対訳文書からの機械翻訳専門用語辞書作成

熊野 明† 平川秀樹†

機械翻訳システムのカスタマイズ手段であるユーザ専門用語辞書作成を自動化する目的で，日英対訳コーパスから機械翻訳用専門用語辞書を作成する方法を開発し評価した．本方法では，(1)日本語文書と英語文書から対応の単位となるユニットを抽出，(2)日本語ユニットと英語ユニットの対応関係を推定，(3)日本語文書から合成名詞と未知語を専門用語として抽出，(4)専門用語を含む日本語ユニットの対応英語ユニットから訳語候補を生成，(5)複数の訳語候補を評価して最も確かなものを選定することで対訳データを作成する．対訳コーパス中の語句の対応関係の推定には，既存の対訳辞書知識から得られる言語情報とテキスト中の頻度から得られる統計情報を統合して利用した．この2種類の情報を利用することにより，構成語間に直接対訳関係のない合成名詞に対する対訳データや未知語に対する対訳データなど，言語情報のみを利用する方法では得られないデータも抽出できた．日英間で文章構成や表現の大きく異なる特許明細書を対象に専門用語辞書の作成実験を行った結果，300文程度の小規模な文書からでも，合成名詞に対する訳語を70%以上の精度で推定できた．未知語の訳語推定は2,000文程度の文書で50%以上の精度が得られた．これまで人手で行っていたユーザ専門用語辞書の作成作業の半分以上を自動化でき，機械翻訳利用の効率を向上した．

Building an MT Technical Term Dictionay from Parallel Texts

Akira Kumano† and Hideki Hirakawa†

A method for generating a machine translation (MT) technical term dictionary from parallel texts has been developed and evaluated. In our approach, (1)parts of documents ("units") are extracted from both Japanese and English texts, (2)each Japanese unit is mapped into English units, (3)Japanese compound nouns and unknown words are extracted from the original text as technical terms, (4)English translation candidates for Japanese terms are extracted from the corresponding English units, and (5)the translation candidates are evaluated to obtain the best one. In order to obtain corresponding words or phrases from parallel texts, this method utilizes both the linguistic information from an existing bilingual dictionary and the statistical information based on the frequency in text. By combining these two types of information, translation pairs which cannot be obtained by a linguistic-based method can be extracted. Over 70% accurate translations for compound nouns are obtained as the first candidate from small (about 300 sentences) Japanese/English parallel texts (patent specifications) containing severe distortions. The accuracy for translations of unknown words is over 50%. Thus, build-up of an MT user technical term dictionary has been semi-automated to improve the efficiency of MT.

1. はじめに

機械翻訳システムの有効利用には，カスタマイズ機能によるユーザ知識の構築が不可欠である．ユーザカスタマイズ機能が対象とする知識にはいろいろな種類がある[1),2)]が，とりわけ語彙知識（専門用語辞書）の整備が重要である．ユーザによる専門用語辞書作成手段としては，あらかじめ準備した対訳用語リストをもとに一括登録する方法と，翻訳原文・訳文の編集時に用語を対話的に登録する方法が一般的であった．しかし，専門用語の対訳リストを事前に準備することは容易ではない．また，対話的な辞書登録では用語ごとに人間による確認が必要であり，翻訳自体の処理に比較してコストのかかる作業である．いずれの場合にも，機械翻訳システムの運用に際してユーザの大きな負担となっていた．したがって，この専門用語辞書作成を効率化できれば，機械翻訳システムにおける作業全体のコストパフォーマンスを向上することができる．

ところで，対訳文書データ（対訳コーパス）は自然言語処理における情報の源として注目されており，各種の知識獲得に関する研究が行われている[3),4)]．この

† (株)東芝 研究開発センター
Research and Development Center, Toshiba Corporation

中では，機械翻訳システムのユーザカスタマイズのために，統計情報や言語情報を利用して辞書データを抽出する研究も盛んである．

統計情報を利用した処理は，対訳コーパスから文の対応関係や語句の対応関係を抽出するために有効であることが示されている[5)~7)]．Kupiec は，文対応のとれた英仏対訳コーパスから名詞句表現の対応関係を得る方法を提唱し，上位 100 語の対応関係が約 90% の精度で得られたと報告している[8)]．また，日英対訳コーパスから n-gram を作成して対訳辞書を半自動生成する研究[9)]では，70% の用語の訳語候補中に正解が含まれている．これらの結果は，言語情報を利用しなくてもある程度の知識獲得が可能であることを示している．このように，大量の対訳文書が利用可能な状況では，統計情報に基づいた処理は有効性が期待できる．

一方，言語情報を利用したものとして，機械翻訳における訳語選択の自動学習が提案されてきた[10)~12)]．さらに，山本らは，既存の機械翻訳辞書を利用して英日対訳コーパスから専門用語対訳辞書を自動作成する研究を行い，人手の作業に比べて用語の網羅性と処理速度に関して有効な手法であると報告している[13)]．この手法では，英語と日本語の各文書中の名詞連続を専門用語候補としてあらかじめ抽出し，既存の機械翻訳辞書を参照することで言語間の対応関係を推定する．ただし抽出する名詞句の種類は利用する機械翻訳辞書の能力で制限されるため，そこから生成可能な訳語候補以外は抽出することができない．また用語の現れる文の対応関係を考慮しないので，対処可能なエラーを残している．

本稿では，既存の対訳文書から合成名詞と未知語を専門用語として抽出し，その対訳辞書を作成する方法について述べる．この方法では，原文訳文間の語句の対応関係を得るために，言語的な情報と統計的な情報を利用する．2種類の情報を利用することにより，言語情報だけでは抽出できない未知語の対訳情報を抽出することができ，また，比較的小規模の文書からでも対訳辞書を作成することができる．

2. 辞書作成のアプローチ

本研究は，比較的少量の対訳データからでも，また，対訳文書間の文対応が単純でないものからでも専門用語対訳辞書の作成ができることを目的としている．現実の応用場面において，ある分野の対訳文書データを大量に入手することは，通常かなり困難であり，仮に多くの対訳文書が存在しても，文の対応関係が単純でない場合が多い．特に，日本語と英語のように語族の異なった言語間においては，この問題はかなり深刻である．

我々は，このような文書の典型例として日本語・英語の対訳特許文書（明細書）を対象として選択した．特許文書は，新しい技術に関する記述が豊富で，新しい技術用語を多く含み，専門用語の抽出対象として価値が高い．しかし，日本語と英語で同じ内容が記述されるものの，文章構成が大きく異なり，また，翻訳に際して表現上かなりの変更が施されていることが多い．このため，対訳特許文書は日英間で対応する辞書データ抽出という観点からは，非常に困難な対象である．

このような対訳文書に対してもロバストな方式を実現するため，我々は言語的情報と統計的情報を統合して利用する方式を採用した．言語的情報には辞書的・構文的・意味的知識が含まれているので，文書の断片からでも語句の対応を判断できるという特徴がある．一方，統計的情報には多くの実例から抽象化した知識が含まれているので，雑音に強いという特徴がある．両者を統合することにより，機械翻訳対訳辞書などの言語情報利用だけでは対訳抽出の困難な未知語に対しても処理が可能になる．

本方式では，以下の手順で機械翻訳用専門用語対訳辞書を作成する．この処理の流れを**図1**に示す．

[1] ユニットの抽出

日本語文書と英語文書から，対応の単位となるユニットを抽出する．

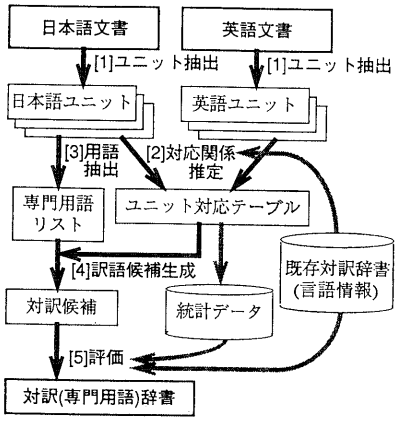


図 1 辞書作成処理の流れ
Fig. 1 Flow of dictionary building.

[2] ユニット間の対応関係の推定
日本語ユニットと英語ユニットの対応関係を推定する.

[3] 専門用語の抽出
日本語文書から専門用語の候補を抽出する.

[4] 用語の英訳語候補の生成
専門用語を含む日本語ユニットに対応する英語ユニットから,訳語候補を生成する.

[5] 用語の英訳語候補の評価
複数の訳語候補を評価して最も確かなものを選定する.

3. ユニット対応関係の推定

対訳文には対応する言語表現が含まれているという仮定は文献 8) の方式の前提であり,統計的手法による情報の源である.専門用語辞書作成においても,この種の情報は積極的に利用すべきである.ただし,我々が対象とする特許文書では日英間の文書構成が大きく異なるため,文の記述順序の対応に基づいた bead model[5] によって表現の対応関係推定を行うことはできない.そこで我々は,文の断片をユニットという概念でとらえ,記述順序によらず言語的知識を主として利用することでユニット間の対応関係を抽出する方法を採用した.

3.1 ユニットの抽出

最初に,日本語・英語両方の文書からユニットの抽出を行う.ユニットとは言語間で対応を付けることのできる単位で,文,節,句などに相当する.専門用語は各ユニットから独立に抽出されるが,専門用語以外の語句は,専門用語に対する文脈情報と呼ぶ.文脈情報は,専門用語の使われる環境,つまり用法を反映しており,専門用語・訳語の対応関係を他の語との関係から判断する材料となる.

ユニットの粒度は,以下の対応付けの精度を大きく左右するものである.仮に,専門用語に相当する名詞句そのものをユニットとすると,ユニットから抽出する用語以外の文脈情報が得られない.その結果,言語間のユニット対応付けを行う際に使用環境や用法を考慮することができなくなり,意味的に正確な対応関係の見付かる可能性が低い.

3.2 ユニットの対応付け

日本語のユニットに対して対応する英語のユニットが存在すると仮定すると,日本語ユニットに含まれる語彙と,英語ユニットに含まれる語彙は近い内容のものが多いと予想できる.辞書に登録すべき専門用語もいずれかの日本語ユニットに含まれており,その訳語は対応する英語ユニットから得ることができる.そのためには,ユニットの対応付けが必要である.

ユニット間の対応付けには,商用の機械翻訳システムの日英対訳辞書のもつ訳語情報を利用した.日本語ユニット中の各内容語に対して機械翻訳対訳辞書を参照し,そこから得られる訳語候補群と英語ユニット中の内容語との対応度をユニット間の対応確信度とした.日本語ユニット JU と英語ユニット EU の対応確信度の計算方法を以下に示す.

(1) 日本語ユニット JU 中の内容語(重複を除く)のリスト $\mathbf{J}$ を作成する.この語数を m とする.
$\mathbf{J}=\{J_1, J_2, \cdots, J_m\}$

(2) 英語ユニット EU 中の内容語(重複を除く)のリスト $\mathbf{E}$ を作成する.この語数を n とする.
$\mathbf{E}=\{E_1, E_2, \cdots, E_n\}$

(3) リスト $\mathbf{J}$ 中の各 J_i に対して日英対訳辞書を参照し,$J_i \Rightarrow E_j$ なる関係にある E_j をすべて選定する.ここで $J_i \Rightarrow E_j$ は,J_i の訳語候補に E_j が含まれていることを示す.

(4) リスト $\mathbf{J}$ 中の J_i のうち,いずれかの E_j に対して $J_i \Rightarrow E_j$ なる関係の発見されたものの語数を x とする.

(5) リスト $\mathbf{E}$ 中の E_j のうち,いずれかの J_i に対して $J_i \Rightarrow E_j$ なる関係の発見されたものの語数を y とする.

(6) JU と EU の対応関係確信度を $P(JU, EU)=(x+y)/(m+n)$ で計算する.

各日本語ユニットに対して,対応関係確信度が大きい英語ユニットから順に対応ユニット候補として推定し,ユニット対応テーブルに格納する.

4. 対訳候補の生成

4.1 用語の抽出

専門用語のみならずその訳語候補を文書中から抽出する精度が低いと,有効な対訳辞書情報が得られない.文献 8) や文献 13) の実験では,最初に原文書・訳文書の両方から名詞連続などの表層的な特徴を利用して専門用語候補を抽出し,対応を推定している.これに対して我々は,原言語である日本語文書だけを構文解析して語彙的・構造的特徴から専門用語を抽出し,その後,英語テキスト中の出現頻度に基づいた統計情報を利用して推定訳語候補を抽出した.日本語で

の解析を優先する理由は，多品詞語の割合が高い英語に比べて，名詞句の認識精度が一般に高いからである．英語での構文解析による用語抽出を前提として処理を行うと，辞書作成処理全体の精度が日本語解析と英語解析の両者の精度に依存するため，高い精度を期待できない．本方法では次節で述べるように，英語の訳語はテキスト中の統計情報を活用し，柔軟に抽出できるようにした．

現在のシステムでは，合成名詞と未知語の2種類を専門用語の候補として扱っている．合成名詞と未知語の例を以下に示す．

A．合成名詞

A1．名詞連続（サ変動詞を含む）

[例]「オープンビット線方式」,「最小加工寸法」,「最密充填」

A2．名詞＋接辞「化」によるサ変動詞

[例]「平坦化する」

A3．動詞連用形＋名詞

[例]「折り返しビット線」

B．未知語

B1．未知語（名詞およびサ変動詞）

[例]「積層する」,「ポリッシング」

B2．サ変動詞の未知語（名詞としては既知）

[例]「センスする」

すべての日本語ユニットからこれらの用語を抽出し，専門用語リストを作成する．日本語の解析には商用の機械翻訳システムの機構を利用した．形態素解析に使用した語彙辞書は，約 70,000 語の見出し語を含んでいる．

4.2 訳語候補の生成

各専門用語の訳語候補は，その用語を含む日本語ユニットに対応する英語ユニットに含まれていると考える．訳語のもつべき制約を最小限に考えると，対応ユニット中の任意の英語単語列が，ある専門用語 JW の訳語である可能性をもつ．この考えにしたがい，訳語候補の生成を次の2段階の処理で行う．

手順 1：対応英語ユニットの選択

手順 2：ユニットからの単語列の抽出

手順 1：

ある専門用語 JW の日本語文書におけるユニット出現頻度が FJU(JW) 個，英語文書の全ユニット数が N 個の場合，FJU(JW)×N 個の組み合わせに対する対応関係確信度を計算し，うち上位 FJU(JW) 個の値をもつ英語ユニットを対応候補ユニット群として選択する．

手順 2：

日本語専門用語 JW の正しい訳語を EW とする．訳語は対応候補ユニット群 $EU_1, EU_2, \cdots, EU_{FJU(JW)}$ に含まれているという仮定から，EW は対応候補ユニット群中で頻繁に現れる単語列であると考える．

対応候補ユニット中で頻繁に現れる単語列を得るために，選択した英語ユニット群から，n-gram データを作成する．ここで n は，用語 JW の構成語数を k としたとき，$1 \leq n \leq 2k$ とする．すなわち，対応候補英語ユニット群中の $2k$ 語以下からなる任意の単語列を訳語候補とした．

抽出対象英文全体から作成した n-gram データのうち，英語ユニット群における出現頻度の高いものから順に訳語候補 EWc_1, EWc_2, … EWc_j とする．出現頻度の低い候補は JW の訳語である可能性が低いという予想から，FJU(JW) に対して一定の割合（6章の実験では 1/4）に満たない頻度の単語列は，ここで訳語候補から除外した．また，今回専門用語として抽出した合成名詞や未知語は名詞であることから，動詞 be を含む単語列，最初か最後が前置詞・接続詞・冠詞である単語列は訳語になり得ないものであり，訳語候補から除外した．

5. 訳語候補の評価

ある訳語候補 EWc_i が専門用語 JW の訳語である確信度 $TL(JW, EWc_i)$ を，二つの確信度の関数として定義する．

$$TL(JW, EWc_i) = F(TLS(JW, EWc_i), TLL(JW, EWc_i))$$

ここで，$TLS(JW, EWc_i)$ は統計情報に基づいた対訳確信度 (translation likelihood based on statistical information)，$TLL(JW, EWc_i)$ は言語情報に基づいた対訳確信度 (translation likelihood based on linguistic information) である．

5.1 統計情報の利用

$TLS(JW, EWc_i)$ はコーパス中の統計情報に基づいた対訳確信度である．言語間の語句の対応関係を統計処理で推定する方法には文献 8) で利用している方法等があるが，文単位の対応関係を前提としており，今回の対象文書に適しているかは検討が必要である．ここでは，単純に出現ユニット頻度を利用し，訳語候補が対応候補ユニットに現れる確率で与える．

$$TLS(JW, EWc_i) = FEU(EWc_i)/FJU(JW)$$

ここで FEU(EWc$_i$) は，EWc$_i$ が現れる対応候補ユニットの数である．

5.2 言語情報の利用

TLL(JW, EWc$_i$) は専門用語 JW と訳語候補 EWc$_i$ の語彙的な対応度に基づく対訳類似スコアであり，機械翻訳辞書から得られる言語情報を利用して計算する．いま専門用語 JW が k 個の構成語からなり，訳語候補 EWc$_i$ が l 個の構成語からなるとする．すなわち，

JW＝wj$_1$, wj$_2$, …wj$_k$

EWc$_i$＝we$_1$, we$_2$, …we$_l$

と表す．

対訳類似スコアを考えるために，次の仮説を利用する．

[仮説]

（a）専門用語 JW と訳語候補 EWc$_i$ とは構成要素単語数が近いほど対応が確からしい．

（b）専門用語 JW 中の各構成語と訳語候補 EWc$_i$ 中の各構成語の間に対訳関係が多いほど対応が確からしい．

この仮説により，言語情報的観点からは，次の仮想的対訳関係（1）を満たす訳語候補 EWc$_i$ が，最も確からしいと考える．

（1） wj$_{j1}$⇒we$_{e1}$, wj$_{j2}$⇒we$_{e2}$, …, wj$_{jk}$⇒we$_{ek}$

ここで wj$_{jx}$⇒we$_{ex}$ は，3.2 ユニットの対応付けで示したのと同様に，wj$_{jx}$ の訳語候補リストに we$_{ex}$ が含まれていることを示す．

JW と EWc$_i$ の構成語間の関係は，一般に（2）に示す 4 種類に分類できる．

（2） ⅰ）wj⇒we

ⅱ）wj→we

ⅲ）wj→ϕ

ⅳ）ϕ→we

(ϕ は語が存在しないことを示す)

ⅰ)の⇒は，対訳関係のある構成語の関係を示し，ⅱ)の→は，対訳関係はないが語数の対応の付く構成語の関係を示し，ⅲ）とⅳ）は，JW と EWc$_i$ の一方に構成語が存在し，他方に語数の対応する構成語が存在しないことを示している．

いま JW と EWc$_i$ の構成語の順序を無視し，次のように構成語の集合とみなす．

JW＝{wj$_1$, wj$_2$, …wj$_k$},

EWc$_i$＝{we$_1$, we$_2$, …we$_l$}

仮説により，すべての we がいずれかの wj と過不足なく対訳関係のある訳語候補（（1）で示した）を，最も確かな仮想訳語と仮定する．対訳類似スコアの計算は，以下に示す方法により仮想訳語との比較で行う．

EWc$_i$ の構成語 we のうちいずれかの wj と語数の対応がとれるものに得点 P（定数）を与える．このうち，対訳関係のあるものには得点 αP ($\alpha>0$) を加点する．したがって，（2）のⅰ）を満たす構成語 we には $P+\alpha P=(1+\alpha)P$，ⅱ）を満たす構成語 we には P が与えられる．つまり，構成語間に明示的な対訳関係が存在しない場合でも対訳確信度を評価する．なお，α の値は予備実験の結果から $\alpha=2$ と決めた．

TLL(JW, EWc$_i$) は，EWc$_i$ の得点と仮想訳語の得点（上の定義により $k\times(1+\alpha)P$）との比で与える．下の例は，TLL(JW, EWc$_i$) の計算を示したものである．訳語候補の構成語のうち，太字で示したもの (open, bit, line) は JW の構成語と対訳関係のあるもの（（2）のⅰ）を満たす構成語），それ以外 (configuration) は語数の対応がとれるもの（（2）のⅱ）を満たす構成語）である．

[例] JW＝オープン／ビット／線／方式 ($k=4$)

open bit line: $(3\times3P)/12P=0.75$

bit line configuration: $(2\times3P+P)/12P=0.58$

open bit line configuration: $(3\times3P+P)/12P=0.83$

5.3 統計情報と言語情報の統合

専門用語 JW に対する訳語候補 EWc$_i$ の確信度 TL(JW, EWc$_i$) を，次のように統計情報による確信度と言語情報による確信度の加重平均で統合する．

$$\begin{aligned}&\mathrm{TL}(\mathrm{JW},\mathrm{EWc}_i)\\&=F(\mathrm{TLS}(\mathrm{JW},\mathrm{EWc}_i),\ \mathrm{TLL}(\mathrm{JW},\mathrm{EWc}_i))\\&=\{p\mathrm{TLS}(\mathrm{JW},\mathrm{EWc}_i)+q\mathrm{TLL}(\mathrm{JW},\mathrm{EWc}_j)\}/\\&\quad(p+q)\end{aligned}$$

$p=q$ とした予備実験の結果，JW のユニット出現頻度 FJU(JW) が小さい場合に TLS(JW, EWc$_i$) の値が極端に低くなるため，正しい EWc$_i$ に対して TLL(JW, EWc$_i$) の値が高くなる場合でも，全体の対訳確信度 TL(JW, EWc$_i$) の値を低くすることがわかった．このため，$\beta=q/p$ を FJU(JW) の関数として与えた．すなわち，FJU(JW) が大きくなるにつれて β が小さくなるような関数を仮定して使用した．

6. 評　価

6.1 実　験

同一分野に関する日本語の特許明細書7件とそれを技術者が翻訳した英訳文を使って評価実験を行った．この実験では，文をユニットとして扱っている．日本語文書のサイズは，全体で2,148文，99,286文字（平均307文，14,184文字）である．

機械翻訳辞書作成経験者が別途選定した正解に対して，本方式で推定した訳語が一致する例を**図2**に示す．合成名詞の訳語だけでなく，未知語の訳語も一部取り出すことができた．これは言語的情報だけによる方法[13]では実現できないものであり，専門用語辞書作成には大きな効果がある．

次の対訳例は，本方式の言語情報の効果を示すものである．

オープンビット線方式＝
open bit line configuration

日本語用語の構成語と英語用語の構成語の間の対訳関係を条件とした方式[13]では，参照した対訳辞書に「方式＝configuration」という対訳知識がない限りこの対訳を抽出することができない．本方式では，言語情報を柔軟に利用することによって抽出ができた．

次の対訳例は，統計情報の効果を示すものである．

カラムアドレスストローブ
＝*column address strobe*

この例では，「ストローブ」が未知語で「スト」が辞書見出しであったために「カラム，アドレス，スト，ローブ」と単語分割を誤ったにもかかわらず，正しい訳語を推定している．言語情報による対応関係の低い確信度を，統計情報による確信度が補った例である．

表1は，第1推定訳語が正解に一致する率と上位3位推定訳語に正解が含まれる率を示したものである．統計情報が訳語の推定精度に及ぼす影響を調べるために，処理対象文書量に変化を与えて実験を行った．上段は1文書ごとの処理結果の平均，下段は7文書を統合して処理した結果を示す．いずれも7文書での実験の結果が平均を上回っているが，特に未知語の正答率の上昇が大きい．本方式での未知語に対する訳語推定処理は，統計情報を利用している部分が大きいため，文書量を増やすことにより統計情報の効果が高まることがわかる．

6.2 考　察

正しい訳語が推定できなかった場合を分析した．主な原因は次の2点である．

1. ユニットの対応の誤り
2. 未知合成名詞の単語切りの誤り

ユニットの対応誤りは，現在のシステムでは実際に1文対1文の対応関係がない場合に起きる．実験に用いた対訳文書の1例を調べたところ，98の日本語文のうち12文は日英間に1対1の文対応がなかった．1対1対応のない場合の多くは，日本語1文に対して英語2文が対応しているものであった．日本語のあるユニット JU_i に対して英語の2ユニット EU_j+EU_{j+} が対応している場合，EU_j が EU_{j+1} より長いと JU_i^1 の対応候補ユニットとして EU_j が選定される確率が高いが，JU_i 中の専門用語の訳語が EU_{j+1} に含まれる場合もあり，この状況では対応候補ユニットから正しい訳語を推定できない．

ユニットの対応関係を正確にする方法の一つとし

合成名詞	
最小加工寸法	*minimum featuring size*
素子分離領域	*element separation region*
オープンビット線方式	*open bit line configuration*
カラムアドレスストローブ	*column address strobe*
セルアレイ	*cell array*
未知語	
ポリッシング	*polishing*
コレクタ	*collector*
積層する	*to form*

図2 正しく推定された訳語
Fig. 2 Correctly obtained translations.

表1 推定訳語の正答率
Table 1 Accuracy of obtained translations.

処理単位		合成名詞			未知語		
		総頻度	第1訳語	上位3訳語	総頻度	第1訳語	上位3訳語
1文書	*306.9*文	*460.6*	71.7% (*330.3*)	82.5% (*380.1*)	*55.6*	30.1% (*16.7*)	52.4% (*29.1*)
7文書	2,148文	3,224	72.9% (2,349)	83.3% (2,680)	389	54.0% (210)	65.0% (253)

（注） 斜体の頻度は，7文書の平均値

て，現在ユニットとして扱っている単位を文ではなく節や動詞句など，より小さな単位にすることが考えられる．ユニットを小さくすることで，JU_i は EU_j と EU_{j+1} に対応する部分に分割される可能性が高く，言語情報による対応が確かになる．また，文脈情報を共有する表現の頻度が大きくなることで統計情報の効果が高まり，統合的に訳語推定の精度を向上することが期待できる．

未知語の合成名詞の誤りは，特にカタカナ語で問題となる．カタカナ語が連続する場合，その分割を誤ると，言語情報を利用して正しい訳語を推定することができない．このためには，カタカナ語を多く登録した辞書を利用することにより，形態素解析精度を向上させる必要がある．

今回実験した方法では，言語情報として対訳辞書に基づいた対訳関係だけを利用している．このため，対応ユニットの選択後は，未知語に関して言語情報をまったく利用できない．ユニットを構文解析してその構造を参照すれば，日英間の専門用語の対応関係がかなり確かになると予想される．この方法は，合成名詞に対しても効果があると考える．また，日本語用語と英語訳語候補の頻度比以外の統計情報を用いた手法により，訳語推定の精度を向上できる．

7. 結　　論

既存の機械翻訳用対訳辞書から得られる言語情報と文書中の頻度情報をもとにした統計情報を利用することで，対訳文書から機械翻訳用専門用語辞書を作成できた．利用した文書は特許明細書7件であり，サイズとしても比較的入手しやすいものである．

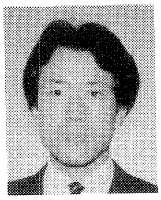
日本語文中の合成名詞と未知語を専門用語としてとらえ，その両者に対して訳語を推定することができた．合成名詞では約 70% の精度が得られ，第3候補まで提示してユーザが選択する方式を仮定すると，80%以上の割合でデータが利用できる．未知語に対する訳語は 2,000 文程度の対訳文書を利用した場合に 54%推定できた．これは合成名詞に比べて低いが，対象文書量をさらに増やすことで改良できる見通しを得た．結果的に，機械翻訳システムのカスタマイズに不可欠ながらこれまで人手で行っていたユーザ専門用語辞書の作成作業の多くの割合を自動化でき，機械翻訳利用の効率を向上することができた．

今後は，機械翻訳辞書としての評価のために，人手で作成した辞書との比較を行う．また，今回は合成名詞・未知語だけを専門用語としたが，普通名詞・動詞などの訳語選択情報を抽出することでカスタマイズ効果は高まると考えられる．また，本方法で作成した辞書が，実際の機械翻訳作業においてどの程度有効であるか，実際の翻訳作業に利用して調べる予定である．

参 考 文 献

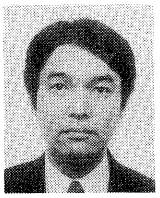
1) 熊野　明，吉村裕美子，平川秀樹，天野真家：機械翻訳文法のカスタマイズ，情報処理学会研究報告，NL 84-11 (1991).
2) 熊野　明，木下　聡，平川秀樹：機械翻訳のユーザ規則によるカスタマイズ，人工知能学会研究会資料，SIG-SLUD-9301-6 (1993).
3) Dagan, I., Itai, A. and Schwall, U.: Two Languages Are More Informative Than One, *Proc. of the 29th Annual Meeting of the ACL*, pp. 130-137 (1991).
4) Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Parallel Texts, *Proc. of the 31st Annual Meeting of the ACL*, pp. 23-30 (1993).
5) Brown, P.F., Lai, J.C. and Mercer, R.L.: Aligning Sentences in Parallel Corpora, *Proc. of the 29th Annual Meeting of the ACL*, pp. 169-176 (1991).
6) Gale, W.A. and Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics*, Vol. 19, No. 1, pp. 75-90 (1993).
7) Chen, S.F.: Aligning Sentences in Bilingual Corpora Using Lexical Information, *Proc. of the 31st Annual Meeting of the ACL*, pp. 9-16 (1993).
8) Kupiec, J.: An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, *Proc. of the 31th Annual Meeting of the ACL*, pp. 17-22 (1993).
9) 野美山浩：コーパスからの対訳辞書の半自動生成，第 47 回情報処理学会全国大会論文集，6P-8 (1993).
10) 野上宏康，熊野　明，田中克己，天野真家：既存目的言語文書からの訳語の自動学習方式，第 42 回情報処理学会全国大会論文集，2C-6 (1991).
11) 野美山浩：目的言語の知識を用いた訳語選択とその学習性，情報処理学会研究報告，NL 86-8 (1991).
12) 加藤直人：目標言語のフルテキスト検索による機械翻訳の訳語選択，電子情報通信学会技術報告，NLC 93-32 (1993).
13) 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).

(平成6年4月21日受付)
(平成6年7月14日採録)



熊野　明（正会員）

1959 年生．1982 年東京工業大学工学部情報工学科卒業．同年東京芝浦電気(株)(現，(株)東芝)入社．以来，機械翻訳，電子化辞書などの自然言語処理システムの研究開発に従事．1986-1989 年(株)日本電子化辞書研究所研究員．現在(株)東芝研究開発センター情報・通信システム研究所研究主務．人工知能学会会員．



平川　秀樹（正会員）

1956 年生．1980 年京都大学大学院工学研究科電気工学専攻修士課程修了．同年東京芝浦電気(株)（現，(株)東芝）入社．以来，機械翻訳，談話解析などの自然言語処理システムの研究開発に従事．1982-1985 年(財)新世代コンピュータ技術開発機構研究員．1993 年より(株)日本電子化辞書研究所第五研究室室長．現在(株)東芝研究開発センター情報・通信システム研究所主任研究員．人工知能学会，言語処理学会，ACL 各会員．