Vol.2012-NL-208 No.2
2012/9/2

パテントファミリーを用いた専門用語訳語獲得における対訳文対非抽出部分の利用

豊田 樹生[1]　高橋 佑介[1]　牧田 健作[1]　宇津呂 武仁[2]　山本 幹雄[2]

概要：専門用語の対訳辞書は特許文書翻訳の過程において必要不可欠なものである．本論文では，日米パテントファミリーを情報源として，専門用語対訳辞書を生成する手法を提案する．従来より，日米パテントファミリーの対応特許文書中において，「背景」および「実施例」の部分の日英対訳文対の対応付けを行い，これを情報源として専門用語の対訳辞書を生成する手法が提案されている．しかし，この方式では，対訳文対が抽出される部分は，「背景」及び「実施例」全体の約 30%であり，約 70%は利用されていなかった．そこで，本論文では，「背景」及び「実施例」の残りの 70%の部分から専門用語の対訳辞書を生成する方式を提案する．NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に登録されていない日本語複合名詞を対象として，既存の対訳辞書を用いた要素合成法を適用したところ，約 13%については，対応特許文書の英語側に出現する英訳語が生成可能であり，その精度は 90%以上であった．提案方式を日英対訳特許文書 1,000 文書対に適用したところ，一特許文書対あたり平均二組の対訳専門用語対を収集することができた．

キーワード：対訳専門用語，パテントファミリー，統計的機械翻訳

Utilizing Portion of Patent Families with No Parallel Sentences Extracted in Estimating Translation of Technical Terms

Itsuki Toyota[1]　Yusuke Takahashi[1]　Kensaku Makita[1]　Takehito Utsuro[2]　Mikio Yamamoto[2]

Abstract: A bilingual lexicon for technical terms is necessary in the process of translating patent documents. This paper studies a method of generating bilingual lexicon for technical terms from parallel patent documents. In the previous methods of generating bilingual lexicon from parallel patent documents, the portion from which parallel patent sentences are extracted is composed of the parts of “Background” and “Embodiment”. However, this portion is about 30% out of the whole “Background” and “Embodiment” parts and about 70% are not used. Considering this situation, this paper proposes to generate bilingual lexicon for technical terms from the remaining 70% out of the whole “Background” and “Embodiment” parts. The proposed method employs the compositional translation estimation technique which uses an existing bilingual lexicon. We show that, for 13% of the Japanese compound nouns that are not included in the phrase translation table trained with parallel patent sentences nor in the existing bilingual lexicon, translation candidates can be generated through the compositional translation estimation technique, and can be found in the English part of the patent family. On the average, we generate about two pairs of bilingual technical terms per patent family and we achieve over 90% accuracy.

Keywords: bilingual lexicon for technical terms, patent family, statistical machine translation

1. はじめに

特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて不可欠である．特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源であり，これまでに，対訳特許文書を情報源として，専門用語対訳対を自動獲得する手法の研究が行われてきた．文献 [10] では，NTCIR-7 特許翻訳タスク [3] において配布された日英 180 万件の対訳特許文を用いて，対訳特許文からの専門用語対訳対獲得を行った．この研究では，句に基づく統計的機械翻訳モデル [7] を用いることにより，対訳特許文から学習されたフレーズテーブル，要素合成法，Support Vector Machines (SVMs) [15] を用いることによって，専門用語対訳対獲得を行った．また，文献 [8] においては，文献 [10] の専門用語訳語推定タスクの後段のタスクとして，同義対訳専門用語の同定と収集を行っている．

ここで，上述の日英 180 万件の対訳特許文は，文献 [14] の手法により，日米パテントファミリーの対応特許文書中において，「背景」および「実施例」の部分の日英対訳文対を対応付けたものであるが，実際に良質な対訳文対が抽出できた部分の割合は約 30%にとどまっている．そこで，本論文においては，「背景」および「実施例」のうちの残りの 70%の部分を言語資源として，専門用語の訳語推定を行った結果について報告する．具体的には，NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に登録されていない日本語複合名詞を対象として，既存の対訳辞書を用いた要素合成法 [12] を適用したところ，約 13%については，対応特許文書の英語側に出現する英訳語が生成可能であり，その精度は 90%以上であった．提案方式を日英対訳特許文書 1,000 文書対に適用したところ，一特許文書対あたり平均二組の対訳専門用語対を収集することができた．

2. 日英対訳特許文

本論文では，フレーズテーブルの訓練用データとして，NTCIR-7 の特許翻訳タスク [3] で配布された約 180 万対の日英文対応データを使用した．なお，この文対応データは以下に示す手順で作成されたものである．

(1) 1993-2000 年発行の日本公開特許広報全文と米国特許全文を得る．
(2) 米国特許の中から日本に出願済みのものを優先権番号より得て，日英対訳特許文書を取得する．
(3) 日英対訳特許において日英間で比較的直訳されている関係となっている度合が大きい「背景」及び「実施例」の部分を抽出する．
(4) 抽出した部分に対して，[14] によって日英間で文対応をつける．

[1] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305–8573, Japan
[2] 筑波大学 システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, 305–8573, Japan

3. 句に基づく統計的機械翻訳モデルのフレーズテーブル

本論文で用いるフレーズテーブルでは，日英の句の組，及び，日英の句が対応する確率を推定し記述する．このとき，句に基づく統計的機械翻訳モデルのツールキットである Moses [6] を前節で述べた文対応データに対して適用する．Moses によってフレーズテーブルを作成する過程を以下に示す．

(1) 単語の数値化，単語のクラスタリング，共起単語表の作成などの処理を文対応データに対する前処理として行う．
(2) 文対応データを利用し，最尤な単語対応を英日・日英の両方向において得る．
(3) 英日・日英両方向における単語対応を利用し，ヒューリスティックスを用いることにより，対称な単語対応を得る．
(4) 対称な単語対応を用いて，可能な全ての日英の句の組を作成する．そして，各組に対して，「文単位の句対応制約」の条件に対する違反の有無をチェックする (違反しない句の組を有効な対応とみなす)．
(5) 文対応データにおける日英の句の対応の数を集計する．このとき，各句の対応に翻訳確率を付与する．

手順 (4) について，以下に「文単位の句対応制約」の条件を示す．

> 日本語文の形態素列中の形態素を文頭から順に $V_1, V_2, \ldots, V_n$，英文の単語列中の単語を文頭から順に $W_1, W_2, \ldots, W_m$ として，日本語句を $P_J (= V_p \cdots V_{p'})$ とし，英語句を $P_E (= W_q \cdots W_{q'})$ とする．ここで，日英句の組 $\langle P_J, P_E \rangle$ が含まれるある一つの対訳文対 $\langle T_J, T_E \rangle$ 中において得られているあらゆる単語対応 $\langle V_i, W_j \rangle$ について，「$p \leq i \leq p' \Leftrightarrow q \leq j \leq q'$ 」が成り立つ場合に，P_J と P_E は対訳文対 $\langle T_J, T_E \rangle$ において「文単位の句対応制約」に違反しない，と定義する．

4. 要素合成法による訳語推定

4.1 訳語推定の概要

要素合成法による訳語推定の例として，日本語専門用語 “並列態様” の英語訳を推定する様子を図 1 に示す．まず，専門用語 “並列態様” を構成要素 “並列” と “態様” に分割し，それぞれの構成要素を既存の対訳辞書を利用して目的言語に翻訳する．このとき，それぞれの構成要素の訳語対

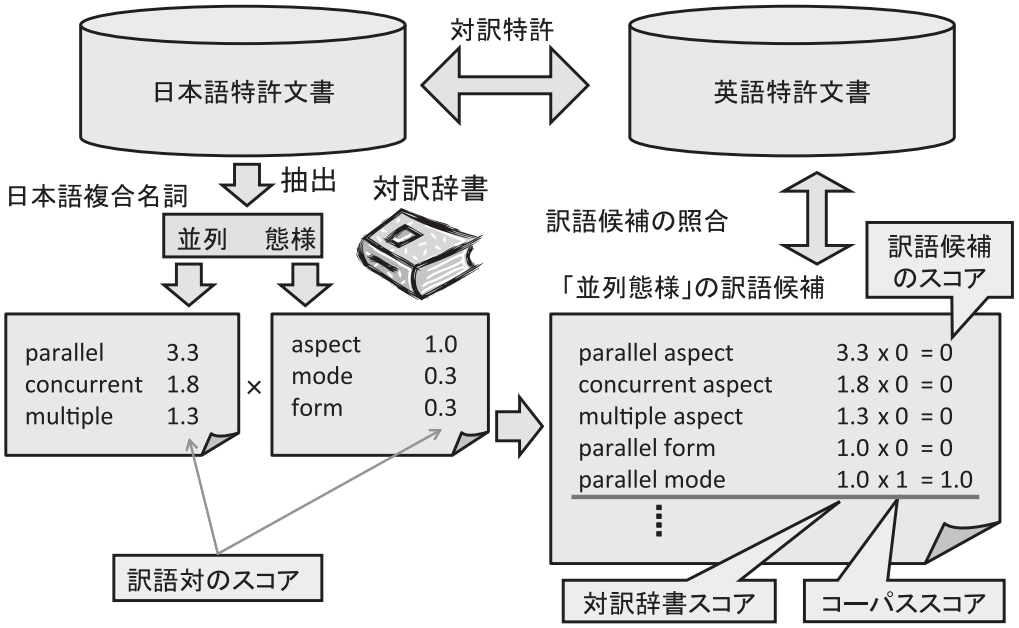

図 1 日本語の専門用語「並列態様」の要素合成法による訳語推定

表 1 英辞郎における見出し語数及び訳語対数

辞書	見出し語数		訳語対数
	英語	日本語	
英辞郎	1,631,099	1,847,945	2,244,117
前方一致部分対応対訳辞書	47,554	41,810	129,420
後方一致部分対応対訳辞書	24,696	23,025	82,087

にはスコアが付与されている．次に，前置詞句の構成を考慮した語順の規則にしたがって，それらの構成要素の訳語を結合し，訳語候補を生成する．ここで，訳語候補のスコアは，対訳辞書スコアとコーパススコアの積で計算される．対訳辞書スコアは訳語対のスコアの積で計算され，コーパススコアはその訳語が目的言語コーパスにおいて生起しているか否かによって計算される．最後に，スコア 1 位の訳語候補が選択される．つまり，この例の場合，“parallel mode” が訳語として獲得されることになる．

本論文では，既存の対訳辞書として「英辞郎」[*1] [*2]に加えて，英辞郎の訳語対から作成した部分対応対訳辞書 [12] を用いる．まず，既存の対訳辞書から，日本語及び英語の用語がそれぞれ 2 つの構成要素 (具体的には，日本語の場合は JUMAN[*3]による形態素解析によって得られる形態素列，英語の場合は単語列) からなる訳語対を抽出し，これを別の対訳辞書 P_2 とする．次に，P_2 中の訳語対の第一構成要素から前方一致部分対応対訳辞書 B_P を作成し，第二構成要素から後方一致部分対応対訳辞書 B_S を作成する．

本論文においては，英辞郎については Ver.131 を使用し，前方一致部分対応対訳辞書及び後方一致部分対応対訳辞書については，Ver.79 及び Ver.131 を統合したものを用いた．

[*1] http://www.eijiro.jp/
[*2] 本論文では，英辞郎 Ver.79 及び Ver.131 を用いる．
[*3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

英辞郎および部分対応対訳辞書の見出し語数および訳語対数を表 1 に示す．

4.2 訳語候補のスコア

本節では，要素合成法による訳語推定における訳語候補のスコアを定義する．

まず，訳語推定すべき専門用語を y_S，この y_S に対して生成された訳語候補を y_T とする．このとき，y_S を以下に示すように構成要素 s_i に分解できると仮定する．

$$y_S = s_1, s_2, \cdots, s_n$$

次に，s_i の訳語を t_i とすると，y_T も同様に以下に示すように分解される．

$$y_T = t_1, t_2, \cdots, t_n$$

このとき，訳語対 $\langle y_S, y_T \rangle$ は以下のように表される．

$$\langle y_S, y_T \rangle = \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \cdots, \langle s_n, t_n \rangle$$

次に，訳語候補 y_T にスコアを与えることを考える．

具体的には，まず，対訳辞書を用いて y_S と y_T の対応の適切さを推定し，その適切さに応じたスコアを与える (これを対訳辞書スコアと呼ぶ．)．ただし，y_T 全体の対訳辞書スコアは，訳語対 $\langle s_i, t_i \rangle$ のスコア $q(\langle s_i, t_i \rangle)$ の積で構成される．次に，目的言語コーパス中で訳語候補 y_T が生起

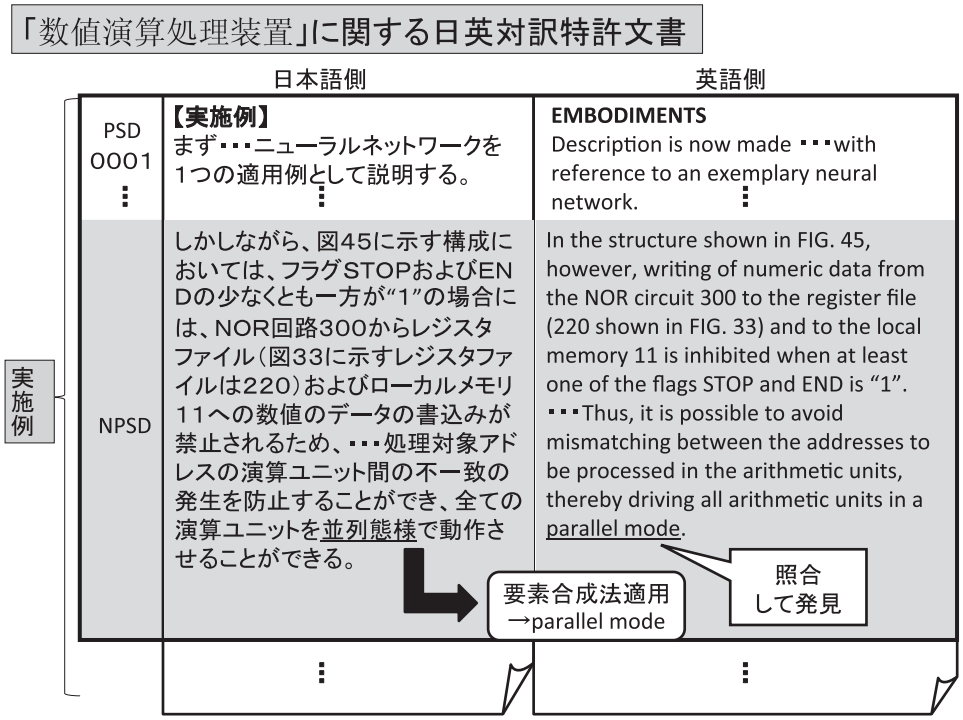

図 2 「実施例」における対訳文対抽出部分 (PSD 部) および非抽出部分 (NPSD 部) の例

するか否かに基づいて y_T の適切さを評価し，スコアを与える (これをコーパススコア $Q_{corpus}(y_T)$ と呼ぶ．)．訳語候補 y_T のスコアは，以下に示すように，この 2 つのスコアの積により定義される．

$$\prod_{i=1}^{n} q(\langle s_i, t_i\rangle) \cdot Q_{corpus}(y_T)$$

実際には，ある訳語候補が 2 つ以上の系列の訳語対から生成される場合があるので，本論文では，以下に示すように，それぞれの系列のスコアの和によって $Q(y_S, y_T)$ を定義する．

$$Q(y_S, y_T) = \sum_{y_S = s_1, s_2, \ldots, s_n} \prod_{i=1}^{n} q(\langle s_i, t_i\rangle) \cdot Q_{corpus}(y_T)$$

訳語対のスコア

訳語対 $\langle s, t\rangle$ のスコア $q(\langle s, t\rangle)$ は $\langle s, t\rangle$ がどの対訳辞書に出現するかで場合分けを行った以下の式によって定義される．

$$q(\langle s, t\rangle) = \begin{cases} 10^{(compo(s)-1)} & \text{英辞郎の場合} \\ \log_{10} f_p(\langle s, t\rangle) & B_P\text{の場合} \\ \log_{10} f_s(\langle s, t\rangle) & B_S\text{の場合} \end{cases}$$

ここで，$compo(s)$ は s の構成要素数，$f_p(\langle s, t\rangle)$ は，対訳辞書 P_2 中に第一要素として $\langle s, t\rangle$ が出現する回数，$f_p(\langle s, t\rangle)$ は，P_2 中に第二要素として $\langle s, t\rangle$ が出現する回数を表す．

コーパススコア

コーパススコア $Q_{corpus}(y_T)$ は，訳語候補 y_T の適切さを，その訳語候補が目的言語コーパスにおいて生起するか否かに基づいて値を定める．

$$Q_{corpus}(y_T) = \begin{cases} 1 & \text{生起する場合} \\ 0 & \text{生起しない場合} \end{cases}$$

5. 対訳文非抽出部分における訳語推定

本論文で用いる日英対訳特許文書の日本語側は，「背景」より前に存在する部分 N_J^1，「背景」B_J，「背景」と「実施例」の間に存在する部分 N_J^2，「実施例」M_J，「実施例」より後に存在する部分 N_J^3 から構成されている．そして，これらの部分のうち，「背景」B_J および「実施例」M_J は，対訳文抽出部分 PSD_J，及び，対訳文非抽出部分 $NPSD_J$ に分割される．この特許文書の構成の例を図 2 に示す．また，英語側の特許文書の全体 D_E に対しても，同様に，対訳文抽出部分 PSD_E，及び，対訳文非抽出部分 $NPSD_E$ に分割されるとする[*4]．

$$D_J = \langle N_J^1, B_J, N_J^2, M_J, N_J^3\rangle$$
$$B_J \cup M_J = \langle PSD_J, NPSD_J\rangle$$
$$D_E = \langle PSD_E, NPSD_E\rangle$$

本論文では，このうちの「背景」B_J 及び「実施例」M_J における対訳文非抽出部分 $NPSD_J$ から日本語専門用語 t_J を抽出した．

[*4] 本論文では，英語側の特許文書において，「背景」部分および「実施例」部分を高精度に抽出することが容易ではなかったため，「背景」部分および「実施例」部分の抽出を行わず，英語側特許文書全体をそのまま利用した．

表 2 日英対訳特許文書 1,000 組における日本語複合名詞の分類

分類	件数 (割合 (%))
英辞郎の英訳が英語側特許文書中に含まれる	345 (1.5)
要素合成法の訳語が英語側特許文書中に含まれる	2,914 (13.0)
英辞郎または要素合成法により，英訳語候補生成可能であるが英語側特許文書中には含まれない	13,165 (58.8)
英辞郎または要素合成法により生成不能	5,972 (26.7)
合計	22,396 (100)

表 3 要素合成法の訳語候補が英語側特許文書中に出現する 500 例の内訳

分類		件数 (割合 (%))
一般語		5 (1.0)
評価対象外		71 (14.2)
専門用語	正解	423 (84.6)
	不正解	1 (0.2)
合計		500 (100)

次に，その日本語専門用語 t_J に対して，日英対訳特許文書の英語側 D_E における対訳文非抽出部分 $NPSD_E$ を英語側コーパスとみなして要素合成法を適用し，英語訳語候補の集合 $TranCand(t_J,\ NPSD_E)$ を作成した．

$$TranCand\Big(t_J,NPSD_E\Big) = \Big\{t_E \in NPSD_E \ \Big|\ t_Jに対して要素合成法により\ t_Eを生成し\ Q(t_J,t_E) > 0\Big\}$$

そして，この $TranCand(t_J,\ NPSD_E)$ を用いて，以下の関数 $\mathrm{CompoTrans}_{\max}$ によりスコア最大となる訳語候補を得る．

$$\mathrm{CompoTrans}_{\max}(t_J,NPSD_E) = \underset{t_E \in TranCand(t_J,NPSD_E)}{\arg\max} Q(t_J,t_E)$$

以上の手順により，日英対訳特許文書の英語側 D_E における対訳文非抽出部分 $NPSD_E$ から英語専門用語 t_E を獲得する．

6. 評価

パテントファミリーである日英対訳特許文書 1,000 文書対を対象として日本語複合名詞を抽出し，その英語訳語を獲得する評価実験を行った．

まず，日英対訳特許文書 1,000 組において，NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に登録されていない日本語複合名詞の分類を表 2 に示す．このうち，要素合成法の訳語が英語側特許文書中に含まれる日本語複合名詞 2,914 件の内の任意の 500 例を抽出し，それの内訳を調査した．また，英語訳語が要素合成法によって生成可能であるが，それが英語側特許文書中に出現しなかった日本語複合名詞 13,165 件，および，英語訳語が英辞郎によって生成不可能である日本語複合名詞 5,972 件の合計 19,137 件の内の任意の 100 例を抽出し，それらの内訳を調査した．

まず，要素合成法の訳語が英語側特許文書中に含まれる日本語複合名詞 500 例を，一般語，評価対象外，専門用語に分類した．この内訳を表 3 に示す．この結果，専門用語は 500 例中 424 例 (84.8%) 含まれており，正解であった専門用語は 424 例中 423 例 (99.8%) であった．ここでの正解とは該当専門用語が日本語特許文書において名詞句として使われており，且つ，その訳語が英語特許文書において名詞句として使われている状態を指す．どちらか一方でも満たしていない場合は不正解とした．また，以下に示す日本語複合名詞は評価対象外とした．

- 接頭辞又は接尾辞が不適切である日本語複合名詞．具体的には「上記〜，下記〜，当該〜，該〜，各〜，〜等，〜毎」が接頭辞又は接尾辞に付いている日本語複合名詞．
- 部分文字列である日本語複合名詞．具体的には，例えば「直角二相変調回路」という全体の文字列の内，部分文字列である「相変調回路」の部分が抽出された日本語複合名詞．
- 末尾が識別子である日本語複合名詞．具体的には「データバッファ装置 DB」のように末尾に「DB」などの識別子の付いている日本語複合名詞．

次に，英辞郎または要素合成法により，英訳語候補生成可能であるが英語側特許文書中には含まれない，もしくは，英辞郎または要素合成法により生成不能である日本語複合名詞 100 例を，同様に一般語，評価対象外，専門用語に分類した．この内訳を表 4 に示す．この結果，専門用語は 100 例中 79 例 (79%) 含まれており，英語側特許文書中に対応する英語表現が存在した専門用語は 79 例中 52 例 (65.8%) であった．

最後に，上述の英語側特許文書中に対応する英語表現が存在した専門用語 52 例から，正解だと想定される日本語

表 4 要素合成法の訳語候補が英語側特許文書中に出現しない，または，要素合成法により訳語候補が生成不能である 100 例の内訳

分類		件数
一般語		4
評価対象外 (名詞句抽出失敗)		17
専門用語	英語側特許文書中に対応する英語表現が存在	52
	英語側特許文書中に対応する英語表現が存在しない	27
合計		100

表 5 英語側特許文書中に対応する英語表現が存在する 52 例の内訳

分類		件数 (割合 (%))
英辞郎に対訳関係が有り過不足なく照合		6 (11.6)
英辞郎に対訳関係無し	冠詞が必要	1 (1.9)
	and が必要	1 (1.9)
	語順のミス	1 (1.9)
	上記以外のエントリー不足	43 (82.7)
合計		52

複合名詞及びその訳語の対を抽出し，英辞郎における対訳関係の有無を調査した．この内訳を表 5 に示す．ここで，英辞郎中に対訳関係が存在するとは，英辞郎・前方一致辞書・後方一致辞書の日英方向の辞書において，当該日本語複合名詞，及び，構成形態素のエントリーが存在しており，かつ，想定した正解における構成要素の語順が日英間で一致している状態を指す．その結果，英辞郎に対訳関係が存在した対数は 52 例中 6 例 (11.6%) であった．

7. 関連研究

既存の辞書には含まれない単語又は表現の訳語対を獲得するための研究はこれまでにも盛んに行われてきた．主にこれらの研究では，パラレルコーパス (同じ内容で言語が異なる文同士が対応) やコンパラブルコーパス (同じ分野で言語が異なる文書同士が対応) を利用している．

パラレルコーパスはコンパラブルコーパスと比べると入手が難しい．しかし，その反面，言語資源としての価値はパラレルコーパスの方に分がある．また，ウェブは，雑音が多いというデメリットを有している反面，多様な分野及び言語の文書データを得ることができ，汎用性が高いというメリットを有している．

パラレルコーパスを利用する方式としては，初期の統計的機械翻訳モデルである IBM モデル [1,2]，および，句に基づく統計的機械翻訳モデルによって自動構築されるフレーズテーブル [6] 等がある．また，その他には，対応する単語対を多く含んでいる文同士は対応する確率が高いという仮定に基づく，統計的共起測定法 [9] がよく知られている．

フレーズテーブル及び既存対訳辞書を用いた専門用語の訳語推定 [10] では，NTCIR-7 特許翻訳タスク [3] において配布されたパラレルコーパス (日英 180 万件の対訳特許文) を利用している．この研究では，句に基づく統計的機械翻訳モデルを用いることにより，対訳特許文から学習されたフレーズテーブル，要素合成法，Support Vector Machines (SVMs) [15] を用いることによって，専門用語対訳対獲得を行っている．また，この方式の後段のタスクとしては，同義対訳専門用語を同定・収集する手法 [8] が提案されている．

これらの先行研究と本論文との違いは，本研究においては，専門用語獲得の際に対訳文非抽出部分を用いているのに対して，先行研究においては，対訳分抽出部分を用いている点である．

一方，コンパラブルコーパスを知識源とする研究においては，初期の頃より，文脈ベクトル [9] がよく用いられてきた．文脈ベクトルとは，文書中における，あるフレーズの近傍位置に出現する語の頻度を並べたベクトルのことである．既存の辞書に登録されていないフレーズの訳語を獲得する場合に，意味的に近いフレーズ同士は似た文脈に出現するという仮説に基づき，文脈ベクトル間の類似度を用いて訳語を獲得する．

また，近年においては，ウェブからコンパラブルコーパスや特定分野のコーパスを収集し，訳語獲得の知識源とする研究なども広く行われている．

スニペット (ウェブ検索結果の要約) からの対訳キーフレーズ抽出 [5] においては，Google[*5]を使用した検索において得られるスニペットを関連語収集の為のコーパスとして利用している．具体的には，キーフレーズの関連語の訳語によって拡張された検索質問を用いてスニペットを獲得し，そのスニペット中の訳語候補から，キーフレーズとその訳語候補間の発音類似性により訳語を獲得している．

ウェブを用いた外国人名の英訳自動獲得 [4] においては，カタカナ表記の外国人名の訳語を，関連語をキーワードと

[*5] http://www.google.com/

する言語横断情報検索及び，発音類似性を利用した訳語推定により獲得している．一方，ウェブを用いた外国人名事典の自動編纂 [11] においても同様の技術を用いているが，人名の収集を含む人名事典の自動編纂全体の自動化を目標としている点が異なる．

また，ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定 [13] においては，要素合成法を適用して生成した訳語候補の検証において，ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定を用いている．

8. おわりに

本論文においては，日米パテントファミリーの対応特許文書中において，対訳文が抽出されなかった「背景」および「実施例」のうちの 70%の部分を言語資源として，専門用語の訳語推定を行った結果について報告した．具体的には，NTCIR-7 特許翻訳タスク [3] において配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に訳語対が登録されていない日本語複合名詞を対象として，既存の対訳辞書を用いた要素合成法を適用した．その結果，約 13%については，対応特許文書の英語側に出現する英訳語が生成可能であり，その精度は 90%以上であった．提案方式を日英対訳特許文書 1,000 文書対に適用したところ，一特許文書対あたり平均二組の対訳専門用語対を収集することができた．

参考文献

[1] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roosin, P. S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85 (1990).

[2] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311 (1993).

[3] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting*, pp. 389–400 (2008).

[4] 後藤功雄，加藤直人，田中英輝，江原暉将，浦谷則好：World Wide Web を用いた外国人名の英訳自動獲得，情報処理学会論文誌， Vol. 47, No. 3, pp. 968–979 (2006).

[5] Huang, F., Zhang, Y. and Vogel, S.: Mining Key Phrase Translations from Web Corpora, *Proc. HLT/EMNLP*, pp. 483–490 (2005).

[6] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th ACL, Companion Volume*, pp. 177–180 (2007).

[7] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. HLT-NAACL*, pp. 127–133 (2003).

[8] 梁　冰，宇津呂武仁，山本幹雄：対訳特許文を用いた同義対訳専門用語の同定と収集，言語処理学会第 17 回年次大会論文集，pp. 963–966 (2011).

[9] Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, *Handbook of Natural Language Processing* (Dale, R., Moisl, H. and Somers, H., eds.), Marcel Dekker Inc., chapter 24, pp. 563–610 (2000).

[10] 森下洋平, 梁　冰, 宇津呂武仁, 山本幹雄：フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定，電子情報通信学会論文誌， Vol. J93–D, No. 11, pp. 2525–2537 (2010).

[11] 榊原洋平，佐藤理史：ウェブを用いた外国人名事典の自動編纂，言語処理学会第 13 回年次大会論文集，pp. 879–882 (2007).

[12] 外池昌嗣，木田充洋，高木俊宏，宇津呂武仁，佐藤理史：要素合成法を用いた専門用語の訳語候補生成・検証，言語処理学会第 11 回年次大会論文集，pp. 17–20 (2005).

[13] 外池昌嗣，宇津呂武仁，佐藤理史：ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定，自然言語処理， Vol. 14, No. 2, pp. 33–68 (2007).

[14] Utiyama, M. and Isahara, H.: A Japanese-English Patent Parallel Corpus, *Proc. MT Summit XI*, pp. 475–482 (2007).

[15] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).