Vol.2013-NL-212 No.12
2013/7/19

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およびフレーズテーブルの利用

豊田 樹生[1]　龍 梓[1]　董 麗娟[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．ここで，我々は，これまで，文献 [8] において，「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，既存の対訳辞書を用いて専門用語の訳語推定を行う方式の有効性を実証した．一方，本論文では，特に，人手で作成された辞書である英辞郎及びその部分対応対訳辞書に加えて，全体の約 30%を訓練例として学習したフレーズテーブルを併用して要素合成法を適用し，専門用語の訳語推定を行う方式を提案する．実際に，評価実験において，パテントファミリー 1,000 組当たり約 7,300 対の専門用語訳語対が獲得できることを示す．

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，フレーズテーブル

Utilizing a Phrase Translation Table and Portion of Patent Families with No Parallel Sentences Extracted in Estimating Translation of Technical Terms

Itsuki Toyota[1]　Zi Long[1]　Lijuan Dong[1]　Takehito Utsuro[2]　Mikio Yamamoto[2]

Abstract: In the previous methods of generating bilingual lexicon from parallel patent sentences extracted from patent families, the portion from which parallel patent sentences are extracted 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 not only from the 30% but also from the remaining 70% out of the whole “Background” and “Embodiment” parts. The proposed method employs the compositional translation estimation technique utilizing the remaining 70% as a comparable corpus for validating translation candidates. As the bilingual constituent lexicons in compositional translation, we use an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the 30%. Finally, we show that about 7,300 technical term translation pairs can be acquired from 1,000 patent families.

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

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

[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

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
フレーズテーブル	33,845,218	33,130,728	76,118,632

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 [6] では，NTCIR-7 特許翻訳タスク [1] 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 [3] 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) 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 [4] においては，文献 [6] 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 [9] 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．文献 [8] では，「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，既存の対訳辞書を用いた専門用語の訳語推定を行った．本論文では，既存の対訳辞書に加えてフレーズテーブルを用い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，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，人手で作成された辞書である英辞郎及びその部分対応対訳辞書とフレーズテーブルを併用した要素合成法を適用した．評価実験において，パテントファミリー 1,000 組当たり約 7,300 対の専門用語訳語対が獲得できることを示す．

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 [1] 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，文献 [9] の手法によって日英間で文対応をつける．

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，前節で述べた文対応データに対して，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 [3] を適用する．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．

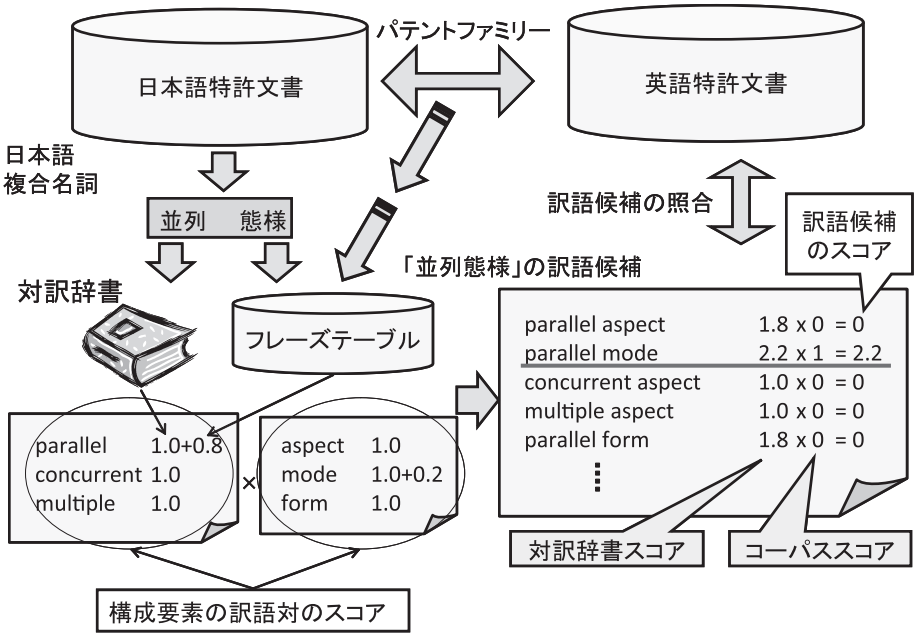

図 1　日本語の専門用語「並列態様」の要素合成法による訳語推定

4. 要素合成法による訳語推定

4.1 既存の対訳辞書及びフレーズテーブル

本研究では，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 [7]，及び，前節で述べたフレーズテーブルを用いる．両者における見出し語数および訳語対数を表 1 に示す．

部分対応対訳辞書生成の手順は以下のとおりである．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．

4.2 訳語候補のスコア

訳語候補のスコアを $Q(y_S, y_T)$ とする．このとき，y_S は日本語専門用語を，y_T は生成された訳語候補を表し，y_S は構成要素 $s_1, s_2, \cdots, s_n$ に，y_T は構成要素 $t_1, t_2, \cdots, t_n$ に分解できると仮定する．

すると，$Q(y_S, y_T)$ は対訳辞書スコア $\prod_{i=1}^{n} q(\langle s_i, t_i \rangle)$ とコーパススコア $Q_{corpus}(y_T)$ の積で定義される．

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．

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

このとき，対訳辞書スコアはこの構成要素同士のスコアの積によって求まり，コーパススコアは訳語候補が目的言語側のコーパスに出現するか否かによって求まる．

4.2.1 対訳辞書スコア

構成要素の訳語対 $\langle s, t \rangle$ の対訳辞書スコア $q(\langle s, t \rangle)$ は訳語対が英辞郎, 前方一致部分対応対訳辞書 B_P，または，後方一致部分対応対訳辞書 B_S に含まれる場合のスコア q_{man} 及び訳語対がフレーズテーブルに含まれる場合のスコア q_{smt} の和によって定まる．

$$q(\langle s, t \rangle) = q_{man}(\langle s, t \rangle) + q_{smt}(\langle s, t \rangle)$$

$$q_{man}(\langle s, t \rangle) = \begin{cases} 1 & (\langle s, t \rangle \text{ が英辞郎，} B_P, \\ & \text{または，} B_S \text{ に含まれる場合}) \\ 0 & (\text{それ以外の場合}) \end{cases}$$

$$q_{smt}(\langle s, t \rangle) = \begin{cases} P(t|s) & (\langle s, t \rangle \text{ がフレーズ} \\ & \text{テーブルに含まれ，} \\ & \text{かつ } P(t|s) \geq p_0 \\ & \text{である場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \quad (1)$$

上記の定義においては、訳語対 $\langle s, t \rangle$ が英辞郎，B_P，また

[*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/index.php?JUMAN

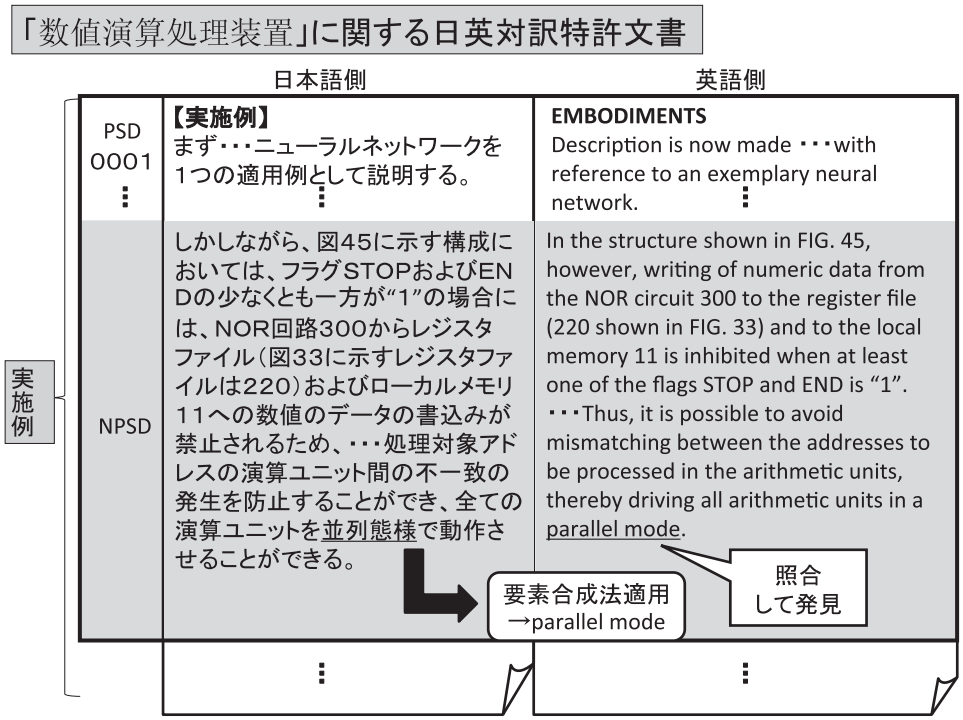

図 2 「実施例」における対訳文対非抽出部分

は，B_S に含まれる場合，$q_{man}(\langle s,t\rangle)$ は 1 となり，それ以外の場合は 0 となる[*4]．一方，$\langle s,t\rangle$ がフレーズテーブルに含まれる場合は，翻訳確率の下限 p_0 のパラメータに従い，スコアを決定する．このパラメータ p_0 は，6 節において，評価用データ以外の調整用データを用いて最適化される．

4.2.2 コーパススコア

コーパススコアは，訳語候補 y_T が目的言語側のコーパスに出現する場合にのみ 1 となり，出現しない場合には 0 となる．

$$Q_{corpus}(y_T) = \begin{cases} 1 & (y_T \text{ が目的言語側コーパスに出現する場合}) \\ 0 & (y_T \text{ が目的言語側コーパスに出現しない場合}) \end{cases}$$

4.2.3 例

例として，専門用語 “並列態様” の対訳 “parallel mode” を獲得する様子を図 1 に示す．本論文では，まず，この日本語専門用語 “並列態様” を構成要素 s_1 の “並列” と s_2 の “態様” に分解し，これらを既存の対訳辞書及びフレーズテーブルを利用して目的言語に翻訳する．そうすると s_1 からは t_1 として “parallel”,“concurrent”,“multiple” が，s_2 からは t_2 として “aspect”,“mode”,“form” が生成され，さらに各々に訳語の参照元に応じたスコアが付与される．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．このとき，各々の訳語候補の対訳辞書スコアは t_1 と t_2 のスコアの積となる．例えば，“parallel aspect” の対訳辞書スコアは $(1.0 + 0.8) \times 1.0 = 1.8$ である．

最後に，これら訳語候補を対訳辞書スコア順に，目的言語側のコーパスに対して照合を行い，もし照合すればそのコーパススコアは 1，照合しなければ 0 となる．この場合，結果的に，訳語候補のスコアが最も高い “parallel mode” が獲得される．

[*4] 文献 [7] においては，英辞郎中の訳語対のスコアは構成要素の数が多いほど大きくなり，一方，部分対応対訳辞書中の訳語対のスコアは，英辞郎における部分対訳対の頻度が大きいほど大きくなるというスコア体系が採用されている．本論文において，このスコア体系の評価を行なった結果では，本論文で用いているスコア体系との間で大きな性能差は観測されなかったため，本論文では，より簡易なスコア体系を採用した．

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」B_J，「実施例」M_J，および，「背景・実施例以外の部分」N_J 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．また，英語側の特許文書の全体 D_E に対しても，同様に，「背景」B_E，「実施例」M_E，および，「背景・実施例以外の部分」N_E から構成され，「背景」B_E および「実施例」M_E は，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．この特許文書の構成の例を図 2 に示す．

表 2 パテントファミリー 1,000 組における日本語複合名詞の分類

(1) 対象日本語複合名詞の集合 = 全日本語複合名詞 61,133 個の集合

対訳辞書の種類		英辞郎のみ	フレーズテーブルのみ	英辞郎及びフレーズテーブル
(a)	英辞郎の英訳が英語側特許文書中に含まれる	5,449 (8.9%)		
(b)	フレーズテーブルの日本語側と完全一致	32,516 (53.2%)		
(c)	要素合成法 (提案手法) の訳語が英語側特許文書中に含まれる	4,004 (6.6%) (集合 E)	14,310 (23.4%) (集合 P，ただし， $\|P\|$ 最大の場合 ($p_0 = 0$))	14,575 (23.8%) (集合 EP，ただし， $\|EP\|$ 最大の場合 ($p_0 = 0$))
(d)	英辞郎または要素合成法 (提案手法) 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	397 (0.6%)	993 (1.6%)	1,041 (1.7%)
(e)	英辞郎または要素合成法 (提案手法) により生成不能	18,767 (30.7%)	7,865 (12.9%)	7,552 (12.4%)
合計		61,133 (100%)		

(2) 対象日本語複合名詞の集合 = 全日本語複合名詞 61,133 個の集合 − 集合 (a) − 集合 (b) − 集合 E

対訳辞書の種類		フレーズテーブルのみ	英辞郎及びフレーズテーブル
(c)	要素合成法 (提案手法) の訳語が英語側特許文書中に含まれる	10,375 (17.0%) (集合 $P - (E \cap P)$)	10,571 (17.3%) (集合 $EP - (E \cap EP)$)

$$
\begin{aligned}
D_J &= \langle B_J,\, M_J,\, N_J \rangle \\
B_J \cup M_J &= \langle PSD_J,\, NPSD_J \rangle \\
D_E &= \langle B_E,\, M_E,\, N_E \rangle \\
B_E \cup M_E &= \langle PSD_E,\, NPSD_E \rangle
\end{aligned}
$$

ここで，本論文では，英訳語推定対象となる日本語専門用語 t_J(実際に，6 節において評価を行なう際には，日本語複合名詞を抽出し訳語推定を行なう) を抽出するにあたっては，対訳文抽出部分 PSD_J 中の日本語専門用語の英訳語の多くは対訳文から学習したフレーズテーブル中に含まれると予測し，「背景」B_J 及び「実施例」M_J における対訳文非抽出部分 $NPSD_J$ を抽出元とした．

次に，その日本語専門用語 t_J に対して，英語側の「背景」B_E 及び「実施例」M_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 $TranCand(t_J,\, B_E \cup M_E)$ を作成した[*5]．

$$
\begin{aligned}
&TranCand\Big(t_J,\, B_E \cup M_E\Big) \\
&= \Big\{ t_E \in B_E \cup M_E \;\Big|\; t_J\text{に対して要素合成法により} \\
&\qquad\qquad t_E\text{を生成し } Q(t_J,t_E) > 0 \Big\}
\end{aligned}
$$

そして，この $TranCand(t_J,\, B_E \cup M_E)$ 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．

$$
\begin{aligned}
&\mathrm{CompoTrans}_{\max}(t_J,\, B_E \cup M_E) \\
&= \operatorname*{argmax}_{t_E \in TranCand(t_J,\, B_E \cup M_E)} Q(t_J,t_E)
\end{aligned}
$$

以上の手順により，日英対訳特許文書の英語側の「背景」B_E 及び「実施例」M_E から英語専門用語 t_E を獲得する．

6. 評価

6.1 評価対象日本語複合名詞集合の作成

提案手法を評価するため，以下の 3 通りの比較を行った．

(i) 「英辞郎のみ」 ··· 対訳辞書として，英辞郎及びその構成要素から生成される辞書を用いる．

(ii) 「フレーズテーブルのみ」 ··· 対訳辞書としてフレーズテーブルを用いる．

(iii) 「英辞郎及びフレーズテーブル」 ··· 対訳辞書として，英辞郎及びフレーズテーブルを用いる．

はじめに，パテントファミリー 1,000 組を取り出し，そこから 61,133 例の日本語複合名詞を抽出した．次に，これら

[*5] ここで，比較評価として，英語側の「背景」B_E 及び「実施例」M_E における対訳文非抽出部分 $NPSD_E$ のみを英語側コーパスとみなして要素合成法を適用する評価実験も行ったが，英語側コーパス中において適切な訳語候補を照合できる割合が下がったため，本論文においては，英語側の「背景」B_E 及び「実施例」M_E を英語側コーパスとみなして要素合成法を適用する方式を採用した．

表 3 提案手法の評価結果及びパテントファミリー 1,000 組当たりの訳語対獲得数の推定値

(1) 要素合成法において用いた対訳辞書の種類別

対訳辞書の種類	英辞郎のみ	フレーズテーブルのみ	英辞郎及びフレーズテーブル
評価対象日本語 複合名詞の集合	$E' \subset E$, $\|E'\| = 93$	$P' \subset P - (E \cap P)$, $\|P'\| = 224$	$EP' \subset EP - (E \cap EP)$, $\|EP'\| = 230$
再現率 (%) 適合率 (%) F 値 (%)	97.8 97.8 97.8	(再現率 $> 20\%$，適合率最大の場合 ($p_0 = 0.07$)) 26.8 / 89.0 / 41.2	(再現率 $> 30\%$，適合率最大の場合 ($p_0 = 0.15$)) 32.2 / 92.5 / 47.7
訳語対 獲得数の 推定値	$3,918$ ($= 4,004 \times 0.978$) (対象日本語 複合名詞集合 E, $\|E\| = 4,004$)	(再現率 $> 20\%$，適合率最大の場合 $2,779$ ($= 10,375 \times 0.268$) (対象日本語複合名詞集合 $P - (E \cap P)$, $\|P - (E \cap P)\| = 10,375$)	(再現率 $> 30\%$，適合率最大の場合) $3,401$ ($= 10,571 \times 0.322$) (対象日本語複合名詞集合 $EP - (E \cap EP)$, $\|EP - (E \cap EP)\| = 10,571$)

(2) 全日本語複合名詞 61,133 個を対象とした合計値

	集合 E に対して英辞郎のみを用いて訳語推定 + 集合 $P - (E \cap P)$ に対して フレーズテーブルのみを用いて訳語推定	集合 E に対して英辞郎のみを用いて訳語推定 + 集合 $EP - (E \cap EP)$ に対して 英辞郎及びフレーズテーブルを用いて訳語推定
訳語対 獲得数の 推定値	6,697 (= 3,918+2,779)	7,319 (= 3,918+3,401)

61,133 例の日本語複合名詞に対して 4 節で述べた要素合成法を適用し，表 2-(1) に示すように，以下の 5 つのカテゴリに分類した．

(a) 日本語複合名詞が英辞郎に含まれ，その訳語が英語側特許文書中に含まれる．
(b) (a) 以外の場合で，日本語複合名詞がフレーズテーブルの日本語側と完全一致する．
(c) (a)，(b) 以外の場合で，日本語複合名詞に対して提案手法の要素合成法を適用した結果，その訳語が英語側特許文書中に含まれる．
(d) (a)，(b)，(c) 以外の場合で，日本語複合名詞に対する訳語が英辞郎及び，提案手法の要素合成法によって生成されるが，英語側特許文書中に含まれない．
(e) (a)，(b)，(c)，(d) 以外の場合で，日本語複合名詞に対する訳語が英辞郎及び提案手法の要素合成法により生成されない．

以下では，まず，表 2-(1) の (c) 欄に示すように，上記 (i)∼(iii) の 3 通りの対訳辞書を用いて要素合成法を適用した結果，訳語候補が英語側特許文書に含まれる日本語複合名詞の集合を求める．

- 対訳辞書として上記 (i)「英辞郎のみ」を用いた場合について，求められた日本語複合名詞の集合を集合 E (4,004 例) とする．
- 対訳辞書として上記 (ii)「フレーズテーブルのみ」を用いた場合について，求められた日本語複合名詞の集合を集合 P とする．ただし，4.2.1 節，式 (1) の翻訳確率の下限 p_0 については，要素数 $|P|$ が最大となる場合の値，$p_0 = 0$ を用い，$|P| = 14,310$ となる．
- 対訳辞書として上記 (iii)「英辞郎及びフレーズテーブル」を用いた場合について，求められた日本語複合名詞の集合を集合 EP とする．ただし，4.2.1 節，式 (1) の翻訳確率の下限 p_0 については，要素数 $|EP|$ が最大となる場合の値，$p_0 = 0$ を用い，$|EP| = 14,575$ となる．

次に，表 3-(1) に示すように，集合 E，$P - (E \cap P)$，$EP - (E \cap EP) = EP - E$ より，評価用の日本語複合名詞の集合 E' (93 例)，P' (224 例)，EP' (230 例) をそれぞれ作成する．ただし，ここで，集合 E 中の日本語複合名詞については，対訳辞書として上記 (i)「英辞郎のみ」を用いることにより，大半の日本語複合名詞の英訳語を正しく推定できることが分かっているため，集合 P および EP に対しては，集合 E に含まれる日本語複合名詞を除外し，集合 $P - (E \cap P)$，および，$EP - (E \cap EP) = EP - E$ を作成した後，これらを母集団として評価用の日本語複合名詞の集合 P'，および，EP' をそれぞれ作成した．

6.2 評価結果

前節で作成した評価用の日本語複合名詞の集合 E'，P'，EP' を対象として，上記 (i)∼(iii) の 3 通りの対訳辞書を用いて要素合成法を適用し，再現率，適合率，F 値を求めた結果を表 3-(1) に示す．ただし，集合 P'，および，EP' を対象とした評価においては，上述の評価用の日本語複合名

詞の集合以外のパラメータ調整用の集合を用いて，再現率が 20∼30%程度，適合率が 80∼90%程度となるパラメータ (翻訳確率の下限値 p_0) の調整を行なった．具体的には，集合 P' に対して，対訳辞書として上記 (ii)「フレーズテーブルのみ」を用いた場合の評価においては，調整用の集合 $P''(\subseteq P-(E\cap P))$ を用いて，再現率 >20%，適合率最大となるパラメータ p_0 を求め，評価を行なった．一方，集合 EP' に対して，対訳辞書として上記 (iii)「英辞郎及びフレーズテーブル」を用いた場合の評価においては，調整用の集合 $EP''(\subseteq EP-(E\cap EP))$ を用いて，再現率 >30%，適合率最大となるパラメータ p_0 を求め，評価を行なった．

さらに，

「適合率が 80∼90%程度の性能のもとで訳語推定を行なった結果に対して，人手で訳語候補の正誤判定を行ない，訳語候補が適切であると判定できた場合にのみ，訳語対を獲得する」

という方式を仮定し，評価対象の日本語複合名詞の集合の母集団である集合 (E，$P-(E\cap P)$，および，$EP-(E\cap EP)$) の要素数と訳語推定の再現率の積によって，訳語対獲得数の推定値を求めた結果を表 3-(1) の「訳語対獲得数の推定値」欄に示す．上記 (i)∼(iii) の 3 通りの対訳辞書を用いた場合の結果を以下に示す．

- 対訳辞書として上記 (i)「英辞郎のみ」を用いた場合は，評価集合 E' における再現率，適合率，F 値とも 97.8%となり，訳語対獲得数の推定値は 3,918 対となった．この結果から，集合 E の大半の日本語複合名詞に対して，英語訳語を正しく推定できることが分かった．
- 対訳辞書として上記 (ii)「フレーズテーブルのみ」を用いた場合は，評価用集合 P' における再現率は 26.8%，適合率は 89.0%となり，訳語対獲得数の推定値は 2,779 対となった．
- 対訳辞書として上記 (iii)「英辞郎及びフレーズテーブル」を用いた場合は，評価用集合 EP' における再現率は 32.2%，適合率は 92.5%となり，訳語対獲得数の推定値は 3,401 対となった．

以上より，対訳辞書として上記 (iii)「英辞郎及びフレーズテーブル」を用いた場合に，全体として 90%以上の適合率で訳語候補の推定を行なうことができ，表 3-(2) に示すように，全日本語複合名詞 61,133 例を対象とした場合には，合計で約 7,300 対の専門用語対訳対を獲得できることが分かった．この結果から，フレーズテーブルと英辞郎を提案手法により組み合わせて用いることで，フレーズテーブルあるいは英辞郎を単独で用いる場合と比較して，より多くの専門用語訳語対を高精度に獲得できることがわかった．

7. 関連研究

文献 [5, 6, 10] では，パテントファミリーから抽出された対訳特許文を用いて，訳語対の獲得を行っている．しかし，本論文では，対訳特許文が抽出されなかった残りの部分を利用して新たな専門用語訳語対の獲得を行っている点が異なる．また，文献 [4] では，複数の対訳特許文において，ある日本語専門用語に対して複数の訳語が出現するという状況を考えて，同義対訳専門用語の同定と収集を行っている．上記の手法と本論文の手法を併用することは比較的容易であると考えられる．

一方，提案手法と比較して，コンパラブルコーパスからの訳語対獲得手法 (例えば，文献 [2]) においては，通常，文脈ベクトル等の類似性を言語間で測定した情報を手がかりとする点が特徴である．これに対して，本論文の手法において訳語推定の情報源として用いているパテントファミリーは，一般のコンパラブルコーパスと比較すると，対訳となっている部分の割合がかなり高い点にその特徴がある．本論文の手法においては，この利点を生かして要素合成法を適用することにより，比較的容易に訳語対の獲得を実現している．

8. おわりに

本論文では，パテントファミリーから専門用語の対訳辞書を生成する方法について述べた．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，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，人手で作成された辞書である英辞郎及びその部分対応対訳辞書とフレーズテーブルを併用した要素合成法を適用した．パテントファミリー 1,000 組から約 7,300 対の専門用語訳語対を獲得できることを示した．今後の課題として，専門用語訳語推定の過程に対して，句に基づく統計的機械翻訳モデルを直接適用してその性能を提案手法と比較し，提案手法の有効性を評価する予定である．

参考文献

[1] 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).

[2] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL*, pp. 414–420 (1998).

[3] 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).

[4] 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).

[5] Lu, B. and Tsou, B. K.: Towards Bilingual Term Extraction in Comparable Patents, *Proc. 23rd PACLIC*, pp. 755–762 (2009).

[6] 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).
[7] 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).
[8] 豊田樹生，高橋佑介，牧田健作，宇津呂武仁，山本幹雄：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，情報処理学会研究報告，Vol. 2012–NL–208 (2012).
[9] Utiyama, M. and Isahara, H.: A Japanese-English Patent Parallel Corpus, *Proc. MT Summit XI*, pp. 475–482 (2007).
[10] Yasuda, K. and Sumita, E.: Building a Bilingual Dictionary from a Japanese-Chinese Patent Corpus, *Computational Linguistics and Intelligent Text Processing*, LNCS, Vol. 7817, Springer, pp. 276–284 (2013).