非対訳コーパスからの日英機械翻訳ルールの自動獲得

田中 貴秋　　松尾 義博　　大山 芳史
NTTコミュニケーション科学研究所
{takaaki,yosihiro,ooyama}@cslab.kecl.ntt.co.jp

1 はじめに

機械翻訳の性能を左右する要素として両言語間の変換ルールが挙げられるが、日本語と英語のように言語構造が大きく異なる場合には単語単位の変換だけでなく、句単位以上のルールが必要になる。現実の文章の分野に対応して変換ルールを用意するには、コーパスからの自動獲得の機構が不可欠であり、多くの研究が行われている。

宇津呂ら [1] は構文解析結果を素性構造に変換し両言語間で照合することにより対訳コーパスから動詞の表層格パターンを獲得する方法を提案している。また Haruno ら [2] は対訳コーパスから各言語内での共起頻度を用いて collocation を抽出し、両言語間の共起情報を利用して対訳表現を獲得している。Kaji ら [3] は、句構造木で現された構文解析結果から両言語で句構造を対応づけることにより変換テンプレートを獲得する方法を提案している。

これらはいずれも文単位で両言語の対応が取れた対訳コーパスから対訳表現を得るには有効な方法であるが、実際にはこのようなコーパスは数が限られており、また作成するには膨大な作業量を必要とする。

一方、単言語のコーパスであれば大量に利用することが可能である。例えば、田中ら [4] は対訳関係のないコーパスから、両言語での共起情報の類似性を仮定して訳語関係の抽出を行っている。しかし、日本語と英語を対象とした場合、両者の言語構造が大きく異なっているために句より大きな単位の変換ルールを獲得することは容易ではない。

本稿では、正解の対訳例を手がかりとして、独立した2言語のコーパスから機械翻訳の変換ルールとして用いるのに適切な対訳表現を収集する方法を提案する。辞書に記載されている対訳表現を例として与え、それに類似する対訳表現を獲得する実験を行った結果について報告する。

2 基本的な考え方

2言語の表現対を機械翻訳のための変換ルールとして用いるためには、次の2点を満たす必要がある。

1. 二つの表現に意味的な対応があること
2. あるまとまった意味を表す単位として適切であること

コーパスからこのような表現を自動的に収集する場合、通常 1. の手がかりとして対訳コーパスにおける文の対応関係や対訳辞書などを用い、2. を満たすために構文情報や統計情報を用いる。

対象とする表現を対訳コーパスからではなく、独立した 2言語のコーパスから獲得しようとする際に問題となるのは、1. の手がかりが与えられない点である。したがって、直接 1. の精度を追求していくことは困難であるが、対訳となりうる表現の組を収集し、2. を満たすものを選択することで適切な表現の組をある程度絞り込むことができると考えられる。

本手法では、初めに 1., 2. を満たす対訳例を与えてこれに類似する表現を収集して対訳候補とすることを考える。例えば英和辞書には

- ◇ …に答えて : in response to ～
- ◇ …に比べて : in comparison to ～
- ◇ …に従って : in obedience to ～

などの対訳表現が記載されている [5]。これらは、形の上では日本語が節の一部[1]、英語が前置詞句の一部となっており2言語間では大きな相違があるが日本語から英語に変換するパターンはいずれも類似している。

このような対訳表現と類似した表現を日本語のコーパスから抽出し、英語訳を対訳例から類推して英語表現を作れば、新しい対訳表現の候補ができる。これらの中には、1., 2. を満たす表現が含まれていることが期待される。類推した表現の中には、2. を満たさない不適切なものが含まれるので除外する必要がある。これらは、英語のコーパスを検索して該当の表現が出現するかどうかを調べることにより取り除くことができる。こうして 2. を満たさないものを削除していくことで間接的に 1. を満たさない表現の組も除外されると予想される。

以上のことを仮定して、対訳関係のない独立したコーパスから変換ルールとするのに適切な対訳表現を獲得する実験を行った。本稿ではその獲得方法と、実験結果について述べる。

3 対訳表現の獲得方法

3.1 全体の構成

本手法では、対訳ではない独立したコーパスを表現の収集対象とするので対訳表現を探す手がかりとして対訳例を与える。初めにこの対訳例から日本語から英語への変換パターンを抜きだす。もし、同様なパターンで翻訳可能な表現が存在するならば、それぞれのコーパス中に該当する表現が出現すると考えられる。

次に対訳例の変換パターンに類似した表現を収集する。対訳例から抜き出した変換パターンの日本語の条件に当てはまる表現を日本語のコーパスから抽出し、対訳辞書などを使い変換パターンに従って英語表現を生成し、この日本語と英語の組合せを新たな対訳の候補とする。

生成された英語表現から妥当なものを選択するためにそれぞれの表現を英語のコーパスで検索する。妥当

[1] 新納らは日本語の助詞相当表現 (関連表現) を抽出する方法について提案している [6]

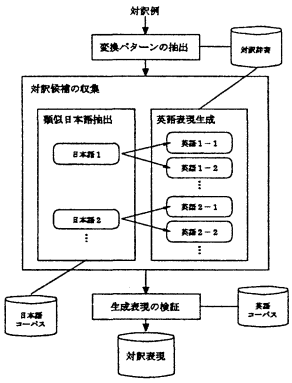

図 1: 構成概略図

な表現であれば実際の文章中で使用される可能性が高く、不適切なものは出現しないのでその分別ができると考えられる。また同時に、収集された日本語について適当でないものも、生成された英語表現がコーパスに現れないために結果的にここで除外されることが期待される。

全体構成の概略を図1 に示す。

変換パターンの抽出 与えられた対訳例からその変換パターンを抽出する

対訳候補の収集 対訳例と類似した表現対を収集する

類似日本語抽出 変換パターンの条件を満たす日本語の類似表現をコーパスから収集する

英語表現生成 抽出した変換パターンに従って、収集した日本語表現から英語表現を生成する

生成表現の検証 生成した英語表現の妥当性をコーパスを使って検証する

以下にそれぞれの処理についての詳細を示す。

3.2 対訳例からの変換パターンの抽出

与えられた対訳例の結び付きを変換パターンとして抽出し、収集しようとする表現のパターンを決定する。本稿では日本語表現 J から英語 E への変換パターン $T(\cdot)$ ((1) 式) は以下のようにして抽出する。

$$T(J) \rightarrow E \qquad (1)$$

1. 対訳例の日本語の構成単語のうち自立語を、対訳辞書などを使って英語に変換する。その際、日本語の品詞以外の派生形にも変換する。
2. 変換した英語と一致するものが対訳例の英語表現にあれば、両言語の単語を対応付けて変数化しその変換方法（派生形など）を記述する。一致する英語が存在しない場合、その単語は対応付けを行わない。

例えば「に答えて : in response to」の対訳表現の場合、日本語の構成単語のうち自立語「答える」について対訳辞書を引くと、その派生語を含めて “answer”、“respond”、“response” などが得られる。そのうち ”response” が英語表現中の単語に一致するので、両者を対応させて変数化し、「動詞から名詞に変換する」ことを記述する。その結果、$T(\cdot)$ は (2) 式 のように抽出される。

$$T\begin{pmatrix} \text{に（助詞）}, \\ \alpha_{1j}\text{（動詞, 連用形）}, \\ \text{て（助詞）} \end{pmatrix} \rightarrow \begin{pmatrix} in, \\ \alpha_{1j} \rightarrow \alpha_{1e}\text{（名詞）}, \\ to \end{pmatrix} \qquad (2)$$

3.3 対訳表現候補の収集

対訳例と同様のパターンで翻訳することのできる日英の表現の組を探すためにその候補を収集する。日本語コーパスから対訳例の日本語と類似したものを収集し、それから変換パターンに従って英語表現を生成する。この日本語と英語の組を新たな対訳表現の候補とする。

日本語類似表現の抽出

3.2 で抽出した変換パターン $T(\cdot)$ の日本語の条件を満たす表現を、対訳例と類似した表現として日本語コーパスから抽出する。形態素解析済みのコーパスから変換パターンの条件部を参照して

- ◇ 変数は、品詞と活用形が等しいもの
- ◇ 字面の部分は、字面、品詞、活用形がいすれも等しいもの

の並びで構成されている表現を類似表現として収集する。抽出の条件として構文情報を用いず単語列を抽出するので適切でない表現も多数収集されるが、後の処理で除外する前提で量を集めることを優先する。

英語表現の生成

収集された日本語の表現を、変換パターン $T(\cdot)$ に基づいて変換し、対応する英語表現の候補を生成する。このとき一つの日本語に対して複数の英語表現を生成することも許す。例えば、(2) 式に当てはまる表現「にしたがって」で、動詞「したがう」に対して和英辞書に対訳語が “follow”,”obey” と記載されていたとすると

$$\text{にしたがって} \rightarrow \begin{cases} \text{in follow to} \\ \text{in obedience to} \end{cases}$$

の 2 組の表現対を生成する。

3.4 収集した対訳候補の取捨

3.3 で生成された英語表現は日本語から機械的に変換パターンを適用して作られたものであるので、生成された表現から妥当でない表現を除外する必要がある。ルール化するのに不適切な日本語表現からは誤った英語表現が生成される可能性が高いので、英語表現が妥

当でないものを取り除くことで不適切な日本語も同時に削除されることが期待できる。

生成された英語表現を英語（目的言語）のコーパスで検索し、実際の文中で使用されているかを調べる。この段階で、変換ルール化するのに適切でない表現は大部分がコーパスに出現しないので取り除くことができると考えられる。

しかし、検索は字面の単語列として行い構文情報を考慮していないので、その単語並びがコーパス中に偶然現れれば現れれば、不適切な表現でも削除されない。例えば、もしコーパス中に次のように "in advance" という表現が出現すると、それを含んだ "in advance to" という表現は変換ルール化するのに相応しくないにもかかわらず残ってしまう。

> They were prepared *in advance* to violate the law in order to provoke.

そこで、このような部分表現を含むものを除外するために各構成語が共起する確率を計算し、結び付きが弱いと考えられる表現を削除する。

この例では 3 単語の列 w_1, w_2, w_3 に対して次の (3)(4) 式のような条件付き確率を考え、これらがそれぞれ設定した閾値 Th_1、Th_2 以上になる表現を残し、小さいものを削除することにより、上の例のような偶然に現れた組合せを除外できると考えられる。

$$p(w_1w_2 \cdot w_3 \mid w_1w_2) \geq Th_1 \quad (3)$$

$$P(w_1 \cdot w_2w_3 \mid w_2w_3) \geq Th_2 \quad (4)$$

4 実験と考察

4.1 対訳表現候補の収集結果

次の 3 組の対訳例

(a) **に答えて** : **in** response **to**

(b) **を考えて** : **in** consideration **of**

(c) **と協力して** : **in** cooperation **with**

を使って実験を行った。コーパスとして、日本経済新聞 5 年分[2](約 1GB) に日英機械翻訳システム ALT-J/E[7] の形態素解析処理を行ったものを使用した。変換パターンの抽出や英語表現の生成には ALT-J/E の日英対照辞書（対訳辞書）を用いた。

対訳例 (a) ～ (c) からは、それぞれ太字の部分が字面で残り、他の部分が変数化されたものが、変換パターンとして抽出された。例えば 対訳例 (a) からは (2) 式のような変換パターンが抽出された。この変換パターンの条件に当てはまる日本語の表現をコーパスから収集し、その表現から対訳辞書を用いて対訳の候補となる英語表現を生成した。

結果を表 1 に示す。収集された日本語に対して約半数について何らかの英語表現が生成され、英語の種類はさらにその半分となっている。つまり複数の日本語から同一の英語表現が生成されている割合が高い。

[2] 日本経済新聞 CD-ROM 90 版～94 版を使用した

表 1: 対訳表現候補収集の結果

変換パターン		(a) 型	(b) 型	(c) 型	計
収集日本語表現		3208	4194	1495	8897 種
対訳表現候補		5041	5953	2954	14308 組
	日本語種類	1844	2050	881	4775 種
	英語種類	991	1110	784	2885 種

表 2: 生成された英語表現の検証結果

変換パターン	(a) 型	(b) 型	(c) 型	計
英語がコーパスに出現	197	272	81	550 組
共起確率の条件を満たす	43	106	72	221 組

4.2 生成英語表現の検証

4.1 で示したように、(a) 型の場合で約 1800 の日本語表現に対して 1000 弱の英語表現が生成されるが、これらの表現を英語コーパスで検索し、実際の英語の文章中で使用される表現であるかどうかによってその判定を行った。

英語コーパスとして新聞記事 1 年分 (約 40MB) を使い、対訳例 (a) ～ (c) それぞれのパターンから生成された英語表現を検索した。さらにコーパス中に一度以上出現した表現について (3)(4) 式を計算し絞り込みを行った。本実験では $Th_1 = 0.3$、$Th_2 = 0.1$ と設定した。

結果を表 2 に示す。英語コーパスの検索を行った段階で (a) ～ (c) 型合わせて 14000 組ほどあった対訳候補を 4 % 弱に絞り込むことができた。(a),(b) のパターンの場合にはその中の約半分に英語表現の単位として不適切なものが含まれているが、英語表現の共起確率を計算すると図 2のように分布するので、適当な閾値で分離できる。本実験で設定した値では英語表現の単位として適切なもののみに絞り込まれた。

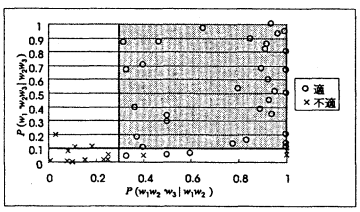

図 2: 英語表現の共起確率の分布

4.3 獲得された表現

本方法により最終的に獲得された表現を表 3に示す。4.2 の検証で残った日英の表現対のうち 3 割から 5 割のものが変換ルールに使用するのに適当な対訳表現となっている。対訳関係のないコーパスからこのような

表 3: 獲得された対訳表現数

変換パターン	(a) 型	(b) 型	(c) 型	計
獲得された表現	43	106	72	221 組
妥当な表現 (割合)	18 (42 %)	35 (33 %)	35 (49 %)	88 組 (40 %)

表 4: 収集表現と学研英和辞書収録表現の比較

	収集された対訳数	割合
同一対訳が収録	6	7%
同一日本語のみ収録	3	3%
同一英語のみ収録	42	48%
日英どちらも未収録	37	42%
計	88	100%

表 5: 辞書記載語と類似の表現

辞書記載日本語	収集日本語	対訳
をさがし求めて	を探して	in search of
を見越して	を予想して	in anticipation of

数の表現が獲得できたことは本手法の有効性を表していると言える。

本手法で収集された表現について検討するため、学研電子英和辞書 (見出し語約 6 万語) に収録されている同様のパターンに当てはまる対訳表現と比較したものを表 4に示す。

獲得された妥当な対訳表現 88 組のうち、すでに同一の日本語と英語の対訳表現が収録されていたのは、約 7% で、残りは日本語か英語の少なくとも一方は英和辞書に収録されていなかったものである。最も数が多かったのは 表 5 に示すように英語表現は同一のものが収録されているが、対訳として付けられた日本語と収集した日本語が異なっているものである。このようなものは、同一の英語表現に翻訳される日本語のヴァリエーションを収集したものとなっており、日英機械翻訳に利用するには都合が良い。また、日本語と英語のどちらもが辞書に収録されていなかったものも 42% に上っており、新しい対訳表現も多く獲得されている。

4.4 獲得できなかった表現

辞書に記載されているが獲得できなかった表現もあり、これは 2つの原因による。一つは「に反ばくして」や"in obedience to" のようにコーパス中に出現しなかった表現を含むものである。ある程度の規模のコーパスを使用した結果出現しなかった場合には、使用したコーパスの分野では用いられない表現が除外されたと見ることができる。

もう一つは対訳表現「に優先して : in preference to」のように日本語は抽出されたが、英語表現を生成することに失敗したものである。使用した辞書に「優先 (する)」の訳語として "preference" が無かったのが原因である。これは、生成に使用する辞書を増やしたり、同義語辞書などを使用することによってある程度向上すると思われる。

4.5 日英表現の対応妥当性

英語表現は、コーパスから抽出した日本語表現と対訳辞書を使って生成するが一般に各単語は多義をもつため、一つの日本語表現から複数の英語表現が生成されたり、逆に複数の日本語表現から同一の英語表現が生成されたりするため、日英の対訳としてどの組合せが相応しいかを判定する必要がある。例えば、「と比較して」「と合わせて」「と比べて」からいずれも英語表現 "in comparison with" が生成された。特に「と合わせて」は「合わせる」に多義があり、機械的に対応付けることは危険である。このように英語として妥当な表現が得られても対応する日本語の妥当性を検証する必要がある。

これらの用いられ方の違いは共起する語に反映すると予想される。日本語、英語の各表現の共起語を日本語語彙大系 [8] に記載されている名詞の意味属性などに基づいて分類することが考えられる。例えば、ある表現 c に共起する語のうち意味属性 i に属する数 a_i に対して特徴ベクトルを $\mathbf{f}_c = (r_0, r_1, ..., r_n)$ (ただし $r_i = a_i / \sum_k a_k$) のように定義し、日本語 j と英語 e の対応する尤度をそれぞれの特徴ベクトルの内積 $(\mathbf{f}_j \cdot \mathbf{f}_e)$ で表すことにより、対応可能性を順序付けることなどが考えられる。

5 おわりに

本稿では、日本語のコーパスを表現の抽出に使い、そこから生成した表現を英語のコーパスを用いて検証することによって、対訳例と類似した対訳表現を獲得する方法を提案した。結果として対訳関係のない独立した 2 言語のコーパスから 3 割から 5 割程度の精度で対訳となる表現を獲得できることを示した。また、この方法により与えた対訳例以外の対訳表現が収集できるだけでなく、同一の英語表現に翻訳されるべき日本語の表現のヴァリエーションも同時に収集できることを示した。

今後は、適合率を向上させるため収集した日本語と英語の対応妥当性を検証する方法を検討する予定である。また、本手法を他の対訳例 (変換パターン) に適用することも考察したいと考えている。

参考文献

[1] 宇津呂武仁, 松本裕二, 長尾真. 二言語対訳コーパスからの動詞の格フレーム獲得. 情報処理学会論文誌, Vol. 34, No. 5, pp. 913–924, 1993.

[2] Masahiko Haruno, Satoru Ikehara, Takefumi Yamazaki. Learning bilingual collocations by word-level sorting. In *Proc. of COLING*, 第 1 巻, pp. 525–530, 1996.

[3] H. Kaji, Y. Kida, Y. Morimoto. Learning translation templates from bilingual text. In *Proc. of COLING*, pp. 672–678, 1992.

[4] 田中久美子, 岩崎英哉. 非対訳コーパスを用いた訳語関係の抽出. 情報処理学会研究報告 110-13, 1995.

[5] 田中貴秋, 松尾義博, 大山芳史. 英和辞書からの日英翻訳ルールの自動獲得. 情報処理学会研究報告 119-15, 1997.

[6] 新納浩幸, 井佐原均. コーパスからの関係表現の自動抽出. 情報処理学会論文誌, Vol. 35, No. 11, pp. 2258–2264, 1994.

[7] S.Ikehara. Multi-level machine translation method. *Journal of Future Computing Systems*, Vol. 2, No. 3, 1989.

[8] 池原悟, 宮崎正弘, 白井 諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.