パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定*

豊田 樹生† 梁 冰‡ 宇津呂 武仁§ 山本 幹雄§
筑波大学 理工学群工学システム学類†
筑波大学大学院 システム情報工学研究科‡ 筑波大学 システム情報系§

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 [4] 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 [2] 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) [7] 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 [3] においては，文献 [4] 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 [6] 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，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，既存の対訳辞書を用いた要素合成法 [5] を適用し，90% 以上の高い精度で訳語の推定が可能であることを示す．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．

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

†Itsuki Toyota, College of Engineering Systems, School of Science and Engineering, University of Tsukuba
‡Bing Liang, Graduate School of Systems and Information Engineering, University of Tsukuba
§Takehito Utsuro, Mikio Yamamoto, Faculty of Engineering, Information and Systems, University of Tsukuba

2 日英対訳特許文

本論文では，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，フレーズテーブルの訓練用データとして使用した．この文対応データは，1993-2000 年発行の日本公開特許広報全文と米国特許全文を対象として，文献 [6] によって日英間で文対応を付けたものである．

3 要素合成法による訳語推定

3.1 既存の対訳辞書

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 [5] 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

[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

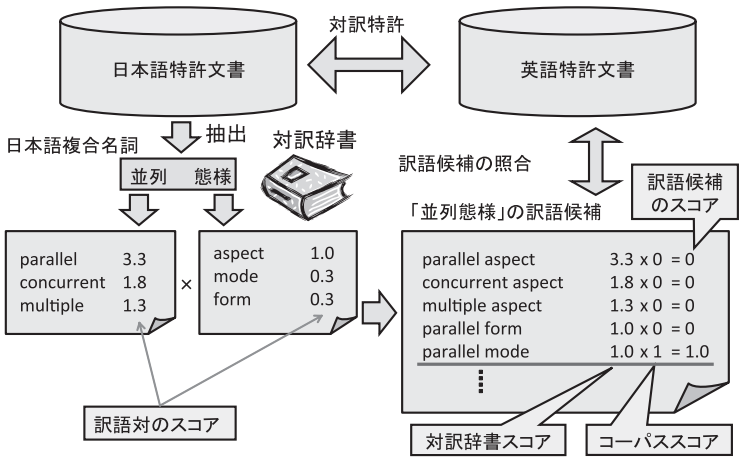

図 1: 日本語の専門用語「並列態様」の要素合成法による訳語推定

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

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．

3.2 訳語候補のスコア

訳語候補のスコアは，対訳辞書スコア $Q_{dict}(y_S, y_T)$ とコーパススコア $Q_{corpus}(y_T)$ の積で定義される．

$$Q(y_S, y_T) = Q_{dict}(y_S, y_T) \cdot Q_{corpus}(y_T)$$

ここで 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．また，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生起しているか否かによって求まる．

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 $3.3 \times 1.0 = 3.3$ 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 になる．この場合，結果的に，訳語候補のスコアが一番高い “parallel mode” が獲得されることになる．

4 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]．

$$\begin{aligned} D_J &= \langle N_J^1, B_J, N_J^2, M_J, N_J^3 \rangle \\ B_J \cup M_J &= \langle PSD_J, NPSD_J \rangle \\ D_E &= \langle PSD_E, NPSD_E \rangle \end{aligned}$$

[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．

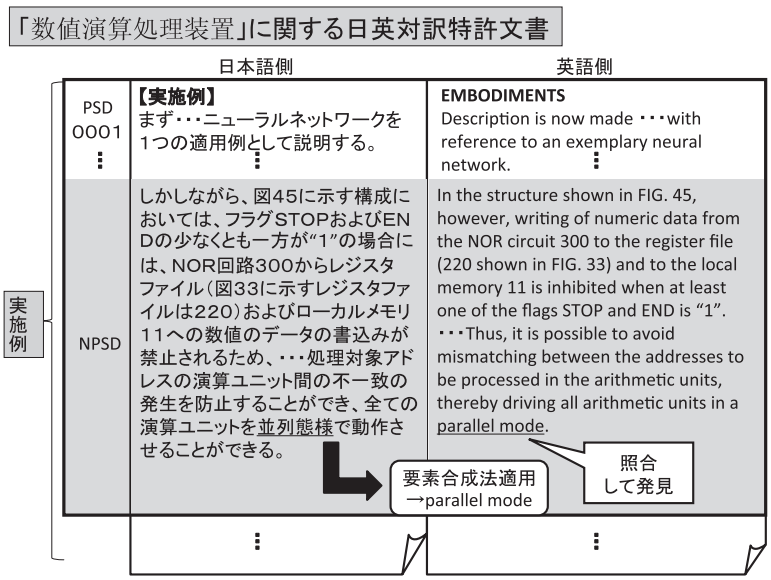

図 2: 「実施例」における対訳文非抽出部分

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．

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．

5 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 組における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 100 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 件を抽出し，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 100 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 100 例中 89 例 (89%) 含まれており，正解であった専門用語は 89 例中 89 例 (100%) 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，(i) 接頭辞又は接尾辞が不適切である，(ii) 部分文字列である，(iii) 末尾が識別子である，の場合は評価対象外とした[5]．

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，

[5]接頭辞又は接尾辞が不適切とは「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専門用語を指す．部分文字列であるとは，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専門用語を指す．末尾が識別子とは，例えば「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専門用語を指す．

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 100 例の内訳

分類		件数
一般語		3
評価対象外		8
専門用語	正解	89
	不正解	0
合計		100

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．

6 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 [1] 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，既存の対訳辞書を用いた要素合成法を適用し，90% 以上の高い精度で訳語の推定が可能であることを示した．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．

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

参考文献

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

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[3] 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. 言語処理学会, 2011.

[4] 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.

[5] 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.

[6] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pp. 475–482, 2007.

[7] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.