうち，2 文節以上の約 36 億文を利用した．このコーパスは，JUMAN[17]，KNP[18] で構文解析済のものである．

英語 Web コーパスには，2004 年にクローリングした 3 単語以上，15 単語以下で文末にピリオドがある約 15 億文を利用した．このコーパスは，辻井研 POS tagger v1.0[9]，MSTParser v0.4.3（Penn Treebank 3 からモデルを構築）で構文解析を実施した．

種の訳語選択では，1 位の訳語とその間の格関係の組の頻度がしきい値以上の場合，動詞・格要素それぞれ 2 位までの訳語も辞書中の語義（辞書に登録されていた EDR の概念識別子

Vol.2009-NL-192 No.7
2009/7/22

または JMDict の sense タグ）が 1 位と同じで動詞の場合は式 (1)，格要素の場合は式 (2) により計算されるスコアが 1 位のスコアの 0.8 倍以上で Web コーパス中の頻度がしきい値以上であれば，それらの組も訳語とした．これによって，日本語動詞と格要素の組 1 つに対して，最大で 4 組の英訳が得られる．訳語を 2 位まででスコアが 0.8 倍以上に制限したのは，訳語の種類を増やすことよりも正しくない訳が含まれるのを防ぎたかったためである．なお，この制限に用いたパラメータの値は経験的に設定したものであり，最適値とは限らない．

構文構造の頻出部分木の出現頻度の下限は 5 とした．なお，この論文では研究の初期段階としてこのように固定したが，最適値とは限らない．

日本語部分木の根の動詞の表現は基本形にした状態で頻度を数えた．そして，日本語動詞の文末表現は基本形のままで対訳を獲得した．この際に，対訳となる英語の時制や否定などは対応するかの判断では問わないこととした．対訳候補を取得した後で，英語側の時制や否定，モダリティなどを識別し，対応する日本語側の動詞の文末表現を英語側に合うように変形した．これは，日本語の文末表現が多様なために日本語の部分木の頻度が少なくなってしまうのを避けるために行った．日本語動詞の変形は，例えば，過去形にする場合は，動詞の活用形を「タ形」に変更するといった人手で作成した 9 つのルールを用いて，動詞の活用型と活用形のテーブルを利用して行った．

対訳部分木の選定での相対的な頻度による選択では，選択する上位の数は次のようにした．最初の日本語の動詞と格要素の組毎の選択では，同じ日本語に対して上位 3 位までの英語，同じ英語に対して上位 2 位までの日本語を選択した．最初の選択結果全てから選択する 2 段階目の選択では，同じ日本語に対して上位 6 位までの英語，同じ英語に対して上位 4 位までの日本語を選択した．日本語よりも英語の数を多く設定している理由は，英語の場合，同じ内容語の並びでも単数形や複数形，冠詞の a と the の違いなど表記の異なりが日本語より多いと思われたからである．また，日本語の表現が同じでも対応する英語側で識別した時制や否定などが異なるものは，異なる日本語表現として扱った．

日英とも辞書登録の有無を調べる際には基本形を用いた．英語の lemma は XTAG ツール[⋆1]の英語形態素データを用いて取得した．

3.1.2 実験結果

得られた対訳 treebank の数は約 221 万，対訳 treebank に含まれる日本語の動詞とヲ格の格要素の組の異なり数は，34,042 であった．得られた対訳 treebank からランダムにサンプルした例を表 2 に示す．このサンプル中の丸括弧は構文構造を示している．次に，対訳 treebank の長さの分布を図 4 に示す．日本語の形態素数は 4 から 6 が多いことが分かる．英語の形態素数は 3 から 5 が多いことが分かる．さらに，対訳 treebank が格フレーム中の動詞とヲ格の格要素をどの程度含むかを調べた．ヲ格の格要素の総頻度のうち対訳 treebank に含まれる格要素の頻度の割合は 0.496 であった． 1 つの格フレームでの頻度 100 以上など今回対訳を獲得しようとした動詞と格要素の組の頻度の割合は 0.832 であった．このうち対訳 treebank に含まれる頻度の割合は 0.596 であった．

表 2 対訳 treebank の例

頻度	日本語	頻度	英語
49	(国 を も 滅ぼす こと)	54	(destroying(the nation))
41	((再度) 会議 を 開く)	9	(hold(a meeting)(again))
7	((彼 の) 手 を 濡らす)	10	(dip((his)hand))
251	(方法 を 選定 する だろう)	27	((will)choose(method))
42	((私 が) 報告 を 受けて いる だろう)	9	(((you)will)be receiving(this report))
23	(アドバイス だけ を 求め ない かも しれ ない)	7	((may(not))seek(advice))
18	(と いう レポート を 書く)	5	(Write(a report(on what(has meant))))
12714	(お 問い合わせ を 受け付け なかった)	11	(have(not)received(any inquiries))
24	((まさに) 金 を 捨てて いる だろう)	6	((will(just))be throwing away)
7	((企業 が) 関心 を 抱く)	9	((This Corporation)holds(the interest))

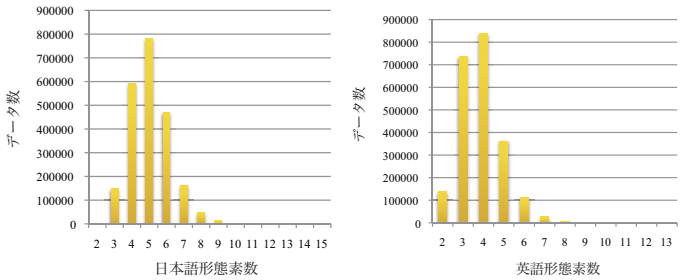

図 4 長さの分布

3.2 翻訳実験

3.2.1 実験設定

翻訳モデルの学習に利用した対訳コーパスの種類とその内容は以下の通りである．NTT 機械翻訳機能試験文[⋆2]（NTT，日本語文異なり数 3,718）：機械翻訳システム評価用の文例集である．読売新聞（Yomiuri，250,000 文対）：読売新聞と The Daily Yomiuri から自動作成さ

⋆1 ftp://ftp.cis.upenn.edu/pub/xtag/morph-1.5/morph-1.5.tar.gz

⋆2 http://www.kecl.ntt.co.jp/mtg/resources/mt-test-set-1.txt

Vol.2009-NL-192 No.7
2009/7/22

れた日英対応付けコーパス[10] である．JST（994,500 文対）：科学技術論文の抄録の対訳文である．BTEC（442,738 文対）：旅行会話の対訳文である．また，対訳コーパスの一種として用いた対訳辞書には，EDR 日英対訳辞書 V3.0 とクロスランゲージ日英対訳辞書の品詞を除いて見出し語の重複を除いたものを用いた．Web から構築した対訳表現は，頻度の情報は利用せずにそれぞれの対訳の頻度は 1 となる状態で対訳コーパスとして利用した．

テストデータは，各対訳コーパス中で Web から構築した対訳表現に含まれる日本語の動詞とヲ格の格要素の組を含む 500 文（ただし NTT は 455 文．この条件を満たすものが 455 文であったため）を，読売新聞は対応付けスコアの上位順，JST はデータの並び順，BTEC はデータが文字コード順になっていたためランダムに選択した．ここで日本語の動詞は基本形での一致により含まれるかどうかを判定した．開発データはそれぞれのコーパスから 500 文を選択した．言語モデル用データには，3.1.1 節の英語 Web コーパスのうち先頭からの約 2 億文を利用した．全ての実験で同じ言語モデルを利用した．

翻訳システムには moses を利用した．翻訳結果の評価値（BLEU，NIST）の計算には NIST のツール[⋆1] を用いた．日英の形態素解析には，3.1.1 節で用いたものと同じツールを用いた．

3.2.2 実験結果

翻訳結果を 1 リファレンスによる BLEU4 と NIST の翻訳評価スコアで評価した結果を表 3 に示す．表 3 の対訳コーパスの dic は対訳辞書，web は Web から構築した対訳表現を表す．dic+web とは，対訳コーパスとして dic と web を利用したことを表す．太字の部分は Web コーパスから構築した対訳表現（web）を追加して利用することにより，BLEU の値が 0.005 ポイント以上改善した部分である．また，灰色部分は学習データからテストデータを除いているため，白色部分と学習データが異なる．

表 3 より，テストデータが NTT の場合は，Web コーパスから構築した対訳表現を追加することで BLEU の値が向上している．テストデータが Yomiuri の場合は，対訳コーパスとして対訳辞書のみの場合に Web コーパスから構築した対訳表現を追加すると BLEU の値が向上している．その理由は，対訳辞書に不足している語順や語彙の共起に基づく訳語選択結果や機能語を，Web コーパスから構築した対訳表現が補ったためと考えられる．BTEC については，Web から構築した対訳表現を追加することによる効果は見られなかった．これは，Web で中・高頻度な表現と比べてドメインが遠いためと思われる．

各対訳コーパス中でのカバレッジと動詞が持つ格の割合を表 4 に示す．ここで，key とは，

⋆1 ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

表 3 翻訳実験の評価

	テストデータ							
	NTT		Yomiuri		JST		BTEC	
対訳コーパス	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
dic	0.0432	2.5688	0.0578	2.9988	0.0501	3.3442	0.0373	1.3856
dic+web	**0.0660**	3.1331	**0.0646**	3.2395	0.0533	3.5543	0.0274	1.4398
NTT	0.0479	2.7825	0.0171	1.6817	0.0093	1.7597	0.0000	1.0346
NTT+dic	0.0674	3.4961	0.0465	3.0345	0.0402	3.7987	0.0328	1.5342
NTT+dic+web	**0.0887**	3.9130	0.0458	3.3293	0.0397	3.9810	0.0176	1.4972
Yomiuri	0.0699	3.5911	0.1383	5.0272	0.0370	3.2138	0.0352	1.8619
Yomiuri+dic	0.0924	4.0832	0.1222	4.8259	0.0892	4.3978	0.0665	2.1417
Yomiuri+dic+web	**0.1157**	4.3184	0.1326	4.9108	0.0756	4.4374	0.0478	1.8259
JST	0.0670	3.5024	0.0801	4.0573	0.2030	6.6094	0.0000	1.5287
JST+dic	0.0760	3.7844	0.0888	4.2886	0.2062	6.7624	0.0483	1.8199
JST+dic+web	**0.1037**	4.2594	0.0905	4.3522	0.2041	6.6508	0.0313	1.7805
BTEC	0.0060	1.3459	0.0046	1.0226	0.0000	0.7653	0.6018	6.5064
BTEC+dic	0.0335	2.8000	0.0424	2.8461	0.0409	3.2874	0.7133	7.1403
BTEC+dic+web	**0.0664**	3.6321	**0.0555**	3.4072	**0.0496**	3.9138	0.6241	6.7895
web	0.0487	2.6968	0.0460	2.6643	0.0430	2.9327	0.0182	1.1685

表 4 各コーパスの日本語側のカバレッジ

	NTT	Yomiuri	JST	BTEC
key を含む文の割合	0.122	0.188	0.184	0.064
動詞を含む文の割合	0.874	0.952	0.958	0.916
ヲ格を持つ動詞のうち格要素と動詞が key に含まれる割合	0.309	0.227	0.205	0.147

表 5 NTT テストデータの動詞と目的語の翻訳の評価

対訳コーパス	評価		
	◎	○	×
Yomiuri+dic	0.183	0.107	0.710
Yomiuri+dic+web	0.493	0.120	0.387

Web から構築した対訳表現に含まれるヲ格の格要素と動詞の組を示す．

3.3 考　　察

提案手法で Web コーパスから構築した対訳表現を用いることにより，Web で中・高頻度な表現を翻訳できるようになったかどうかについて考察する．

翻訳実験に用いたテストデータのうち，最も特定のドメインへの偏りが少ないと思われる NTT 機械翻訳機能試験文を対象に検討する．テストデータは Web から構築した対訳表現に

表 6 NTT テストデータの翻訳の相対評価

動詞と目的語の翻訳評価	コーパス追加の影響	割合
◎または○	改善	0.387
	どちらともいえない	0.090
	改悪	0.017
×	改善	0.177
	どちらともいえない	0.233
	改悪	0.097

表 7 対訳 treebank を追加することで動詞と目的語の翻訳が良くなった例

翻訳元文（NTT）	翻訳結果	
	対訳コーパス:Yomiuri+dic	対訳コーパス:Yomiuri+dic+web
妻が湯を沸かす。	his wife hot water heat.	wife boil the water.
彼は手を合わせた。	he is.	he joined the hand.
秘書がお客にお茶を入れた。	secretary of tea to customers.	the secretary to customers made the tea.
その言葉は皆の心を動かした。	the word as the heart of the people.	words have changed the minds of all.
浅間山が爆発を起こした。	an agency.	the mountain, asama caused the explosion.
彼女は花に水を掛けた。	she is a flower in the water.	she poured the water into the flowers.
彼らは坂道を上った。	to the slope.	they ascended the hill.
彼は目を開いた。	he is the eyes.	he opened the eyes.
日本はドイツと同盟を結んだ。	the alliance of germany and japan.	japan has formed an alliance with germany.
彼は頭を下げた。	he is.	he lowered the head.

表 8 対訳 treebank を追加することで動詞と目的語の翻訳が悪くなった例

翻訳元文（NTT）	翻訳結果	
	対訳コーパス:Yomiuri+dic	対訳コーパス:Yomiuri+dic+web
私は何回も足を運んだ。	i went to several times.	i moved feet several times.
私は彼の話に耳を貸した。	that i was actually listening to.	he said i lent an ear.
その箱をあけるにはボタンを押しさえすればよい。	in the empty box if the push button.	this button to open the box only if the.
私は娘に本を読んでやる。	i read a book to her.	i play reading this book for daughter.
彼は飛んでいる鳥を打ち落とした。	he hit the bird is out.	birds are flying, he cast the.
私は彼をその部屋で見た。	i had seen him in the room.	the he i watched in the room.
我々のチームは 5 月より研究を開始する。	our team to begin research in may.	we in five team, of the research.
彼女はよく英語を話す。	she is a speak english well.	she is the english as well.

含まれる動詞とヲ格を含むため，この部分の表現は Web 上で中・高頻度な表現といえる．

NTT のテストデータのうち，最初の 300 文を対象に動詞とヲ格の格要素が正しく翻訳できたかどうかを人手で評価した．評価は，動詞とヲ格の格要素の訳語が正しく，格要素が動詞の目的語になっていて，動詞の時制および否定・肯定も正しいものは◎，◎の条件のうち動詞の時制または否定・肯定が正しくないものは○，それ以外は×とした．評価の対象は，対訳コーパスに読売新聞と対訳辞書を用いた場合と，さらに Web から構築した対訳表現を追加した場合の翻訳結果とした．◎と○と×の割合の評価結果を表 5 に示す．

表 5 より，読売新聞と対訳辞書を対訳コーパスとして利用した場合より，さらに Web から構築した対訳表現を追加したほうが動詞と目的語の訳が良くなっていることが分かる．

また，Web から構築した対訳表現を追加した場合の翻訳結果がこの対訳表現を追加しない場合の結果に比べて改善したか，改悪したか，どちらともいえないかを，ヲ格を持つ動詞とそれに係る全ての格要素およびその格要素を修飾する語の範囲で評価した．評価結果を表 6 に示す．評価結果は表 5 での◎または○の場合と×の場合の区別が分かるように示している．Web から構築した対訳表現を追加することで翻訳が改善する場合が多いことが分かる．

以上より，提案手法で Web から構築した対訳表現が Web 上で中・高頻度な表現の翻訳に有効であるといえる．

表 8 に読売新聞と対訳辞書を対訳コーパスとした場合（Yomiuri+dic）より Web から構築した対訳表現を追加したほう（Yomiuri+dic+web）が動詞と目的語の訳が良くなった例を示す．この例から，簡単な文でも新聞と対訳辞書だけを対訳コーパスとして用いると正しく翻訳できない場合があることが分かる．それに対して，Web から構築した対訳表現を対訳コーパスに追加した場合は，構文構造の中心となる動詞と目的語が正しく翻訳できている場合があることが分かる．

Web から構築した対訳表現を追加することで，結果が悪くなった原因について調べた．表 5 の Yomiuri+dic で評価が◎であったものが，Yomiuri+dic+web で評価が×になった 13 件について調べた．その原因は，「足を運ぶ」，「耳を貸す」という慣用表現が適切に訳せなかった（4 件），ヲ格と動詞が連続していなくて Web から構築した対訳表現を活用できなかった（1 件），Web から構築した対訳表現の訳語が一般的でなかったため，既存データで翻訳できる部分に悪影響を及ぼした（3 件），Web から構築した対訳表現の動詞の末尾表現の種類が不足していて入力の動詞末尾表現（さえすればよい，やる，ほしい）をカバーできていない（3 件），「打ち/落とす」という複合動詞が Web から構築した対訳表現になかった（1 件），Web から構築した対訳表現に問題はなく原因不明（1 件），であった．

4. 関連研究

対訳コーパスを自動獲得する手法は，これまでにいくつか提案されている．

対訳の Web ページから対訳コーパスを自動獲得する手法[8]，コンパラブルなコーパスから対訳コーパスを自動獲得する手法[2),6),7),10)] が提案されている．しかし，対訳が存在する Web ページやコンパラブルコーパスは単言語の Web コーパスに比べて量が非常に限られてしまう．また，コンパラブルコーパスが存在するドメインは主に報道記事である．そのため，コンパラブルコーパスだけでは，Web 上の表現を幅広く収集することは難しい．

ルールベース機械翻訳を用いて対訳コーパスを自動構築する手法[5] が提案されている．しかし，ルールベース機械翻訳は，Web コーパスに存在しない表現も対訳表現として生成する．そのため，対訳表現が目的言語側の自然な表現であるとは限らず，翻訳誤りも含まれる．その一方で，Web コーパスに対訳が存在しない場合でも正しい対訳を生成できる可能性がある．提案手法は Web コーパスに中・高頻度で出現した表現のみからコーパスを構築する．相補的なので併用することでそれぞれの利点を活用できると思われる．

単言語のデータから対訳の語彙を獲得する手法[3] が提案されている．しかし，この手法は複数単語からなる表現の対訳を獲得することができない．

また，単言語のデータを対象に複合名詞の対訳を単言語コーパスを利用して自動獲得する手法[14),15)] も提案されている．複合名詞の構成要素を対訳辞書で翻訳して対訳候補を生成し，単言語コーパスを用いて対訳候補を検証するという手法である．Web コーパスに対してこの手法を適用すれば，Web コーパス全体から複合名詞の訳語を獲得することができると考えられる．しかし，動詞を含む表現にこれらの複合名詞の対訳獲得手法を適用するには問題がある．なぜなら，動詞を含む表現には複合名詞にはない以下１）から４）のような特徴があり，複合名詞の対訳を獲得する手法はこれらを想定していないまたはこれらに対応していないためである．１）多くの場合に付属語や機能語が含まれる．２）言語によって語順が変わる場合がある．３）対訳コーパスとして有用な表現は連続する表現に限らない．４）長さの単位に制限がなく，最大で 1 文の長さまで対象となる．これに対し，提案手法では，動詞を含み，上記の特徴を想定し，かつ，Web コーパス全体の中で頻出する部分木の対を獲得する．

5. おわりに

大規模な単言語 Web コーパスから動詞を含む対訳の頻出部分構文木を対訳辞書を用いて幅広く獲得する手法を提案した．対訳 treebank の自動獲得実験と獲得した対訳を用いた翻訳実験により提案手法の評価を行い，有効性を確認した．今後はさらにヲ格のカバレッジを増やす方法を検討するとともに，ヲ格以外の格も対象として対訳 treebank を構築し，さらに多くの表現を網羅できるようにしたい．また，提案手法だけでなく，既存手法も利用して Web コーパスをできるだけ多く対応付けすることで大規模な対訳コーパスを構築したい．

参 考 文 献

1) Abe, Kenji and Kawasoe, Shinji and Asai, Tatsuya and Arimura, Hiroki and Arikawa, Setsuo: Optimized Substructure Discovery for Semi-structured Data, PKDD, pp.1-14, 2002.
2) Fung, Pascale and Cheung, Percy: Mining Very-Non-Parallel Corpora: Parallel Sentence And Lexicon Extraction Via Bootstrapping And EM EMNLP, pp.57-63, 2004.
3) Haghighi, Aria and Liang, Percy and Berg-Kirkpatrick, Taylor and Klein, Dan: Learning Bilingual Lexicons from Monolingual Corpora, ACL, pp. 771-779, 2008.
4) Harris, Z. : Distributional structure, Word, 10(23): pp.146-162. 1954.
5) Hu, Xiaoguang and Wang, Haifeng and Wu, Hua: Using RBMT Systems to Produce Bilingual Corpus for SMT EMNLP, pp.287-295, 2007.
6) Munteanu, Dragos Stefan and Marcu, Daniel: Improving Machine Translation Performance by Exploiting Non-Parallel Corpora, CL, 31(4), pp.477-504, 2005.
7) Munteanu, Dragos Stefan and Marcu, Daniel: Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora, ACL, pp.81-88, 2006.
8) Resnik, Philip: Mining the Web for bilingual text, ACL, pp.527-534, 1999.
9) Tsuruoka, Yoshimasa and Tsujii, Jun'ichi: Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, HLT/EMNLP, pp. 467-474, 2005.
10) Utiyama, Masao and Isahara, Hitoshi: Reliable Measures for Aligning Japanese-English News Articles and Sentences, ACL, pp. 72-79, 2003.
11) Zaki, Mohammed J.: Efficiently Mining Frequent Trees in a Forest, ACM KDD, 2002.
12) IWSLT Evaluation Campaign on Spoken Language Translation, 2008.
13) NIST Open Machine Translation Evaluation, 2008.
14) 田中貴秋，松尾義博：対訳関係のないコーパスからの複合名詞対訳の表現の獲得，電子情報通信学会論文誌, Vol. J84-D-II, No.12, pp.2605-2614, 2001.
15) 外池昌嗣，宇津呂武仁，佐藤理史：ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定，自然言語処理, 14(2), pp.33-68, 2007.
16) EDR 電子化辞書 使用説明書 3.0 版，通信総合研究所.
17) 日本語形態素解析システム JUMAN 使用説明書.
18) 日本語構文解析システム KNP 使用説明書.