共起語集合の類似度に基づく対訳コーパスからの対訳語抽出

梶 博行† 相薗敏子†

対訳辞書は機械翻訳システムや多言語情報検索システムの重要な構成要素である．基本対訳辞書の増補や専門用語対訳辞書の作成を自動化することを目的として，対訳コーパスから語の対訳関係を抽出する新しい方法を開発した．本方法は，コーパス中で共起している語の集合で語を特徴付け，共起語集合の類似度が高い語のペアを対訳語ペアとして抽出する．異なる言語の語を構成要素とする共起語集合の類似度を計算するため，既存の対訳辞書を参照して対訳関係が成立する語を対応付ける．共起語集合の類似度計算という統計処理の中で既知の対訳知識を利用することにより，次の長所をあわせ持つ方法が実現できた．第1に，文レベルの対応付けがなされていない対訳コーパスに適用可能である．第2に，小規模な対訳コーパスから対訳語ペアを抽出することができる．第3に，未知語を含む単純語と複合語の任意の組合せの対訳語ペアを抽出することができる．日英対訳の特許明細書コーパスを用いて，既存の対訳辞書（50,000語の見出し語を持つ日英機械翻訳システムの対訳辞書）に未登録の対訳語ペアを抽出する実験を行った．33.8%の抽出率，76.7%の正解率を達成し，提案方法が実用に供しうるとの結論を得た．本方法は，大規模な対訳コーパスを要求せず，対訳文書を個別に処理していけばよいので，実際的である．今後の課題として，コーパスからの複合語抽出精度を向上させることがあげられる．

Extracting Word Translations from Bilingual Corpora Based on Similarity of Co-occurring Word Sets

Hiroyuki Kaji† and Toshiko Aizono†

A new method has been developed for extracting pairs of words that are translations of each other from a parallel corpus. First, for each word of both languages, the set of words co-occurring with it is extracted from the corpus. Then, the similarity between each pair of co-occurring word sets, one for a word of the first language and the other for a word of the second language, is calculated with the assistance of an existing bilingual dictionary of basic words. Finally, pairs of words that bear much similarity are selected. The method has the following features due to the combined use of co-occurrence information given by a corpus and bilingual knowledge given by an existing dictionary. It can extract word translations from rather small, unaligned corpora; it can extract a variety of word translations including pairs of simple words, pairs of compound words, and mixed pairs of simple and compound words. An experiment using Japanese-English patent specification documents achieved 33.8% recall and 76.7% precision; this demonstrates that the method is useful both for improving the coverage of an existing bilingual dictionary and for creating a bilingual dictionary of technical terms. A further problem is to improve the method for extracting compound words from corpora.

1. はじめに

機械翻訳システムや多言語情報検索システムでは，対象分野の膨大な語彙をカバーする対訳辞書を作成することが必要である．専門用語が次々と生み出される分野の場合，辞書のメンテナンスも重要である．対訳辞書の作成・メンテナンスをできるだけ自動化することが望まれる．一方，最近は多くの文書が電子的な手段で作成され，電子化された対訳文書が増加している．このため，対訳コーパスから対訳語の知識を自動的に抽出する技術の研究がさかんになっている．

対訳コーパスから対訳語知識を抽出する研究は，統計処理によるアプローチ[1)~6)]と言語知識を利用するアプローチ[7)~9)]に大別される．

統計処理のアプローチは，対訳コーパスでの出現頻度や出現位置に基づいて両言語の語の相関を計算し，相関度の高いペアを抽出する．良い結果を得るには，一般に大規模なコーパスが必要である．しかし，大規模な対訳コーパスは意外に少なく，適用分野が限られ

† 日立製作所中央研究所
Central Research Laboratory, Hitachi, Ltd.

るという問題がある．また，これまでに提案された多くの方法は，文と文が対応付けられた対訳コーパスを前提としている．対訳コーパスの文レベルでの対応付けを自動化する技術もよく研究され，一定の成果が得られている[10)~13)]．しかし，文対応が 1 対 1 でなかったり，一方の言語のテキストに欠落があったりすると，対応付けの精度が低下するという問題がある．

言語知識を利用するアプローチは，既存の対訳辞書を参照することにより，構成要素レベルで対訳関係が成立する複合語のペアを抽出する．この方法は，小規模な対訳コーパスからでも対訳語のペアを抽出することができる．しかし，単純語の対訳ペアを抽出することはできない．また，複合語の対訳ペアでも，構成要素のレベルで素直な対応関係が成立しないものは抽出することが困難である．

本論文では，上記従来技術の問題点を解決する新しい方法を提案する．すなわち，文の対応付けがなされていない小規模な対訳コーパスから，単純語と複合語の両方を対象として対訳語ペアを自動抽出する方法を提案する．基本対訳辞書がすでに存在していることを前提とし，新しい対訳語ペアの追加登録や専門用語対訳辞書の作成というタスクを想定している．以下，2 章で基本的なアイデアを述べ，3 章で提案方法を詳細に説明する．4 章で日英対訳の特許明細書コーパスを用いた評価実験を報告し，5 章で提案方法の特徴と効果，改良の方向について述べる．最後に，6 章で関連研究と比較する．以下の記述では，対訳コーパスの言語対を日本語–英語としているが，提案方法は任意の言語対に適用可能である．

2. 基本アイデア

2.1 着眼点—共起語集合の類似度

対訳コーパスを構成する 2 つの言語のテキストは，同一の内容を記述している．したがって，対訳関係にある語が出現している文脈は同じである．それぞれの言語で表現されているという問題はあるが，出現文脈を語の特徴と考えれば，対訳関係にある語は高い類似度を示す．ただし，近傍に出現している語の出現文脈もほとんど同じであるから，近傍の語と対訳関係にある語とも類似度が高くなる．したがって，語の出現例（token）に対して対訳語を同定することは難しい．

そこで，異なり語（type word）ごとに出現文脈を累積することを考える．ある出現例の近傍に出現している語が，ほかの出現例の近傍にも必ず出現するということはない．したがって，出現文脈を累積すれば，同一言語内の語を区別するのに十分な情報を得ることができる．両言語の語を累積出現文脈で特徴付け，その類似度を計算することによって，対訳語のペアを抽出する．

語の出現文脈を，その語と共起している語の集合として表現すれば，累積出現文脈は，各出現例に対する共起語の集合の和集合となる．ここで，共起頻度が有効な情報となるので，通常の集合ではなく頻度付きの集合を採用する．共起頻度分だけ同一の語を重複して含む集合である．なお，共起語は名詞，動詞，形容詞などの内容語に限定し，文法的な役割を担っている機能語は除外する．

異なる言語の語を構成要素とする共起語集合の類似度を計算するためには，対訳辞書を利用する．対訳辞書を介して対応付け可能な語を共通の要素と考えることにより，類似度を計算することができる．

以上の考えに基づいて，日本語の語 x と英語の語 y の相関度 $\alpha(x, y)$ を次式で定義する．

$$\alpha(x,y) = \frac{|C(x) \cap C(y)|}{|C(x) \cup C(y)|} = \frac{|C(x) \cap C(y)|}{|C(x)| + |C(y)| - |C(x) \cap C(y)|} \quad (1)$$

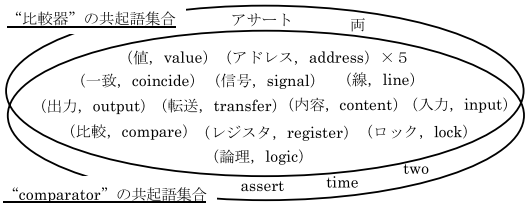
ここに，$C(x)$，$C(y)$ はそれぞれ x，y の頻度付き共起語集合である．式 (1) は Jaccard 係数と呼ばれる類似性の尺度である．ただし，和集合や積集合の演算が，対訳辞書を参照して $C(x)$ の要素と $C(y)$ の要素を対応付ける処理を含むことが，通常と異なっている．

語の相関度，すなわち共起語集合の類似度の計算例を**図 1** に示す．図 1 (a) のサンプル対訳コーパス中の “比較器” の共起語集合と “comparator” の共起語集合，それらの類似度を図 1 (b) に示す．同様に，“比較器” の共起語集合と “assert” の共起語集合，それらの類似度を図 1 (c) に示す．(b) と (c) を比較すると，対訳語ペアの共起語集合が非対訳語ペアの共起語集合より高い類似度を示すことが分かる．図 1 では，同一の文に出現している語を共起語と考えている．この点については 2.2 節で説明する．なお，共起語集合の図の中で，(a, b) は，日本語の語 a と英語の語 b が対訳辞書を介して対応付けられたことを表している．

相関度 α は，対訳コーパスに含まれる語の対応関係に基づいている．したがって，直訳による対訳コーパスに比べて，意訳による対訳コーパスでは有効性が低下する．相関度 α の有効性は，参照する対訳辞書によっても大きく変化する．共起語集合間の語の対応のうち，対訳辞書に登録されているものだけが α の値に寄与するからである．したがって，ある程度のカバー率を持つ対訳辞書が利用できることが前提となる．

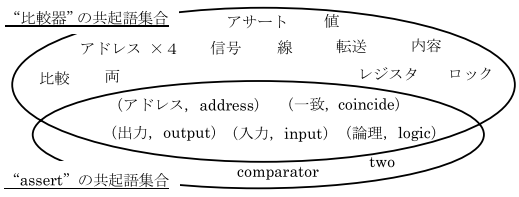
＜日本語テキスト＞
アドレス比較器はアドレス転送中のアドレス信号線の値とロックアドレスレジスタの内容を比較する．アドレス比較器の両入力が一致するとその出力には論理“1”がアサートされる．
＜英語テキスト＞
The address comparator compares the content of the lock address register with the value of the address signal line at the time of address transfer. When the two inputs to the address comparator coincide with each other, logic “1” is asserted as the output.

(a) サンプル対訳コーパス



$$\alpha(\text{比較器,comparator}) = \frac{17}{19+20-17} = 0.77$$

(b) 対訳語ペアの例



$$\alpha(\text{比較器,assert}) = \frac{5}{19+7-5} = 0.24$$

(c) 非対訳語ペアの例

図 1 共起語集合の類似度
Fig. 1 Similarity of co-occurring word sets.

2.2 「共起」の定義

共起語集合の類似度に基づいて対訳語ペアを抽出する場合，共起をどう定義するかが問題である．対訳関係にある語を特徴付ける共起語集合は，互いに対応する範囲から抽出されたものでなければならない．共起語を抽出する範囲が言語間でずれると，共起語集合の類似度が低下し，対訳語ペアを抽出することが困難になる．

共起の定義としては，文共起，ウィンドウ共起，構文共起が考えられるが，それぞれ一長一短がある．

(1) 文共起

同じ文に含まれている語を共起語として抽出する．文の対応が 1 対 1 である場合，言語間で共起語抽出範囲がずれることはない．しかし，文の対応が 1 対 1 でない場合，必ずずれるという問題がある．また，複文の場合を考えると，共起語の抽出範囲として少し広すぎるとも思われる．同一文中の語を区別する情報がほとんど得られないからである．

(2) ウィンドウ共起

一定数の語を収容するウィンドウをテキストに沿って移動させながら，ウィンドウ内に同時に含まれている語を共起語として抽出する．文の対応が 1 対 1 でない場合，文共起より有利であるが，ウィンドウのサイズをどう決めるかが問題である．語を区別する力を高くするという意味では，小さめのウィンドウがよい．しかし，小さいウィンドウでは，言語間で抽出範囲がずれる確率が高くなる．特に，日本語–英語のように構造的な差異が大きい言語対の場合は，信頼性に欠ける．

(3) 構文共起

構文的な依存関係を持つ語を共起語として抽出する．強いつながりを持つ語が抽出されるので，対訳語ペアを抽出する手がかりとしても強力である．しかし，言語間で依存関係が保存されるという保証はない．日本語–英語のように構造的差異の大きい言語対の場合，あるいは意訳の多い対訳コーパスの場合には，厳格すぎると思われる．また，構文解析の処理時間や精度の問題も考慮しなければならない．

以上のとおり，決定的に優位なものはない．本研究では，全体的に安定した動作が期待される文共起を採用する．1 対多あるいは多対多の文対応を含む対訳コーパスでも，全体としては 1 対 1 対応の部分が多い．共起語集合は累積して利用するので，文対応が 1 対 1 でない場合の弱点も致命的ではない．

2.3 対訳語ペアの競合

対訳語ペアの抽出を困難にする要因として，語の対訳関係が 1 対 1 ではなく多対多であることがあげられる．2.1 節で提案した方法では，日本語の語または英語の語を共有する対訳語ペアの競合により，次の問題が発生する．

(1) 抽出対象としての競合

対訳コーパスが 2 つの対訳語ペア (a, b_1)，(a, b_2) を含むとする．a は日本語の語，b_1 と b_2 は英語の語である．このとき，日本語テキストから得られる共起語集合 $C(a)$ は，英語テキストから得られる 2 つの共起語集合 $C(b_1)$，$C(b_2)$ の和集合に対応したものになる．$C(b_1)$，$C(b_2)$ は，ともに $C(a)$ 全体ではなく $C(a)$ の部分集合に対応している．したがって，$\alpha(a, b_1)$，$\alpha(a, b_2)$ はそれほど大きな値にならず，(a, b_1)，(a, b_2) の両方とも対訳語ペアとして抽出されない可能性が高くなる．

(2) 共起語集合の対応付けにおける競合

対訳コーパスが 2 つの対訳語ペア (a, b_1)，(a, b_2) を含むとする．a を含む日本語共起語集合と b_1，b_2 の両方を含む英語共起語集合を対応付ける際，a を b_1 と b_2 に配分しなければならないが，これは局所的に決められない問題である．一般に対訳語ペアの競合は連鎖を形成しているからである．共起語集合の対応付けを

厳密に行おうとすると，組合せ最適化問題になる．本論文では，多少の誤差はやむをえないという立場で，相関度 α の簡易な計算方法を示す．

2.4 コーパスの処理単位

提案方法による対訳コーパスの処理単位は，個々の対訳文書とする．複数の対訳文書を一括して処理することは想定しない．第1の理由は技術的な理由である．1つのテーマについて記述した文書の範囲では，複数の訳語に訳される語の比率が低いので，対訳語ペアの競合による影響を小さくすることができる．第2の理由は利用面からのものである．利用可能な対訳文書を逐次処理し，抽出された対訳語ペアを対訳辞書に登録していくのがよい．大規模な対訳コーパスを用意しないと利用できない方法は実際的でない．

個々の対訳文書を処理単位とすることにより，次の利点が得られる．第1に，抽出率より正解率を重視する方針をとることができる．抽出結果を人手でチェックすることを考えると，ある程度の正解率を確保することが望ましい．抽出できなかった対訳語ペアは，別の対訳文書から抽出されることを期待すればよい．第2に，実装に際して，性能面の工夫をしなくてもよい．1つの対訳文書から抽出される異なり語の数は限られるので，すべての語の組合せについて相関度を計算しても，処理時間やメモリ容量が問題になることはない．

3. 提案方法

3.1 概　　要

提案方法は，**図2**に示すように日本語共起データ抽出，英語共起データ抽出，対訳語抽出の3つのステップから構成される．

(1) 日本語共起データ抽出：日本語テキストを文に分割する．そのあと，各文を構成する語（単純語および複合語）を抽出し，文共起データを抽出する．処理の結果として，各々の語に対する頻度付き共起語集合を出力する．

(2) 英語共起データ抽出：英語テキストに対して(1)と同様の処理を行う．

(3) 対訳語抽出：基本対訳辞書を参照しながら，日本語の語と英語の語のすべての組合せについて相関度を計算する．そして，高い相関を持つペアで基本対訳辞書に含まれていないものを新しい対訳語ペアとして選択する．このようにして抽出した対訳語ペアをフィードバックし（図2中，点線の矢印），対訳語抽出処理を再実行する．

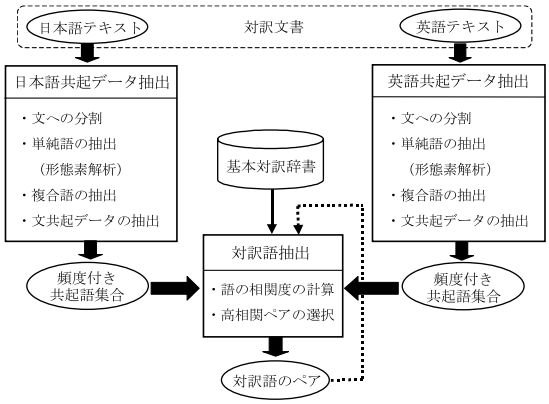


図2 対訳語抽出方法
Fig. 2 Proposed method for extracting word translations.

3.2 語の抽出

(1) 単純語の抽出

形態素解析の結果から，内容語の品詞が付与された語と未知語を抽出する．未知語は名詞である可能性が高いので，内容語と考える．

(2) 複合語の抽出

品詞列のパターンマッチングにより複合名詞を抽出する．次に示すように，日本語と英語で若干異なる品詞列パターンを用いる．

$$JCN := \{N|UK\}\ \{N|UK\}+ \quad (2)$$

$$ECN := \{N|UK|ADJ\}\ \{N|UK\}+ \quad (3)$$

ここに，JCN は日本語複合名詞，ECN は英語複合名詞，N は名詞，ADJ は形容詞，UK は未知語を表す．

上記の品詞列パターンにマッチしても，非最大複合名詞は抽出せず，最大複合名詞のみを抽出する．非最大複合名詞とは，より大きな複合名詞に包含されている複合名詞であり，最大複合名詞とは，ほかの複合名詞に包含されていない複合名詞である．非最大複合名詞を抽出しない理由は，複合名詞の構造的曖昧性の問題があるからである．なお，英語の複合語抽出において，形容詞で始まる複合名詞から形容詞を除いて得られる非最大複合名詞は抽出する．形容詞が複合名詞の単なる修飾語である可能性も高いからである．

(1)，(2)により抽出した日本語の異なり語を，以下では $x_1, x_2, \cdots, x_m$ と記す．同様に英語の異なり語を $y_1, y_2, \cdots, y_n$ と記す．これらの語が対訳語ペア抽出の対象であると同時に，語を特徴付ける共起語集合の要素となる．

ここで，「異なり語」について注意しておく．提案方法では，品詞が違っても，語幹の綴りが同じなら同一の語と考える．たとえば，“増加する”（動詞）と “増加”

（名詞）は同一の語，“increase”（動詞）と “increase”（名詞）も同一の語と考える．対訳語ペアの抽出では品詞は二次的なものであること，英語の多品詞解消の精度が十分でないことが，その理由である．

3.3 共起データの抽出

共起の定義として文共起を採用することを 2.2 節で述べた．出力である頻度付き共起語集合を以下ではベクトル形式で記すことを付け加えておく．

– 日本語共起頻度ベクトル

$$\boldsymbol{f}(x_i) = (f_{i,1}\ f_{i,2}\ \cdots\ f_{i,m})\ (i=1,2,\cdots,m) \quad (4)$$

$f_{i,i'} = x_i$ と $x_{i'}$ の共起頻度

– 英語共起頻度ベクトル

$$\boldsymbol{g}(y_j) = (g_{j,1}\ g_{j,2}\ \cdots\ g_{j,n})\ (j=1,2,\cdots,n) \quad (5)$$

$g_{j,j'} = y_j$ と $y_{j'}$ の共起頻度

3.4 語の相関度の計算

(1) 対訳行列の作成

基本対訳辞書 D から処理対象の対訳文書に関係する情報のみを取り出して，対訳行列 T を作成する．

$$T(i,j) = \begin{cases} 1 & \cdots\ (x_i, y_j) \in D \\ 0 & \cdots\ (x_i, y_j) \notin D \end{cases} \quad (i=1,2,\cdots,m;\quad j=1,2,\cdots,n) \quad (6)$$

(2) 日本語，英語共起頻度ベクトルの補正

– $\sum_j T(i',j) = 0$ なら，$\boldsymbol{f}(x_i)$ の第 i' 要素 $f_{i,i'}$ を 0 にする．

$(i = 1,2,\cdots,m;\quad i' = 1,2,\cdots,m)$

– $\sum_i T(i,j') = 0$ なら，$\boldsymbol{g}(y_j)$ の第 j' 要素 $g_{j,j'}$ を 0 にする．

$(j = 1,2,\cdots,n;\quad j' = 1,2,\cdots,n)$

この処理は，基本対訳辞書を介して相手言語の語に対応付けることができない語を共起語集合から除外することに相当する．その理由は次のとおりである．そのような語は相関度 α の値を低下させるが，低下の度合いは一定でなく，共起頻度によって変動する．したがって，そのような語を共起語集合から除外したほうが，相関度 α の信頼性が高くなる．

(3) 日本語共起頻度ベクトルの擬似共起頻度ベクトルへの変換

$$\boldsymbol{f}'(x_i) = \left(f'_{i,1}\ f'_{i,2}\ \cdots\ f'_{i,n}\right) = \boldsymbol{f}(x_i)\cdot T \quad (i=1,2,\cdots,m) \quad (7)$$

日本語の語を特徴付ける共起語集合は日本語の語を要素とする集合であるが，これを英語の語を要素とする集合に翻訳することに相当する．$f'_{i,j}$ は，x_i の共起語のうち y_j に翻訳可能な語の共起頻度の和である．いわば x_i と y_j の擬似的な共起の頻度である．そういう意味で $\boldsymbol{f}'$ を擬似共起頻度ベクトルと呼ぶ．

ここで，競合する対訳語ペアに関連して擬似共起が過剰に生成される問題について述べる．2 つの対訳語ペア (x_p, y_q)，(x_p, y_r) が基本対訳辞書に含まれているとする．このとき，式 (7) によれば，x_i と x_p の $f_{i,p}$ 回の共起から，x_i と y_q の $f_{i,p}$ 回の擬似共起，さらに x_i と y_r の $f_{i,p}$ 回の擬似共起が生成される．

擬似共起の過剰生成を避けるには，対訳行列を翻訳確率行列にすればよい．$T(i,j)$ の値を，x_i が y_j に翻訳される確率にするのである．しかし，次の理由により式 (6) に示した二値行列を採用した．翻訳確率行列は処理対象の対訳文書に対応したものでなければならないが，翻訳確率を事前に求めることは困難である．そして，対訳文書に対応しない翻訳確率行列を用いると，共起語集合の対応付け（実際には，(4) に示すように擬似共起語集合と共起語集合との対応付け）において，対応付けのもれが頻繁に発生する．

(4) 語の相関度の計算

$$\alpha(x_i, y_j) = \frac{\sum_k \min\{f'_{i,k}, g_{j,k}\}}{\sum_k f_{i,k} + \sum_k g_{j,k} - \sum_k \min\{f'_{i,k}, g_{j,k}\}} \quad (i=1,2,\cdots,m;\quad j=1,2,\cdots,n) \quad (8)$$

式 (8) は式 (1) を近似計算するものである．すなわち，x_i の共起語集合と y_j の共起語集合の積集合演算を，x_i の擬似共起語集合と y_j の共起語集合の積集合演算に置き換えている．

ここで，擬似共起の過剰生成に起因する式 (8) の誤差について述べる．話を簡単にするため，対訳語ペア (x_p, y_q)，(x_p, y_r) が競合し，また (x_p, y_q) や (x_p, y_r) とほかの対訳語ペアは競合しないとする．このとき，x_i の共起語集合が x_p を含み，かつ y_j の共起語集合が y_q と y_r をともに含む場合が問題になる．

式 (8) 中の $\sum_k \min\{f'_{i,k}, g_{j,k}\}$ の項における x_p と y_q，y_r の寄与分は $\min\{f_{i,p}, g_{j,q}\} + \min\{f_{i,p}, g_{j,r}\}$ である．これの正しい値は $\min\{f_{i,p}, g_{j,q} + g_{j,r}\}$ である．したがって，$f_{i,p} < g_{j,q} + g_{j,r}$ のときに過大評価される．ただし，過大評価されても，その値が $g_{j,q} + g_{j,r}$ を超えることはない．一方，(x_i, y_j) が正しい対訳語ペアであれば，多くの場合 $f_{i,p} \fallingdotseq g_{j,q} + g_{j,r}$ であるので，$g_{j,q} + g_{j,r}$ あるいはそれに近い値になる．したがって，$\sum_k \min\{f'_{i,k}, g_{j,k}\}$ が過大評価されたとしても，y_j を固定して比較すると，正しい対訳語ペアに対する値を超える可能性は低い．

以上のように，擬似共起が過剰生成されても式(8)が過大評価されるとは限らない．また，過大評価によって相関度の大小が逆転する可能性は低い．式(8)は式(1)の簡易な計算式として有効と考える．

3.5 高相関ペアの選択

日本語の語と英語の語の組合せ $(x_i, y_j)(i = 1, 2, \cdots, m;\quad j = 1, 2, \cdots, n)$ のうち，基本対訳辞書 D に未登録であって，次の2つの条件を満たすものを選択する．

$$\forall k(\neq j)\quad \alpha(x_i, y_j) > \alpha(x_i, y_k)$$
$$\&\quad \forall k(\neq i)\quad \alpha(x_i, y_j) > \alpha(x_k, y_j) \qquad (9)$$
$$\forall k(\neq j)\quad T(i,k)\cdot\alpha(x_i, y_k) = 0$$
$$\&\quad \forall k(\neq i)\quad T(k,j)\cdot\alpha(x_k, y_j) = 0 \qquad (10)$$

条件(9)は，日本語の語からみても英語の語からみても，相手が最大の相関値を持つペアであることを意味する．また，条件(10)により，基本対訳辞書に既登録の対訳語ペアと競合しないペアに限定する．これらはかなり厳しい条件であるが，抽出率より正解率を重視する考えによる．

3.6 抽出された対訳語ペアのフィードバック

2.1節で述べたように，参照する対訳辞書のカバー率が高いほど，相関度 α の有効性が増す．そこで，抽出された対訳語ペアをフィードバックして，対訳語ペアを再抽出する．具体的には，抽出された対訳語ペアの集合を A とするとき，3.4節の(1)対訳行列の作成における D を $D \cup A$ に置き換える．3.4節の(2)以降と3.5節の処理はフィードバック前と同じである．

なお，抽出された対訳語ペアには正しくないものも含まれるので，フィードバックによって抽出率や正解率が低下する可能性もある．4章の評価実験ではこの点も評価する．

4. 評価実験

4.1 概　　要

提案方法を評価するため実験プログラムを作成し，日英対訳の特許明細書コーパス☆から基本対訳辞書に未登録の対訳語ペアを抽出する実験を行った．

基本対訳辞書としては，ある日英機械翻訳システムの辞書を利用した．日本語の見出し語が約50,000語で，1見出し語あたり平均3.8個，多い語には20〜30個の英訳語が付与されている．

特許明細書は，半導体分野のもの5編を使用した．2.4節で述べた考え方に従って，5編を別々に処理した．実験プログラムによる抽出結果を手作業による抽出結果と比較し，抽出率と正解率を計算した．3.6節で述べたフィードバックの効果を確認するため，フィードバック前後の抽出結果それぞれに対して抽出率と正解率を計算した．

手作業による抽出は，出現例ベースで行った．3.2節で述べた自動抽出の基準に準拠して語を抽出したが，複合語の認定は柔軟に行った．すなわち，式(2)や式(3)のパターンにあてはまらなくても，複合語と思われるものは抽出した．また，最大複合名詞より非最大複合名詞のほうが複合語として適切であると判断される場合，最大複合名詞ではなく非最大複合名詞を抽出した．たとえば，“上記論理回路”や“回路素子数”は抽出せず，“論理回路”や“回路素子”を抽出した．

4.2 特許明細書コーパスの特性

英語明細書は日本語明細書を基に作成されており，基本的には直訳の対訳コーパスである．しかし，英文を推敲しても日本語明細書はそのままであるので，表現や内容が対応しない部分が含まれている．

5編の特許明細書の諸元を**表1**に示す．日本語テキストおよび英語テキストに関する数値は，プログラムで求めた．語の対応関係に関する数値は，手作業で抽出した結果に基づいている．

表1中の特性値 $C4$，$C7$，$C7'$，$C9$ についてコメントしておく．$C4$ は，抽出対象として競合する対訳語ペアが17.4%含まれることを示している．$C7$ と $C7'$ は，語の相関度の計算における誤差の要因となりうる対訳語ペアの比率である．3.4節では，(i)日本語共起頻度ベクトルを擬似共起頻度ベクトルに変換するとしたが，日本語と英語の扱いを逆にし，(ii)英語共起頻度ベクトルを擬似共起頻度ベクトルに変換してもよい．(i)の場合は，$C7$ すなわち24.6%の対訳語ペアが誤差の要因となりうる．(ii)の場合は，$C7'$ すなわち29.5%の対訳語ペアが誤差の要因となりうる．評価実験では，誤差要因が若干少なくなる(i)を採用した．$C9$ は特許明細書コーパスに対する基本対訳辞書のカバー率で，共起語の83.4%が相関度の計算に寄与することが分かる．

4.3 対訳語ペアの抽出結果

抽出結果の評価指標として抽出率（recall）と正解率（precision）を用いた．それぞれ次式で定義される．

☆ 特許明細書は，テキストと図面の間で対応する要素に識別番号を付ける習慣がある．『アドレス比較器504はアドレス転送中のアドレス信号線230の値と ··· / The address comparator 504 compares the content of the lock address register 502 with ···』において下線を付した数字がその例である．識別番号の対応関係は確実であり，共起語集合の要素として有効と思われる．しかし，識別番号は特許明細書特有のものであるので，評価実験には識別番号を除いたテキストを使用した．

表 1 評価実験用特許明細書コーパスの諸元
Table 1 Profile of the patent specification corpus used in the experiment.

		対訳文書	①	②	③	④	⑤	計†)
A1	日本語テキスト	文数	90	120	686	230	178	1,304
A2		内容語総数	1,322	2,089	8,023	3,846	2,449	17,729
A3		平均文長 (A2/A1)	14.7	17.4	11.7	16.7	13.8	13.6
A4		内容語の異なり数	202	273	719	392	524	2,110
A5		平均出現頻度 (A2/A4)	6.5	7.7	11.2	9.8	4.7	8.4
B1	英語テキスト	文数	94	143	704	236	178	1,355
B2		内容語総数	1,463	2,055	9,561	4,326	2,872	20,277
B3		平均文長 (B2/B1)	15.6	14.4	13.6	18.3	16.1	15.0
B4		内容語の異なり数	244	312	936	485	629	2,606
B5		平均出現頻度 (B2/B4)	6.0	6.6	10.2	8.9	4.6	7.8
C1	語の対応関係	対訳語ペアの異なり数	211	316	1,008	608	660	2,803
C2		対訳辞書に未登録のペア	75	126	417	315	302	1,235
C3		うち，他のペアと競合するペア	12	19	85	45	54	215
C4		同上の比率 (C3/C2 [%])	16.0	15.1	20.4	14.3	17.9	17.4
C5		対訳辞書に既登録のペア	136	190	591	293	358	1,568
C6		うち，日本語の語を他のペアと共有するペア	24	28	193	65	75	385
C7		同上の比率 (C6/C5 [%])	17.6	14.7	32.7	22.2	20.9	24.6
C6'		うち，英語の語を他のペアと共有するペア	25	40	245	79	74	463
C7'		同上の比率 (C6'/C5 [%])	18.4	21.1	41.5	27.0	20.7	29.5
C8		単純語どうしのペア	163	221	714	343	438	1,879
C9		辞書カバー率 (C5/C8 [%])	83.4	86.0	82.8	85.4	81.7	83.4

†) ①から⑤の値を単純に加算した．語や対訳語ペアの文書間の重複は考慮していない．

表 2 抽出率と正解率
Table 2 Recall and precision.

(a) 提案方法

		対訳文書	①	②	③	④	⑤	計
C2	文書に含まれる対訳辞書未登録のペア		75	126	417	315	302	1,235
D1	フィードバック前	抽出されたペア	31	53	190	131	100	505
D2		抽出された正解ペア	22	46	144	96	69	377
D3		抽出率 (D2/C2 [%])	29.3	36.5	34.5	30.5	22.8	30.5
D4		正解率 (D2/D1 [%])	71.0	86.8	75.8	73.3	69.0	74.7
E1	フィードバック後	抽出されたペア	31	60	202	140	111	544
E2		抽出された正解ペア	23	50	157	102	85	417
E3		抽出率 (E2/C2 [%])	30.7	39.7	37.6	32.4	28.1	33.8
E4		正解率 (E2/E1 [%])	74.2	83.3	77.7	72.9	76.6	76.7
F1	他の 4 文書によりリカバリされたペア		6	6	9	4	8	33
F2	実質抽出率の向上分 (F1/C2 [%])		8.0	4.8	2.2	1.3	2.6	2.7

(b) 高相関ペア選択時の条件を緩和した場合†)

	対訳文書	①	②	③	④	⑤	計
C2	文書に含まれる対訳辞書未登録のペア	75	126	417	315	302	1,235
G1	抽出されたペア	56	81	329	195	178	839
G2	抽出された正解ペア	27	49	165	115	79	435
G3	抽出率 (G2/C2 [%])	36.0	38.9	39.6	36.5	26.2	35.2
G4	正解率 (G2/G1 [%])	48.2	60.5	50.2	59.0	44.4	51.8

†) 基本対訳辞書に既登録のペアと競合しないという条件をはずした．フィードバックは行わなかった．

– 抽出率 $= |A \cap M| / |M|$
– 正解率 $= |A \cap M| / |A|$

ここに，A は実験プログラムが自動抽出した対訳語ペアの集合，M は手作業で抽出した基本対訳辞書に未登録の対訳語ペアの集合である．A，M とも頻度は付いていない．

表 2 (a) に評価結果を示す．5 編の対訳明細書それぞれに対して，また 5 編の結果の合計に対して抽出率と正解率を計算した．それによると，フィードバックの前では抽出率が 30.5%，正解率が 74.7%であり，フィードバックの後では抽出率が 33.8%，正解率が 76.7%であった．抽出率，正解率ともフィードバックによって向上することが確認できた．なお，フィードバックを 2 回繰り返す実験も行ったが，2 回目のフィードバックで新たに抽出されるペアは少なかった．フィードバックは 1 回だけしか効果がないといえる．

3.5 節で述べた高相関ペアの選択において，既知の対訳語ペアと競合しないという条件 (10) は厳しすぎるとも思われる．そこで，条件 (10) をはずして，条件 (9) のみで対訳語ペアを抽出する実験を行った．結果は，表 2 (b) に示すように，抽出率が 35.2%，正解率が 51.8%であった．表 2 (a) と比較すると，抽出率が少ししか上がらず，正解率が大きく低下している．条件 (10) は必要であるとの結論を得た．

個々の対訳文書では正解率を重視し，抽出率は対訳文書の集合で考えることを 2.4 節で述べた．この方針の妥当性を判断するため，各対訳明細書から抽出できなかった対訳語ペアのうち，ほかの 4 編の対訳明細書から抽出されたものを調べてみた．表 2 (a) の F1，F2 がその結果である．4 編の明細書だけでも，実質的な抽出率が 2.7%向上している．また，5 編の明細書から抽出された対訳語ペアの和集合をとると，少数ではあるが，互いに競合するペアが抽出されていた．これも，個々の対訳文書を処理単位とすることのメリットである．

次に，実験プログラムが抽出した対訳語ペアの例を示す．単純語どうしのペア，複合語どうしのペアに加えて，単純語と複合語のペアも抽出されている．

– 単純語のペア
(排気, pump)，(引き続き, subsequently)，
(フェッチ, fetch)，(容量, capacitance)
– 複合語のペア
(ガス供給機構, gas supplier)，
(桁上げ生成回路, carry generation circuit)，
(高周波加熱, radio frequency heating)
– 単純語と複合語のペア
(圧損, pressure loss)，(薄膜, thin film)
(接続口, connector)，(熱処理, anneal)

抽出された対訳語ペアを構成する単純語は，未知語の場合と既知語の場合がある．上の例において，“フェッチ” や “圧損” は，基本対訳辞書中のどの対訳語ペアにも含まれていなかったという意味で未知語である．これに対し，“排気” や “容量” は，対訳語ペア (排気, exhaust) や (容量, capacity) が基本対訳辞書に含まれていたので，既知語である．ただし，(排気, pump) や (容量, capacitance) は基本対訳辞書に含まれていなかった．このように，既知語であっても未知の訳語があれば抽出の対象となっている．

4.4 ま と め

抽出率 33.8%，正解率 76.7%という結果から，提案方法が基本対訳辞書の増補や専門用語辞書の作成を支援するツールとして実用に供しうるとの結論を得た．提案方法が有効に機能するための，基本対訳辞書のカバー率の下限値，あるいは競合する対訳語ペアの比率の上限値は不明である．しかし，それぞれの値が 80%，20%程度のとき有効に機能することを実験結果は示している．

5. 考 察

5.1 提案方法の特徴とその効果

提案方法は統計処理と言語知識利用の混合型であるが，どちらのアプローチとみても従来の方法と異なっている．

統計的アプローチとしては，従来方法が出現頻度や出現位置で語を特徴付けているのに対し，共起語集合に着目した点が新しい．共起語集合は，出現頻度や出現位置と比べて非常に大きな情報を持っている．このため，1,000 語から 10,000 語程度の小さな対訳コーパスから，頻度 1 のものも含めて，対訳語ペアを抽出することが可能になった．

基本対訳辞書を利用することは，従来の言語知識利用のアプローチと同じである．しかし，利用のしかたがまったく異なっている．従来方法は，語の内部の情報である構成要素の関連度を評価するのに基本対訳辞書を利用している．このため，複合語の対訳語ペアの抽出に限定されている．これに対し，提案方法は，語の外部の情報である共起語集合の類似度を評価するために基本対訳辞書を利用する．ここで，共起語集合による特徴付けは，既知語か未知語かを問わず，単純語にも複合語にも共通に適用することができる．このため，未知語を含む単純語と複合語の任意の組合せの対訳語ペアを抽出することが可能になった．

5.2 改良の方向

(1) 複合名詞の抽出精度向上

3.2 節で述べた簡易な方法では，複合名詞の抽出誤りや抽出もれが避けられない．日本語では，不適切な最大複合名詞（例：“回路素子数”，“絶縁耐圧向上”）が抽出され，抽出すべき複合名詞（例：“回路素子”，“絶縁耐圧”）が抽出されない例が目についた．英語では，品詞列パターンが単純すぎるために抽出できない複合名詞（例：“carry look ahead circuit”，“air operated valve”）が多かった．

4 章の実験で抽出された誤りペアには，部分的には正しいペア（例：(回路素子数, circuit element)）がかなり含まれていた．したがって，複合名詞の抽出精度向上が，対訳語ペアの抽出率，正解率の向上に直結すると思われる．日本語，英語とも，非最大複合名詞を高精度で抽出する必要がある．そのためには，N グ

ラム統計の利用などが考えられる．英語では，多様な複合名詞に対応できるよう，品詞列パターンを補強することも必要である．

(2) ほかの方法との結合

共起語集合の類似度が対訳語ペアの抽出に有効であることを示すことが，本論文の主たる目的であった．このため，ほかの手がかりはあえて利用しなかった．しかし，ほかの方法と結合することで，より良い結果が得られる可能性は高い．すぐに考えられるのは，複合語の構成要素レベルの対応を評価する方法[7),9)]との結合である．この場合，基本対訳辞書は共通に利用することができる．

6. 関連研究との比較

対訳コーパスからの対訳語ペア自動抽出に関する研究は多数報告されているが，パフォーマンスを比較することは難しい．第1の理由は，それぞれ異なるコーパスを用いて評価しているからである．言語対だけでなく，両言語のテキストの対応の程度もさまざまである．第2の理由は，抽出対象とする対訳語ペアの種類など，タスクの設定に違いがあるからである．

ここでは，上記の差異が比較的小さい2つの研究と比較する．

熊野ら[8)]は，既存の対訳辞書と頻度情報を利用する方法を提案している．日英対訳の特許明細書コーパスから日本語の複合名詞と未知語を抽出し，英訳語を推定している．日本語の語の総頻度ベースで正解率を算出し，(a) 総頻度 3,224 の複合名詞に対する正解率が 72.9%，(b) 総頻度 389 の未知語に対する正解率が 54.0%であったと報告している．

提案方法を熊野らの方法と比較するため，4章の評価実験で抽出された対訳語ペアを，(a') 日本語の語が複合名詞，(b') 日本語の語が未知語の単純語，(c') 日本語の語が既知語の単純語の3通りに分類した．そして，分類ごとに，日本語の語の総頻度と頻度重み付きの正解率を算出した．その結果は，(a') 総頻度 1,737 の複合名詞に対する正解率が 88.7%，(b') 総頻度 414 の未知語の単純語に対する正解率が 90.6%，(c') 総頻度 209 の既知語の単純語に対する正解率が 91.4%であった．コーパスサイズの違い（熊野ら：2,148 文，本研究：1,304 文）を勘案して (a) と (a')，(b) と (b') を比較すると，提案方法は，訳語が推定できた語の総頻度が若干少ないが，正解率は高い．語の種類による正解率の変動が小さいことも提案方法の特徴である．

熊野らの方法と提案方法の間には，コーパスの処理単位に関しても違いがある．熊野らの方法では，複数の特許明細書を一括して処理したほうが，個別に処理した場合より，正解率が向上している．上に引用した数値は，一括して処理した場合のものである．一方，提案方法の評価実験では複数の特許明細書を個別に処理した．個別の文書単位で処理する理由や利点は 2.4 節で述べたとおりである．

Fung[5)]は，語の出現位置に注目した統計的方法で，文の対応付けを前提としない方法を提案している．英語–中国語の対訳コーパスから英語の名詞・固有名詞を抽出し，中国語の訳語を推定している．抽出した英語の異なり語 2,779 語のうち，頻度2以上の 661 語に対する正解率が 73.1%であったと報告している．頻度1の語については，訳語を推定することができたものもあるがきわめて難しいと述べていて，正解率は算出していない．

本研究との比較のため，Fung の実験における頻度2以上の語に対する抽出率を推定した．「頻度2以上の英語異なり語が対訳コーパス中でそれぞれただ1つの中国語訳語を持つ」という仮定のもとで，抽出率は正解率と同じ 73.1%である．一方，本研究の実験結果については，対訳語ペアを (i) 日本語の語の頻度が2以上のペアと，(ii) 日本語の語の頻度が1のペアに分け，それぞれについて表2 (a) と同様な表を作成した．それによると，(i) 頻度2以上の語に対して抽出率が 32.7%，正解率が 82.5%，(ii) 頻度1の語に対して抽出率が 35.2%，正解率が 70.5%であった．(i) と Fung の実験結果を比較すると，提案方法は，抽出率が2分の1弱であるが，正解率は高い．なお，(ii) は，提案方法が頻度1の語にも有効であることを示している．

次に，目的のやや異なる関連研究について述べる．

Tanaka ら[14)]は，2つの言語のコーパス中の共起情報の距離が小さくなるように翻訳確率行列を最適化する方法を提案している．対訳に関する知識の獲得に共起情報を利用するという意味で，本研究と着眼点は似ている．しかし，既知の対訳語ペアのうち，コーパス中に高い確率で生起するペアを抽出する方法であり，本研究のように新しい対訳語ペアを抽出するものではない．

共起語集合の類似度で語の関連度を評価するという考え方は，菊井[15)]によるタームリストの翻訳多義解消方法にもみられる．そこでの目的は，複数のタームに対して同時に訳語を決定することである．各タームの既知の訳語を1つずつ選んだ組のうちで，意味的な関連性の高い組を選択する．このため，同一言語（目標言語）の語の間で関連度を計算している．一方，本研究では，未知の対訳語ペアを抽出することが目的であ

るので，異なる言語（原言語と目標言語）の語の間で関連度を計算する．

7. むすび

語をそれと共起している語の頻度付き集合で特徴付け，その類似度を計算することによって，対訳コーパスから対訳語のペアを抽出する方法を提案した．ここで，異なる言語の語から構成される共起語集合の類似度を計算するため，基本対訳辞書を利用する．

本方法の長所は次のとおりである．第1に，文レベルの対応付けがなされていない対訳コーパスに適用可能である．第2に，小規模な対訳コーパスから対訳語ペアを抽出することができる．第3に，未知語を含む単純語と複合語の任意の組合せの対訳語ペアを抽出することができる．

日英対訳の特許明細書コーパスを用いて，基本対訳辞書に未登録の対訳語ペアを抽出する実験を行ったところ，抽出率が33.8%，正解率が76.7%であった．これにより，基本対訳辞書の増補や専門用語辞書の作成を支援するツールとして実用に供しうるとの結論を得た．本方法は，大規模な対訳コーパスを要求せず，対訳文書を個別に処理していけばよいので，実際的である．

今後の課題として，コーパスからの複合語抽出精度を向上させることがあげられる．また，複合語の構成要素レベルの対応を評価する方法などとの結合も有効と思われる．

参考文献

1) Gale, W.A. and Church, K.W.: Identifying word correspondences in parallel texts, *Proc. 4th DARPA Speech and Natural Language Workshop*, pp.152–157 (1991).
2) Kupiec, J.: An algorithm for finding noun phrase correspondences in bilingual corpora, *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, pp.17–22 (1993).
3) Dagan, I., Church, K.W. and Gale, W.A.: Robust bilingual word alignment for machine aided translation, *Proc. Workshop on Very Large Corpora*, pp.1–8 (1993).
4) 井ノ上直己，野垣内出：対訳テキストを用いた日英対訳辞書の自動生成，電子情報通信学会技術報告 NLC93-39 (1993).
5) Fung, P.: A pattern matching method for finding noun and proper noun translations from noisy parallel corpora, *Proc. 33rd Annual Meeting of the Association for Computational Linguistics*, pp.236–243 (1995).
6) Kitamura, M. and Matsumoto, Y.: Automatic extraction of word sequence correspondences in parallel corpora, *Proc. 4th Workshop on Very Large Corpora*, pp.79–87 (1996).
7) 山本由紀雄，坂本 仁：対訳コーパスを用いた専門用語対訳辞書の作成，情報処理学会研究報告 NL-94-12 (1993).
8) 熊野 明，平川秀樹：対訳文書からの機械翻訳専門用語辞書作成，情報処理学会論文誌，Vol.35, No.11, pp.2283–2290 (1994).
9) 石本浩之，長尾 真：対訳文章を利用した専門用語対訳辞書の自動作成—訳語対応における両立不可能性を考慮した手法について，情報処理学会研究報告 NL-102-11 (1994).
10) Brown, P.F., Lai, J.C. and Mercer, R.L.: Aligning sentences in parallel corpora, *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, pp.169–176 (1991).
11) Gale, W.A. and Church, K.W.: A program for aligning sentences in bilingual corpora, *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, pp.177–184 (1991).
12) Kay, M. and Roscheisen, M.: Text-translation alignment, *Computational Linguistics*, Vol.19, No.1, pp.121–142 (1993).
13) Chen, S.F.: Aligning sentences in bilingual corpora using lexical information, *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, pp.9–16 (1993).
14) Tanaka, K. and Iwasaki, H.: Extraction of lexical translations from non-aligned corpora, *Proc. 16th International Conference on Computational Linguistics*, pp.580–585 (1996).
15) 菊井玄一郎：ターム間の意味的関連性に基づくタームリストの翻訳多義解消，自然言語処理，Vol.7, No.3, pp.79–96 (2000).

(平成12年12月22日受付)
(平成13年 6 月19日採録)

梶 博行（正会員）

1973年京都大学工学部電気工学第二学科卒業．1975年同大学院修士課程修了．同年（株）日立製作所入社，システム開発研究所を経て，現在，同社中央研究所勤務．自然言語処理，機械翻訳，情報検索等の研究開発に従事．電子情報通信学会，人工知能学会，言語処理学会，ACM, Association for Computational Linguistics 各会員．

相薗　敏子（正会員）

1989 年聖心女子大学文学部教育学科心理学専攻卒業．1992 年東京工業大学大学院総合理工学研究科システム科学専攻修士課程修了．同年（株）日立製作所入社，システム開発研究所を経て，現在，同社中央研究所勤務．自然言語処理，情報検索等の研究開発に従事．人工知能学会会員．