自 然 言 語 処 理 100−12
（1994. 3. 18）

言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成

熊野 明　平川秀樹
{kmn,hirakawa}@isl.rdc.toshiba.co.jp
(株)東芝 研究開発センター
〒210 川崎市幸区小向東芝町1

対訳文書から機械翻訳用専門用語辞書を作成する方法について述べる。本方法では、対訳コーパス中の語句の対応関係を抽出するために、言語情報と統計情報を統合して利用する。この2種類の情報を利用することにより、従来の言語情報のみによる方法では得られない、未知語の対訳関係なども抽出可能になる。言語間で文章構成の大きく異なる特許文書で実験した結果、300文程度の小規模な文書からでも、合成語に対する訳語を70%以上の精度で推定できた。未知語の訳語推定は小規模の文書では精度が低いが、文書量を増やすことで精度が向上する見通しを得た。

Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information

Akira Kumano　Hideki Hirakawa
Research and Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210, JAPAN

A method for generating a machine translation(MT) dictionary from parallel texts is described. This method utilizes both linguistic information and statistical information to obtain corresponding words or phrases in parallel texts. By combining these two types of information, translation pairs which cannot be obtained by a linguistic-based method can be extracted, and over 70% accurate translations are obtained as the first candidate from small Japanese/English parallel texts containing severe distortions.

1. はじめに

機械翻訳システムを有効に利用するためには、ユーザによるカスタマイズが必要である。ユーザの手でできるカスタマイズ手段には色々なレベルのものがある[熊野91][熊野93]が、とりわけ辞書の整備が重要である。従来、ユーザの辞書作成としては、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、翻訳原文の前編集時または訳文の後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、新規の文書を翻訳する際に、必要な対訳用語リストを前もって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、翻訳作業の効率を上げるための機械翻訳システムの中で、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、機械翻訳システム運用に際してユーザの大きな負担となっていた。そこで、大量の文書を翻訳する際、この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。これは特に、機械翻訳システムの運用開始時に大きな効果がある。

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[Dangan91][Matsumoto93]。この中では、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[Brown91][Gale93][Chen93]。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語の対応精度は約90%であったと報告している[Kupiec93]。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[野美山93]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[野上91][野美山91][加藤93]。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で、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[山本93]。この手法では、英語と日本語文書中の名詞連続をあらかじめ抽出し、既存の機械翻訳辞書を参照することで対応関係を推定する。つまり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ので、機械翻訳辞書から生成可能な訳語以外は抽出することができない。また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語と未知語の専門用語辞書を作成する新しい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、言語情報だけでは取り出すことのできない語句の対応関係を抽出することができ、比較的小規模の文書からでも対訳辞書を作成できる。

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、文書中の文対応が単純でないものからでも利用できる専門用語辞書の作成を目的としている。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においては、その問題はかなり深刻である。

我々は、このような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、技術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のアイデアが記述されるものの、文章の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。特に、文の記述順序を利用した抽出方法は現実的でない。

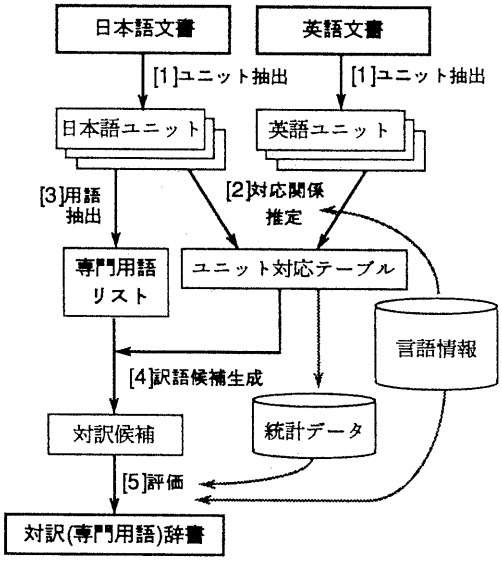

図1 辞書作成処理の流れ

このような対訳文書に対してもロバストな方式を開発するため、我々は言語的な情報と統計的な情報を統合して利用する方式を検討した。言語的な情報には辞書的・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り、統計的な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の処理も可能になる。

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。

[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
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を含んでいるというもっともな仮定は[Kupiec93]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。

しかし、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、表現の対応関係推定には、文の記述順序の対応に基づいたbead model[Brown91]を利用できない。そこで我々は、文の断片をユニットという概念でとらえ、言語的知識を主に利用してユニット間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。ユニットとは日英間で対応を付けることのできる単位で、文、節、あるいは句などに相当し、専門用語はこのユニットごとに抽出する。ユニット中で抽出対象の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いる。

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文脈情報を参照することができなくなり、用語の用法を考慮した正確な対応関係の見付かる可能性が低い。

現在は、最初の評価実験として、文をユニットとして扱っている。

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と英語ユニット中の内容語との対応関係を文間の対応関係とした。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**J**を作成する。
$$\mathbf{J} = \{ J_1, J_2, ..., J_m \}$$
この語数を mとする。

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

(3) リスト**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**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**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
$$P(JU, EU) = \frac{x+y}{m+n}$$
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

専門用語やその訳語の候補を文中から抽出する作業の精度が低いと、最終的に有効な対訳辞書情報が得られなくなってしまう。[Kupiec93]や[山本93]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ている。これに対して我々は、最初に原言語である日本語文書だけを構文解析して専門用語を抽出し、その後、統計情報に基づいて英語の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。本方法では解析処理は日本語文に限定し、英語の訳語は後述するように、統計情報を活用して柔軟に抽出できるようにした。

現在のシステムでは、合成語と未知語の2種類を専門用語の候補として抽出している。ここでいう合成語と未知語には、次のものを含んでいる。

A. 合成語

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」

全ての日本語ユニットを構文解析してこれらの用語を抽出し、専門用語リストを作成する。形態素解析・構文解析には商用の機械翻訳システムの機構を利用した。解析に使用した語彙辞書は、約70,000語の見出しを有している。

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の対応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専門用語 JWの訳語である可能性がある。そこで、訳語候補の生成は次の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 $EU_1, EU_2, ..., EU_{FJU(JW)}$ 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から、n-gramデータを抽出する。ここで$\mathbf{n}$は、用語JWの構成語数をkとしたとき、$1 \leq \mathbf{n} \leq 2k$ とする。この結果、対応候補英語ユニット中の$2k$語以下からなる任意の単語列を訳語候補とすることができる。

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(本実験では 1/4)に満たない頻度の単語列は、この処理以降で使用する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語や未知語は名詞であることから、動詞 be を含む単語列、前置詞・接続詞・冠詞で始まるか終わる単語列は予め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、2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)は統計情報に基いた対訳確信度、TLL(JW, EWc_i)は言語情報に基いた対訳確信度である。

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[Kupiec93]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、単純に出現文頻度を利用し、訳語候補が対応候補文に現れる確率で与える。

$$TLS(JW, EWc_i) = \frac{FEU(EWc_i)}{FJU(JW)}$$

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) $wj_{j1} \Rightarrow we_{e1}, wj_{j2} \Rightarrow we_{e2}, ..., wj_{jk} \Rightarrow we_{ek}$

ここで $wj_{ji} \Rightarrow we_{ei}$ は、3.2 ユニットの対応付け で示したのと同様に、wj_{ji}の訳語候補リストにwe_{ei}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) $wj \Rightarrow we$
ii) $wj \rightarrow we$
iii) $wj \rightarrow \phi$
iv) $\phi \rightarrow we$
(ϕ は語が存在しないことを示す)

i) は、対訳関係のある構成語の関係を示し、ii) 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。

$JW = \{wj_1, wj_2, \ldots wj_k\}$,
$EWc_i = \{we_1, we_2, \ldots we_l\}$

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を与える。このうち、対訳関係のあるものには得点$\alpha P(\alpha>0)$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を与える。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、太字で示したものは JWの構成語と対訳関係のあるもの、それ以外は語数の対応がとれるものである。

[例]「オープン/ビット/線/方式」(k=4)
open bit line:
(3×3P)/12P = 0.75
bit line configuration:
(2×3P+P)/12P = 0.58
open bit line configuration:
(3×3P+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 $TLL(JW, EWc_i)$ と $TLL(JW, EWc_i)$ の加重平均で再定義する。

$$TL(JW, EWc_i) = \frac{p\,TLS(JW, EWc_i) + q\,TLL(JW, EWc_i)}{p+q}$$

pとqの比を一定に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$TLS(JW, EWc_i)$ の値がかなり低くなり、正しいEWc_iに対して$TLL(JW, EWc_i)$ の値が高い場合でも、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、$\beta=\frac{q}{p}$ 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実験を行った。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。また、表1には、第1推定訳語と上位3位推定訳語に正解が含まれる率を示している。統計情報が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、

合成語

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

表1 推定訳語の正答率

処理単位		合成語			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

下段は7文書を統合して処理した結果である。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が上昇している。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る効果が現れていることがわかる。このことは、言語情報を利用できる合成語に関しても同様である。特に出現頻度の低い用語に対しては統計情報がほとんど利用できないので、それを補うためにも文書量の増加による頻度の上昇は効果がある。

次の例は、統計情報の他の効果を示す例である。

「カラムアドレスストローブ」
= *colunm address strobe*
「セルアレイ」= *cell array*

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。「セルアレイ」の例でも、「レイ」のみが辞書見出しであったために、「セルア, レイ」と解析したが、正しい訳語が得られた。

6.2 考察

次に、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. 未知合成語の単語切りの誤り

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つを調べたところ、98の日本語文のうち12文は1対1の対応がないことがわかった。1対1の対応が保証されない場合、本方法による訳語推定の精度は、自ずと低下する。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1ユニットJU_iに対して英語の2ユニットEU_j+EU_{j+1}が対応している場合、EU_jがEU_{j+1}より長いとJU_i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。

未知語の合成語の誤りは、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、誤った言語情報を利用して正しい訳語との対応がとれない。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、未知語に関しては言語情報をまったく利用していないことになる。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、合成語に対しても効果があり、同様に推定精度を向上することができるであろう。

また、現在は統計情報としては日本語用語と英語訳語候補の頻度比のみを利用している。今後は、統計情報の利用で仮定したパラメータの調整を行ったり、他の統計情報を用いた手法を検討することでも、訳語推定の精度を向上させる予定である。

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、対訳文書から専門用語辞書を作成できる見通しを得た。利用した文書は特許明細書の7文書であり、サイズとしても大規模なものではなく、比較的入手しやすいものである。

原文中の語の特徴から合成語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。合成語では約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で考えると、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推定の精度は2,000文程度の対訳文書では合成語に比べて低いが、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辞書作成を一部自動化することにより、実際の運用における辞書作成を加速できる見通しを得た。

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が必要である。今回は合成語・未知語だけを辞書登録対象としたが、一般の名詞・動詞などの訳語選択情報を抽出すればその効果は高まる。また、本方法による辞書作成が、実際の機械翻訳作業においてどの程度の効果があるか、実際の翻訳作業に利用して調べる予定である。

参考文献

[Brown91] Peter F. Brown, Jennifer C. Lai and Robert L. Mercer: Aligning sentences in parallel corpora, *In Proc. of the 29th Annual Meeting of the ACL*, pp. 169-176 (1991)

[Chen93] Stanley F. Chen: Aligning Sentences in Bilingual Corpora Using Lexical Information, In *Proc. of the 31st Annual Meeting of the ACL*, pp. 9-16 (1993)

[Dagan91] Ido Dagan, Alon Itai and Ulrike Schwall: Two languages are more informative than one, In *Proc. of the 29th Annual Meeting of the ACL*, pp. 130-137 (1991)

[Gale93] William A. Gale and Kenneth W. Church: A program for aligning sentences in bilingual corpora, *Computational Linguistics*, Vol. 19, No. 1, pp. 75-90 (1993)

[加藤93] 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, NLC93-32 (1993)

[熊野91] 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, NL84-11 (1991)

[熊野93] 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)

[Kupiec93] Julian Kupiec: An algorithm for finding noun phrase correspondences in bilingual corpora, In *Proc. of the 31st Annual Meeting of the ACL*, pp. 17-22 (1993)

[Matsumoto93] Yuji Matsumoto, Hiroyuki Ishimoto and Takehito Utsuro: Structural Matching of Parallel Texts, In *Proc. of the 31st Annual Meeting of the ACL*, pp. 23-30 (1993)

[野上91] 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, 情報処理学会第42回全国大会, 2C-6 (1991)

[野美山91] 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, NL86-8 (1991)

[野美山93] 野美山浩: コーパスからの対訳辞書の半自動生成, 情報処理学会第47回全国大会, 6P-8 (1993)

[山本93] 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, NL94-12 (1993)