日米パテントファミリーを利用した専門用語訳語推定

—対訳文対抽出部分・非抽出部分の併用—

Estimating Translation of Technical Terms
by Integrating Japanese-English Patent Families as a Parallel Corpus and a Comparable Corpus

筑波大学システム情報系知能機能工学域教授 **宇津呂　武仁**

PROFILE: 1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。京都大学等を経て、2012 年より筑波大学システム情報系知能機能工学域教授。自然言語処理、機械翻訳、ウェブマイニングの研究に従事。

筑波大学大学院システム情報工学研究科知能機能システム専攻 **豊田　樹生**

PROFILE: 2012 年筑波大学理工学群工学システム学類卒業。機械翻訳の研究に従事。

筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻教授 **山本　幹雄**

PROFILE: 1986 年豊橋技術科学大学大学院情報工学系修士課程修了。豊橋技術科学大学等を経て、2008 年より筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻教授。博士（工学）。自然言語処理、機械翻訳の研究に従事。

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、パテントファミリーを情報源として、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。その中で、筆者らは、文献 [4] において、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手法を提案した。この手法では、句に基づく統計的機械翻訳モデル（Moses）[2] によって、日英 180 万件の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 [5]、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、筆者らは、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、日本語専門用語が出現する多数の対訳特許文を利用することにより、同義対訳専門用語を同定･収集する方式 [3] を提案し、その有効性を実証した。

ここで，上述の日英 180 万件の対訳特許文は，文献 [7] 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（以下では、「対訳文対抽出部分」と呼ぶ）の割合は約 30% にとどまっている．そこで、本稿では、この 30% の「対訳文対抽出部分」、および、「背景」および「実施例」のうちの残りの 70% の部分（以下では、「対訳文対非抽出部分」と呼ぶ）を併用して、日本語専門用語の英訳語を推定する方式 [6] を紹介する。この方式においては、「対訳文対非抽出部分」をコンパラブルコーパスとみなし、既知の対訳辞書を用いて専門用語の構成要素の訳語を求め、要素合成法 [5] の考え方に基づき構成要素の訳語を結合することによって専門用語の訳語候補を合成する。そして、コンパラブルコーパス中の目的言語コーパスに対して、合成された訳語候補を検証することによって、訳語推定を行う。なお、要素合成法において用いる既知の対訳辞書としては、人間用の対訳辞書として人手で作成された英辞郎、および、「対訳文対抽出部分」から学習されたフレーズテーブルを併用する。本稿では、評価実験の結果として、パテントファ

ミリー 1,000 組を対象としてこの手法を適用することにより、約 92% 程度の精度で、約 4,750 対の日英専門用語対訳対を獲得できることを示す。

2 要素合成法による訳語推定

2.1 既存の対訳辞書およ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]Ver.131 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]、および、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。ここで、英辞郎中の日英訳語対のうち、日本語側が二形態素、英語側が二単語から構成される訳語対に対して、日英とも第一構成要素となる日本語形態素・英語単語対を抽出し、前方一致部分対応対訳辞書とする。同様に、日英とも第二構成要素となる日本語形態素・英語単語対を抽出し、後方一致部分対応対訳辞書とする。

表 1: 既存の対訳辞書およびフレーズテーブル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

2.2 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、訳語候補のスコアは、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。

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、訳語対が英辞郎または部分対応対訳辞書に含まれる場合のスコア q_{man}、および、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} の和によって求まる。

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

$$q_{man}(\langle s,t\rangle)=\begin{cases}1 & (\langle s,t\rangle \text{ が英辞郎、または、部分対応対訳辞書に含まれる場合})\\ 0 & (\text{それ以外の場合})\end{cases}$$

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

$q_{man}(\langle s,t\rangle)$ は、$\langle s,t\rangle$ が英辞郎、または、部分対応対訳辞書に含まれる場合には 1 となり、それ以外の場合には 0 となる。一方、$q_{smt}(\langle s,t\rangle)$ は、$\langle s,t\rangle$ がフレーズテーブルに含まれ、かつ、翻訳確率 $P(t|s)$ が下限 p_0 以上である場合には、翻訳確率 $P(t|s)$ となり、それ以外の場合には 0 となる。評価実験においては、この下限 p_0 は、調整用データ・セットを用いて最適化される。

一方、コーパススコア $Q_{corpus}(y_T)$ 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。

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

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、

1 http://www.eijiro.jp/

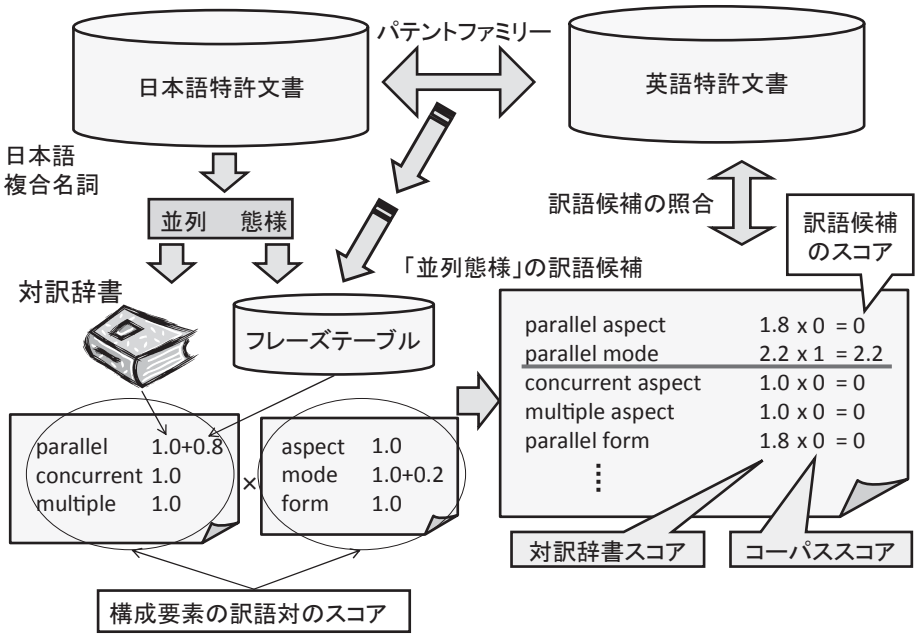

図 1: 要素合成法を用いた日本語専門用語「並列態様」の訳語推定

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とともに、対訳辞書スコアを付与する。具体的には、ここでは、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、各訳語に対して対訳辞書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、各訳語候補の対訳辞書スコアは t_1 の対訳辞書スコアと t_2 の対訳辞書スコアの積となる。例えば、“parallel aspect”の対訳辞書スコアは（1.0+0.8）×1.0=1.8 となる。

最後に、目的言語側のコーパスに対してこれらの訳語候補の照合を行い、照合すればそのコーパススコアを 1、照合しなければ 0 とする。【図 1】の例では、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。

3 対訳文対非抽出部分における訳語推定

本論文で用いる日米パテントファミリーの日本語側 D_J 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、対訳文対抽出部分 PSD_J、及び、対訳文対非抽出部分 $NPSD_J$ 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、対訳文対抽出部分 PSD_E、及び、対訳文対非抽出部分 $NPSD_E$ に分割される。この特許文書の構成の例を【図 2】に示す。

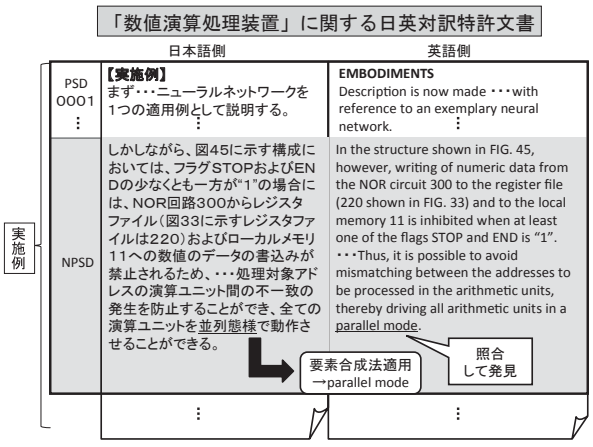

図 2：日米パテントファミリー中の「実施例」における対訳文対非抽出部分の例

$$D_J = \langle B_J, M_J, N_J \rangle$$
$$B_J \cup M_J = \langle PSD_J, NPSD_J \rangle$$
$$D_E = \langle B_E, M_E, N_E \rangle$$
$$B_E \cup M_E = \langle PSD_E, NPSD_E \rangle$$

ここで、本稿では、英訳語推定対象となる日本語専門用語 t_J 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対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 における対訳文対非抽出部分 $NPSD_J$ を抽出元とした。そして、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 を目的言語側のコーパスとみなして要素合成法を適用し、訳語候補の集合を作成した。最後に、スコア最大となる訳語候補を訳語推定結果として獲得した。

4 評価

本稿で述べた手法を日米パテントファミリー 1,000 組に対して適用し、日本語専門用語に対する英訳語推定の評価実験を行った。前節で述べたように、各パテントファミリー中より「背景」および「実施例」部分を抽出し、MeCab[2]（IPAdic）によって日本語文の形態素解析を行った後、名詞・接頭辞・接尾辞・未知語のいずれかの品詞の最長の形態素列に加えて、数字・アルファベットの列が連接することを許容したもの、合計 61,133 例を日本語名詞句として抽出した。このうち、フレーズテーブルの日本語側と完全一致した 32,516 例を除外し、さらに、英辞郎によって英訳が可能な日本語名詞句のうち、英訳語が英語側特許文書中に含まれる 5,449 例を除外し、23,168 例が残った。これらの日本語名詞句のうち、専門用語以外の一般名詞句、形態素解析における区切り位置誤りを伴う語、複合名詞抽出規則の不備の影響を受けた語、等、訳語推定の対象として不適切であるものが約 40-50% 程度含まれていたため、これらを除外した残りの約 50-60% の日本語名詞句を評価実験の対象とした。

2 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

これらの日本語名詞句を対象として、英辞郎、部分対応対訳辞書、および、フレーズテーブルを用いた要素合成法により訳語推定を行ったところ、人間用の対訳辞書である英辞郎および部分対応対訳辞書を用いた場合には、訳語推定可能な語の範囲は相対的に狭いものの、高精度な訳語推定が可能であることがわかった。一方、フレーズテーブルのみを用いた場合、および、英辞郎・部分対応対訳辞書に加えてフレーズテーブルを併用した場合には、フレーズテーブルにおける訳語対の翻訳確率の下限を緩めるほど、訳語推定可能な語の範囲は広がるものの、訳語推定の精度が低下することがわかった。そこで、【表 2】に示すように、英辞郎・部分対応対訳辞書のみを用いた要素合成法により、英語特許側において検証可能な訳語候補が生成できる 3,715 語を対象として訳語推定を行ったところ、94.4% の精度を達成した。さらに、これらの 3,715 語を除外した後、英辞郎・部分対応対訳辞書・フレーズテーブルを併用した要素合成法により、英語特許側において検証可能な訳語候補が生成できる語を対象として、調整用のデータセットを用いて、85% 程度の精度を条件としてフレーズテーブルにおける訳語対の翻訳確率の下限の調整を行ったところ、翻訳確率の下限値が 0.15 となった。また、その場合に、英語特許側において検証可能な訳語候補が生成できる日本語名詞句は 1,460 語となり、訳語推定精度は 85.3% となった。これらの結果を統合すると、合計で 5,175 語を対象として約 92% 程度の精度で訳語推定が行えることがわかった。

表 2: 訳語推定の評価結果

対象日本語専門用語	要素合成法において用いる対訳辞書	精度 (%)
英辞郎・部分対応対訳辞書のみを用いた要素合成法により、英語特許側において検証可能な訳語候補が生成できる 3,715 語	英辞郎・部分対応対訳辞書	94.4
上記の 3,715 語を除いて、英辞郎・部分対応対訳辞書・フレーズテーブルを併用した要素合成法により、英語特許側において検証可能な訳語候補が生成できる 1,460 語	英辞郎・部分対応対訳辞書・フレーズテーブル (翻訳確率 ≧ 0.15)	85.3
合計 5,175 語		91.9

5 おわりに

本稿では、日米パテントファミリーの対応特許文書を知識源として、専門用語訳語対応を獲得する手法について述べた。特に、パテントファミリー中において、「対訳文対抽出部分」および「対訳文対非抽出部分」を併用して、日本語専門用語の英訳語を推定する方式を紹介した。この方式においては、要素合成法[5]の考え方に基づき日本語専門用語の構成要素の訳語を結合することによって訳語候補を合成し、コンパラブルコーパス中の目的言語コーパスに対して、合成された訳語候補を検証することによって、訳語推定を行った。要素合成法において用いる既知の対訳辞書としては、人間用の対訳辞書として人手で作成された英辞郎、および、「対訳文対抽出部分」から学習されたフレーズテーブルを併用した。評価実験の結果においては、人間用の対訳辞書は適用範囲が狭いものの高精度であり、一方、自動学習されたフレーズテーブルは、適用範囲が広いものの精度面で劣るという結果となった。そこで、フレーズテーブル中の翻訳確率に下限を設けることによって、適用範囲と精度の間のバランスをとった結果、パテントファミリー 1,000 組を対象とした場合において、約 92% 程度の精度で、約 4,750 対の日英専門用語対訳対を獲得できることを示した。

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] B. Liang, T. Utsuro, and M. Yamamoto. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. Procedia - Social and Behavioral Sciences, Vol. 27 pp.50-60, 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，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.

[6] I. Toyota, Z. Long, L. Dong, T. Utsuro, and M. Yamamoto. Compositional translation of technical terms by integrating patent families as a parallel corpus and a comparable corpus. In Proc. 5th Workshop on Patent Translation, pp. 16-23, 2013.

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