自然言語処理 114－11
(1996. 7. 18)

対訳コーパス中の共起頻度に基づく対訳表現の自動抽出

北村美穂子
沖電気工業株式会社
研究開発本部 関西総合研究所
kita@kansai.oki.co.jp

松本裕治
奈良先端科学技術大学院大学
情報科学研究科
matsu@is.aist-nara.ac.jp

あらまし

対訳コーパスに出現する任意長の単語列における二言語間の自動対応付けの方法を提案する．片言語内で複数回出現する一語以上の単語からなる単語列を抽出し，それらを二言語間で対応付け，両者の共起頻度に基づいた類似度を求めることによって単語列のペアを抽出する．単語列間の類似度の高いペアから順に繰り返し抽出する方法により，精度の高いペアから順に抽出することができる．類似度の計算は，各単語列の出現回数と各単語列間の二言語間での共起頻度を用いたDiceの式を拡張した式を使用する．分野の異なる三種類のコーパスを用いた抽出実験より，各コーパスの分野に依存した対訳表現が80% 以上の精度で抽出できることを示した．

和文キーワード 統計に基づく自然言語処理, 対訳コーパス, 機械翻訳, 対訳表現, 単語間の類似度

Automatic Extraction of Translation Patterns in Pararell Corpora

Mihoko Kitamura
Oki Electric Industory Co., Ltd.
Kansai Laboratory, Reseach & Development Group

Yuji Matsumoto
Nara Institute of Science and Technology
Graduate School of Information Science

Abstract

This paper proposes a method of finding correspondences of arbitrary length word sequences in aligned parallel corpora of Japanese and English. Translation candidates of word sequences are evaluated by a similarity measure between the sequences defined by the co-occurrence frequency and independent frequency of the word sequences. The similarity measure is an extension of Dice coefficient. An iterative method with gradual threshold lowering is proposed for getting a high quality translation dictionary. The method is tested with parallel corpora of three distinct domains and achieved over 80% accuracy.

key words statistical based NLP, parallel corpus, machine translation, translation patterns, word similarity

1 はじめに

電子化されたテキストが増えるにつれ，自然言語処理におけるテキスト利用技術は不可欠な存在になっている．特に，統計的な手法を利用して，テキストの表層的な解析だけで自然言語処理に有用な言語知識を獲得するための技術の進歩は目覚しい．

さらに，インターネットの普及による情報社会の国際化により，一つの言語を対象にするのではなく，多言語を同時に扱うための研究も盛んになっている．英語，フランス語などのヨーロッパ言語は研究目的のための大規模な対訳コーパスが存在し，文，イディオム，複合名詞，単語単位の対訳間の対応付けに関する種々の研究 [3][6] が行なわれている．これらの言語間の対応付けは，単語の区切りが明瞭である，言語族が近いため一つの概念を表すための言語表現も類似するなど，対応付けにとって有利な特徴を持っているため対応付けは比較的容易である．

一方，日本語と英語間では，大規模な対訳コーパスの入手が困難である，日本語単語の単位が不明瞭である，言語表現の差異が著しいなどの課題があり，前者に比べてあまり研究がなされていない．特に，単語や文といった固定的な単位でなく，単位が不確定な複合語，句単位の対応付けは，文中において意味のあるまとまりを認識し，その単位間で対応付けを行なう必要がある．熊野ら [5] は機械翻訳用の辞書を自動作成するために，日本語の専門用語の対訳を求める方法を提案したが，抽出対象となる日本語の専門用語はあらかじめ抽出しておき，それに対応する英語訳語を求めるという方法を用いている．

本稿は，あらかじめ抽出対象を限定しない．文対応済みの日本語と英語の対訳コーパスにおいて、日本語単語と英単語間の共起頻度を利用して対応付けを行なうことにより，一語以上の任意の数の単語列を単位とした対訳表現の対応付けの方法を提案する．

2 対訳単語間の類似度計算

対訳コーパスが既に文単位で対応が付けられている場合，二言語の単語間の類似度は，各単語がそれぞれ独立に出現する回数と対訳文に同時に出現する回数から求めることができる．2.1 では，相互情報量を用いる方法 と，Dice-coefficient を用いる計算方法について紹介し，2.2 では Dice-coefficient を拡張した我々の計算方法について説明する．

2.1 従来研究

相互情報量を用いた単語間の類似度の計算は一般に以下の式 (1) が用いられる [2]．x, y を任意の 2 つの単語とすれば，$P(x)$ と $P(y)$ は 単語 x，単語 y が独立に起こる生起確率，$P(x,y)$ は，x と y の同時確率である．この式は 2 つの単語が同時に出現する場合のみに着目する相互情報量による 2 単語間の類似度を求める式である．

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Brown らはフランス語から英語への単語の対応付けを行なうために，フランス語から英語への翻訳方向を考慮して上記の式を次のよう拡張にした [1]．ここでフランス語 f が英語 e_j に翻訳される場合，相互情報量は以下 (2) のように定式化される．$P(e_j \mid f)$ と $P(e_j)$ は e_j と f が独立に出現する回数と対訳文に同時に出現する回数から統計的に求めることができる．

$$MI(e_j, f) = \log \frac{P(e_j \mid f)}{P(e_j)} \tag{2}$$

一方，Kay & Röscheisen は，以下のように Dice-coefficient を用いて英語 w_e とフランス語 w_f の類似度を定義した [3]．$f(w_e)$ と $f(w_f)$ は，w_e および w_f が独立に出現する回数であり，$f(w_e, w_f)$ は，対訳文に同時に出現する回数である．

$$sim(w_e, w_f) = \frac{2f(w_e, w_f)}{f(w_e) + f(w_f)} \tag{3}$$

対訳単語間の類似度を求める式として，相互情報量による式と Dice-coefficient による式を比較した 2 つの興味深い研究がある．

Smadja[6] らは，式 (1) と式 (3) を自らの英仏のイディオム対応付けアルゴリズムに適用し，両者の比較を行なった．その結果，対訳単語の対応付けには，相互情報量より Dice-coefficient の方が有効であることを示した．抽出ペア 43 例中，Dice による結果は正解例が 36，不正解の例が 7 であったのに対し，その正解例 36 中，相互情報量による結果は不正解例が 10 もあった，また Dice の不正解の例 7 は，相互情報量によってもすべて不正解であった．

彼らは Dice-coefficient による式が相互情報量による式より優れている理由として，対訳関係では二言語間の対応関係が両方向で同等でなく，片言語方向からの対応が強い場合が多いことを挙げている．式 (1) では最悪の場合 ($P(x,y) = P(x)P(y)$)，類似度は 0 となってしまうが，これは両者に対訳関係がないことを意味しない．

たとえば，極端な場合であるが，“paper” のみしか出現しない英語コーパスにおいて，“paper” は日本語コーパスでは「紙」，「新聞」，「論文」の3つの日本語に翻訳されているが，「紙」はすべて “paper” に翻訳されているとする．この場合，$P(paper, 紙) = P(紙)$, $P(paper) = 1$ となり，“paper” は「紙」の訳語であるにもかかわらず，その類似度は0となってしまう．

一方，大森らは式(2)と式(3)を用いて，仏単語と英単語間の対応付けを行なった[8]．出現回数2回以上の仏単語と英単語を対象にした実験で，ある英単語に対して類似度の大きい順に仏単語を並べ，上位3位までに正しい対訳仏単語が現れた正解率を測定した結果，式(2)では60.6%，式(3)では65.0%の正解率を得ており，本実験でもDice-coefficientの有効性が示された．

2.2 出現回数を重視した対訳単語間の類似度

上記の研究結果に基づいて，適切な結果を効率良く得るために，式(3)を拡張した以下の式(4)(5)を対訳単語間の類似度を求める式として定義する．

$$sim(\langle w_J, w_E \rangle) = (log_2 f_{je}) \frac{2f_{je}}{f_j + f_e} \quad (4)$$

$$sim(\langle w_J, w_E \rangle) \geq log_2 f_{min} \quad (5)$$

w_J: 日本語単語列　w_E: 英語単語列
f_j: w_J の出現回数　f_e: w_E の出現回数
f_{je}: w_J, w_E 対応の出現回数
f_{min}: 出現回数の閾値

式(4)は，式(3)を出現回数 f_{je} で重み付けることにより拡張した．これは次の2つの理由による．第1の理由は，式(3)では英単語2回，日本語単語2回，英日単語同時2回出現も，英単語100回，日本語単語100回，日英単語同時100回出現も類似度は同じ1となるが，出現回数100回の方が2回に比べて類似度の信頼性が高いと言ってよい．このような経験則を式に反映させるために，出現回数による重み付けを行なう．対数をとるのは，出現回数が小さい場合に出現回数の重みが効くようにするための補正である．

第2の理由は計算量の削減である．本研究で扱う対訳コーパスの対応付けの対象は，英語，日本語コーパスの2回以上出現する1語以上の任意長の単語列である．この場合，すべてを一度に計算する方法では計算量の爆発は避けられない．そこで出現回数の条件を徐々に緩めることにより，候補対象を徐々に広げていくという方法をとる．

具体的な方法を述べると，対訳コーパスから出現回数の閾値 f_{min} を満たす日本語単語列と英語単語列を抽出し，それらの対訳単語間の類似度(式(4))を計算する．その際，式(5)の条件を課する．

式(5)を課する理由は次のとおりである．$f_{je} \leq f_{min}$ の場合，$sim(\langle w_J, w_E \rangle) \leq log_2 f_{min}$ である．このため式(5)の条件を満たすペアは，f_{min} 以下の出現回数を持つペアを候補としても，$log_2 f_{min}$ 以上の値にはならない．すなわち，出現回数を下げてもそれ以上の類似度をもつ対訳表現が計算されることはない．したがって，ある出現回数の閾値を満たす単語列ペアにおける類似度を式(4)(5)で計算して対訳表現を求め，対訳表現が抽出されなくなったら出現回数の閾値を下げて計算を繰り返すという処理を行ない，計算の対象を徐々に広げながら，類似度の高い対訳表現から順番に抽出することができる．また，精度が保証されなくなった段階で抽出処理を中止することもできる．

抽出処理の繰り返しにおいて，抽出した対訳表現を計算対象から外しながら処理を進めていくことは計算量の軽減につながる．また，信頼性の高い対訳表現を取り除きながら抽出を進めることにより，残りの候補の対応付けの曖昧性を減らすこともできる．さらに，本処理によって，ある単語列の第一候補以外の複数の訳語も抽出することもできる．

3 対訳表現の自動抽出

以下の手順にしたがって，対訳表現を自動抽出する．図1に自動抽出の概略を示す．使用する対訳コーパスは日本語と英語の対訳コーパスとし，文単位の対応付けは既に行なわれているものとする．

本処理は，

1. 対訳表現候補となる日英単語列の作成
2. 類似度計算による最適ペアの抽出
3. 対訳表現の出現形への復元

の3つの処理からなり，2. の処理は，出現回数の閾値を徐々に下げながら繰り返し行なわれる．

3.1 日英単語列の作成

(1) 形態素解析・自立語抽出　日本語コーパスと英語コーパスを形態素解析し，自立語(日本語は名詞，動詞，形容詞，形容動詞，副詞，英語は名詞，動詞，形容詞，副詞)を抽出する．

(2) 単語列の抽出　各コーパス中の自立語の出現回数を数え，2回以上出現した単語のみを抽出する．次にその単語を先頭とした2単語からなる単語列を作成し，そ

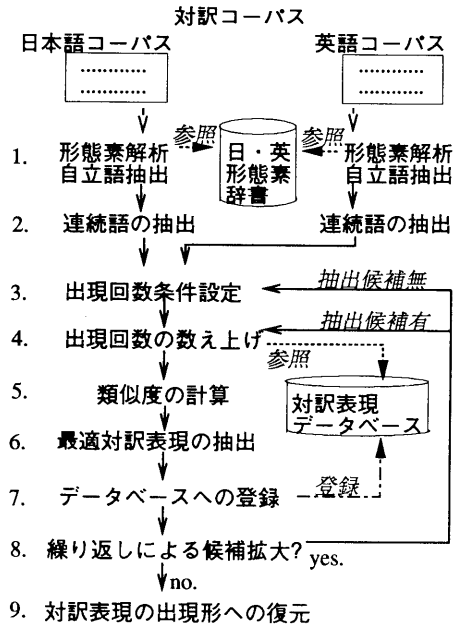

図 1: 対訳表現の自動抽出の概略

の出現回数を数える．さらに2回以上出現した2単語からなる単語列を先頭とする3単語からなる単語列に対しても同様に処理し，以降単語列の構成単語数を拡張しながら各単語列の出現回数を数える．あらかじめ構成単語数の最高数を設定しておき，その値に達すれば処理を終了する．

3.2 類似度計算と最適ペアの抽出

(3) 出現回数条件設定 日本語，英語の単語列の出現回数の閾値 f_{min} を設定する．2回目以降は，前回の f_{min} より小さい値を設定する．

(4) 出現回数の数え上げ (3)の出現回数の閾値を満たす日本語，英語の単語列において，各単語列の出現回数および対訳文内に出現する日本語単語列と英語単語列のペアの出現回数を数える．ただし，両者の対応が対応が既に対訳表現データベースに登録されている場合は，数え上げの対象から省く．たとえば，

a. テキスト解析 :Text analysis
b. 歴史の教科書 :A history text

という2つの対訳文があるとする．データベースに何も登録されていない場合，$sim(\text{教科書}, text) = \frac{2\cdot1}{1+2} = \frac{2}{3}$ であるが，「テキスト:text」がデータベースに登録されている場合は，a. 中の「テキスト」と“text”は，数え上げの対象から省かれるので，$\frac{2\cdot1}{1+1} = 1$ となる．このように，対訳表現データベースに登録されている対訳表現が対訳文に同時に出現する場合を数え上げの対象から外すことにより，類似度が変化し，新たな対訳表現を抽出することができる．

(5) 類似度の計算 (4)における対訳コーパス内の日本語単語列 w_J，英語単語列 w_E の対訳間の類似度を計算する．類似度は，w_J, w_E が出現する回数とその対応が出現する割合を利用するだけでなく，その対応の出現回数も考慮した2.2の式 (4)(5) を用いる．

(6) 最適ペアの抽出 計算された w_J, w_E の対訳単語間の類似度の計算結果から対訳表現として最適な対応を次の方法で求める．

w_E に対応するすべての $WJ = \{w_{J1}, w_{J2}, ..., w_{Jn}\}$ の中で最大の類似度をもつ $\langle w_{Jt}, w_E\rangle (w_{Jt} \in WJ)$ を取り出す．w_{Jt} からも同様に，最大の類似度をもつ対応 $\langle w_J, w_{Es}\rangle$ を取り出す．その結果，$w_E = w_{Es}$ を満たすとき，$\langle w_J, w_E\rangle$ を対訳表現とする．

(7) データベースへの登録 (6)で抽出した対訳表現を対訳表現データベースに登録する．

(8) 繰り返しによる候補の拡大 現在の閾値において対訳表現が抽出できる限り，(4)以降の処理が繰り返される．

一方，新しい対訳表現が抽出できないならば，(3)の処理に戻る．(3)では出現回数の閾値 (f_{min}) が再設定され，(4)で新たな単語列の候補が抽出され，以降 (8) までの処理が繰り返される．

出現回数の閾値 f_{min} が2回に達したならば，(9)の処理に進む．

3.3 対訳表現の出現形への復元

(9) 対訳表現の出現形への復元 対訳表現データベースに登録された対訳表現を，もとの対訳コーパスを参照し，機能語等を補って対訳コーパスに出現した形に復元する．

4 対訳表現の抽出実験

分野の異なる3種類の対訳コーパス (ビジネス例文集，計算機マニュアル，科学技術論文) を用いて，対訳表現

コーパス名	対訳文数	単語				単語列 (10 ≥ 構成単語数 ≥ 1)	
		自立語数		出現回数 ≥ 2		出現回数 ≥ 2	
		英語	日本語	英語	日本語	英語	日本語
ビジネス文書	10,016	2,300	3,739	2,218	3,568	73,026	72,574
科学論文	9,792	7,254	9,415	6,764	8,856	27,329	37,258
マニュアル	11,477	3,701	4,926	3,478	4,799	32,049	38,796

表 1: 単語・単語列の抽出結果

の抽出実験を行なった．以下に実験方法とその結果および考察について述べる．

4.1 実験方法

3種類の対訳コーパス，「取引条件表現辞典例文」[7] (以下ビジネス文書と呼ぶ)，「計算機マニュアル」(以下マニュアルと呼ぶ)，「科学技術論文」(以下科学論文と呼ぶ) [1] は文単位で対応付け，重複する文を一文にまとめたものを使用する．各対訳コーパス中の各文を形態素解析し，自立語のみを抽出した．日本語，英語の形態素解析には，機械翻訳システム PENSÉE[2] の形態素解析部を使用した．対訳コーパスの文数は，重複する文を一文にまとめた結果，ビジネス文書は 10,016 文，科学論文は 9,792 文，マニュアルは 11,477 文であった．

実験における各種の設定は次のとおりである．対訳表現の見出しとなる単語列の最大構成単語数は 10 単語とする．出現回数の閾値 f_{min} は，対訳コーパス中の最高出現回数 Max をもつ単語を求め，$f_{min} = Max/2$ から f_{min} を開始する．$f_{min} > 10$ までは，閾値を $f_{min}/2$ に更新しながら処理を繰り返し，10 以下になれば $f_{min} - 1$ に更新し，f_{min} が 2，または，4.2 で説明する適合率が 80% となった段階で抽出処理を中止する．また，各閾値の抽出処理において，抽出された対訳表現の数が 30 以下になった場合，閾値の更新を行なう．

表 1に，各対訳コーパスに含まれる自立語の出現単語数，および，2 回以上出現した自立語の単語数，その単語列の数を示す．一単語の場合における自立語の出現単語数を比べると，ビジネス文書やマニュアルでは少ないが，科学論文では多い．この理由は，前者の 2 つは統一した内容に関する文書であるため類似の文が数多く存在し，同じ単語が何回も使用されているのに対し，科学論文は分野が多岐に渡っているため出現単語が変化に富み，出現単語数が多くなると考えられる．一方，単語列の場合では，逆に科学論文よりビジネス文書，マニュアルの方が多い．この理由は，科学論文は類似する文がほとんどなく 1 回しか出現しない単語列が大半を占めるためその数は少なくなるのに対し，ビジネス文書やマニュアルは類似の文が多く前後 1 語のみが異なるような単語列が数多く抽出されるためであると考えられる．

4.2 結果と考察

表 2に各対訳コーパスにおいて抽出された対訳表現の数を出現回数の閾値段階別に表した結果を示す．「閾値」には出現回数の閾値 f_{min} の各段階を，「抽出数」は各段階で抽出された対訳表現の数を記す．

本実験では，完全な対訳関係にあり，そのままの形で翻訳用辞書として使用できるものを「正解」とし，日本語，英語のどちらかに足りない単語が存在する，または，対訳関係にあるが日本語，英語どちらかの単位が不適切であるものを「半正解」として，人手により評価した．表 2中の「正解数」，「半正解数」はそれぞれの評価結果の数を表す．なお，表 2中の * 印は，任意の 500 ペアを抜き出して検査した結果の数である．最右桁の「適合率」は，最上位の閾値から各段階まで累積した「抽出数」に対する「正解数」の割合である．なお，() 内は，「半正解」も含めた適合率である．

表 2より，ビジネス文書では出現回数 3 回以上において，マニュアル，科学論文では出現回数 2 回以上において，80% 以上の適合率で対訳表現を抽出することができた．閾値別に抽出結果をみると，マニュアル，ビジネス文書の結果では出現回数の高い閾値での適合率は高く，閾値が低くなるにしたがって適合率が低くなる．しかし，この傾向は科学論文では見られず，全体的に適合率 90% 以上と高精度である．これには次の理由が考えられる．科学論文は固定化された繰り返し出現する表現が少なく，

[1] 「計算機マニュアル」(以下マニュアルと呼ぶ) は沖電気工業株式会社製の計算機のオンラインマニュアルを，「科学技術論文」は "Scientific American" および「日経サイエンス」1 年分を OCR で読み込んで手修正したものを利用した．

[2] PENSÉE は，沖電気工業株式会社，大阪ガス株式会社および株式会社オージス総研の登録商標

ビジネス文書

閾値	抽出数	正解数	半正解数	適合率 (+ 半)
1151	2	2	0	100(100)
575	3	3	0	100(100)
287	4	4	0	100(100)
143	12	12	0	100(100)
71	19	18	1	97.5(100)
35	48	48	0	98.9(100)
17	103	101	2	98.4(100)
10	164	155	8	96.6(99.7)
9	53	51	2	96.6(99.8)
8	67	63	4	96.2(99.8)
7	82	75	6	95.5(99.6)
6	134	114	20	93.5(99.7)
5	163	145	15	92.6(99.4)
4	318	257	50	89.4(98.6)
3	755	502	195	80.4(96.1)
合計	1,927	1,550	302	80.4(96.1)

科学論文

閾値	抽出数	正解数	半正解数	適合率 (+ 半)
68	1	1	0	100(100)
34	21	19	1	90.9(95.5)
17	69	64	5	92.3(98.9)
10	142	133	8	93.1(99.1)
9	52	49	3	93.3(97.9)
8	69	69	0	94.6(99.2)
7	66	63	2	94.7(99.0)
6	105	99	6	94.7(99.2)
5	168	155	12	94.1(99.1)
4	292	263	25	92.9(99.0)
3	536	494	34	92.6(98.4)
2	1,307	(445)*	(46)*	91.7(98.1)
合計	2,828	—	—	91.7(98.1)

マニュアル

閾値	抽出数	正解数	半正解数	適合率 (+ 半)
209	1	1	0	100(100)
104	4	4	0	100(100)
52	19	19	0	100(100)
26	55	54	0	98.7(98.7)
13	145	140	5	97.3(99.6)
10	81	76	5	96.4(99.7)
9	58	55	2	96.1(99.4)
8	75	68	5	95.2(99.1)
7	106	99	7	94.9(99.3)
6	126	118	7	94.6(99.3)
5	214	198	13	94.1(99.1)
4	367	330	26	92.9(98.5)
3	629	519	97	89.4(98.3)
2	1,401	(395)*	(87)*	87.2(97.9)
合計	3,281	—	—	87.2(97.9)

表 2: 各種文書抽出結果

	英語単語列									
	1	2	3	4	5	6	7	8	9	10
1	823/843	43/58	0/6	0/1	0	0	0	0	0	0
2	32/45	401/450	17/55	1/23	0/5	0/4	0/1	0	0/1	0
日 3	0	79/122	72/90	7/23	0/8	0/4	0/4	0	0	0
本 4	0	6/21	29/45	15/23	2/5	1/2	0/1	0	0	0/1
語 5	0	3/10	2/13	7/14	3/10	2/3	0	0/1	0/1	0/1
単 6	0	0	2/4	2/3	0/1	0/2	0	0/1	0/1	0/2
語 7	0	0/1	0	0/2	0	0/1	0	0/1	0	0
列 8	0	0/1	0	0/1	0/1	0/1	0	0	0	0/1
9	0	0	0	0	0	0	0	0	0	0
10	0	0/1	0/1	0	0	1/1	0	0	0	0/6

表 3: 構成単語数別抽出結果 (ビジネス文書)

コーパス名	英語	(再現率)	日本語	(再現率)
ビジネス文書	867	(39.1%)	1,005	(28.2%)
科学論文	2,240	(33.1%)	2,359	(26.6%)
マニュアル	1,922	(55.3%)	2,224	(46.3%)

表 4: 対訳表現として抽出された単語数と再現率

かつ，複数の訳語の可能性を持たない専門用語が数多く存在する．一方，前者の二つは数多くの専門用語を含むが，表 1の説明でも述べたように，狭い分野に限定された統一的な内容を持った文書であるため，固定的な表現を含む類似の文が数多く存在する．したがって構成単語数の異なる数多くの単語列候補が作成されてしまい，対応の候補が絞りにくい．その結果，表 5の「紛争 (、) 論争 (又は) 意見: dispute(,) controversy (or) difference (which may) arise」の “arise” のように不要な語が含まれた対訳表現が抽出されたり，「n 型シリコン:p type」のような常に同時に出現する不適切な対訳表現が抽出される結果となり，適合率が下がると考えられる．

抽出された対訳表現を語の構成単語数別にみた結果を表 3に示す．表 3によると，日本語と英語が一単語で対応する場合の適合率は 823/843(96.2%) と高い．しかし，構成単語数が多くなるほど適合率は下がる．特に種々の品詞が混在した単語列の場合での間違いが多い．これは，表 5の「操作 (が) すでに: attempt(ed) (on a) socket (on which a) connect operation (had) already」のように，本方法では句のまとまりを無視した単語列も対訳表現の抽出候補とするためである．適合率を上げるためには，句読点など句の切れ目を考慮する，形態素解析結果の品詞情報を参照し，その構成品詞によって単語列を限定するなどの方法が考えられる．また，構成単語の差が大きいほど適合率が下がることから，構

成単語の差が大きいものは対訳表現としない方法も有効だと思われる．

抽出された対訳表現で特徴的な例を表5に示す．どのコーパスでも，専門性の高い対訳表現が数多くみられた．マニュアルでは，日本語が英文字，カタカナ文字からなる対訳表現が多く，ビジネス文書では，熟語からなる対訳表現が多かった．一方，科学論文では，「出血熱」のような最近話題となった専門用語も見られる．

マニュアルにみられる「インターネット」を含む3つの対訳表現の例にみるように，ある一単語に関係した複数の対訳表現を抽出することができた．また，一つの対訳コーパス内においても後ろに接続する語によって訳語が異なる「営業秘密:trade secret」と「営業時間:business hour」のような例も抽出することができた．

表6は，ビジネス文書における上位15位の対訳表現である．表6の例のように，どの対訳コーパスでも最初の段階では一単語の対訳表現が抽出される傾向が見られた．ビジネス文書でははは出現回数71回の段階(第21位)ではじめて2語以上の単語から構成される「技術情報:technical information」(類似度 7.08)が抽出された．

表4は，本実験の再現率を示したものである．各日本語，英語の数は，少なくとも1つの対訳表現に出現した単語の総数であり，()内の再現率は，各対訳コーパスに2回以上出現した単語の総数を全体とした割合である．再現率が低い理由は，3.2(6)での条件が厳しいためだと考える．この条件を緩めることにより再現率を向上させることができるが，条件の緩和は適合率の低下を伴ってしまう．これは今後の課題である．

5 まとめ

本稿は，対訳コーパスに出現する任意長の単語列における二言語間の自動対応付けの方法を提案した．本方法は，類似度の高いペアから順に，繰り返し抽出する方法を用いることにより，精度の高いペアから順に抽出することができる．

ある文書の翻訳に用いられる訳語や表現方法は分野に依存する．したがって，本方法で文書ごとに自動抽出した対訳表現を分野限定の翻訳辞書として使用することによって，機械翻訳システムの訳質の向上に直接貢献することができる．さらに，複数語からなる固有名詞，専門用語や固定的な言い回しを 1つの単位とみなし，1つの訳語を与えることは，変換処理だけでなく構文解析処理の軽減にも役立つと思われる．

しかし，3.2(4)で例に挙げた「text:テキスト，text:教科書」のような同一文書内での曖昧性は本方法だけでは解消できない．この曖昧性を解消するためには，[9]で提案した，構造照合に基づく依存構造までも考慮に入れた対訳表現の抽出処理によって，その依存構造や修飾語を抽出し，それらの情報に基づいた訳語選択を必要とする．ただし，その場合でも構造照合時に用いる単語間の類似度の単位を単語列に拡張することによって，本方法で抽出した対訳表現を使用することができる．さらに両者の方法を融合することにより，単語，単語列，依存構造など種々のレベルの辞書情報を持った機械翻訳システム [4] の構築も考えられる．

今回の実験では，形態素解析の品詞結果や市販の翻訳辞書など既存の言語知識はできるだけ用いない方針で実験を行なったが，今後は，既存の言語知識を効果的に利用し，適合率，再現率のさらなる向上を図っていきたいと考える．

参考文献

[1] P.F. Brown. A Statistical Approach to Language Translation. *COLING-88*, volume 1, pages 71–76, 1988.

[2] K.W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[3] M. Kay and M. Röscheisen. Text-Translation Alignment. *Computational Linguistics*, 19(1):121–142, 1993.

[4] M. Kitamura and Y. Matsumoto. A Machine Translation System based on Translation Rules Acquired from Parallel Corpora. *Recent Advances in Natural Language Processing*, pages 27–44, 1995.

[5] A. Kumano and H. Hirakawa. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. *COLING-94*, volume 1, pages 76–81, 1994.

[6] F. Smadja, K.R. McKeown, and V. Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1–38, 1996.

[7] 石上進. 取引条件表現法辞典 電子ブック版 第*1*巻 物品取引. 国際事業開発株式会社, 1992.

[8] 大森久美子, 堤純也, 中西正和. 統計情報を用いた対訳単語辞書の作成. 言語処理学会 第*2*回 年次大会発表論文集, pages 49–52, 1996.

[9] 北村美穂子, 松本裕治. 対訳コーパスを利用した翻訳規則の自動獲得. 情報処理学会論文誌, 37(6):1030–1040, June 1996.

	日本語	英語	類似度
— ビジネス文書 —			
△	紛争 (、) 論争 (又は) 意見	dispute(,) controversy (or) difference (which may) arise	4.34
	営業 秘密	trade secret	3.72
	契約 (の) 発効 日	effective date (of this) agreement	3.12
	営業 時間	business hour	2.92
	確認 (付) 取消 不能 信用状	irrevocable confirm(ed) letter (of) credit	2.81
	技術 製造 ノウハウ	technique manufacture know-how	2.62
	特許 (、) ノウハウ (又は) 技術 情報	patent(s)(,) know-how (or) technical information	1.06
— 科学論文 —			
	出血 熱 ウイルス	hemorrhage fever virus	3.19
	アクリル 酸 メチル	methyl acrylate	3.17
	ロスアラモス 国立 研究所	Los Alamos national laboratory	2
△	n 型 シリコーン	n type	1.78
*	n 型 シリコーン	p type	1.78
	カリフォルニア 大学 デービス 校	university (of) California (at) Davis	1.58
	ワイヤレス ネットワーク	wireless network	1.19
— マニュアル —			
	インタネット アドレス	internet address	2.83
	倍精度 浮動 小数点	double precision float point	1.79
	インタネット プロトコル ＩＰ	internet protocol IP	1.78
	インタネット プロトコル	internet protocol	1.66
*	ホスト テーブル	DoD internet	1.6
	ネーム トゥ アドレス マッピング	name (to) address map(ping)	1.58
*	操作 (が) すでに	attempt(on a)socket(on which a)connect operation already	1.36

△ は「半正解」を * は「不正解」を示す

表 5: 対訳表現の抽出例

日本語	英語	類似度	ペア出現回数	日本語出現回数	英語出現回数
— 1151 回以上 出現ペア —					
会社	company	10.73	3,952	4,081	4,720
ライセンシー	licensee	10.47	2,436	2,521	2,715
— 575 回以上 出現ペア —					
販売店	distributor	9.55	1,471	1,562	1,679
契約 品	product	9.26	2,511	2,996	3,127
売り手	seller	9.24	999	1,039	1,116
— 287 回以上 出現ペア —					
買い手	buyer	8.92	940	970	1,112
当事者	party	8.84	1276	1,394	1,584
書面	writing	8.39	754	860	858
条	article	8.34	778	837	955
— 143 回以上 出現ペア —					
b	b	8.07	332	345	344
a	a	8.01	324	335	340
ＡＢＣ	ABC	7.99	354	362	388
情報	information	7.87	489	549	561
ＸＹＺ	XYZ	7.77	327	333	370
特許	patent	7.65	455	545	505

表 6: 対訳表現の抽出例 (ビジネス文書 (上から 15 位)