日中パテントファミリーを利用した同義対訳専門用語の同定

Identifying Bilingual Synonymous Technical Terms utilizing Japanese-Chinese Patent Families

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**

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 **龍　梓**

2010 年北京航空航天大学ソフトウェア学院卒業。2013 年北京航空航天大学大学院ソフトウェア工学専攻修了。2015 年筑波大学大学院システム情報工学研究科博士前期課程修了。現在、同博士後期課程在学中。機械翻訳の研究に従事。

筑波大学システム情報系情報工学域教授 **山本　幹雄**

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

ここ数年、中国の特許文献数が飛躍的に増大しており、中国語の特許文献を日本語で検索する必要性が高まっており、中国の特許を日本語に翻訳する仕事の重要性が高まっている。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、日英パテントファミリーから作成された日英対訳特許文を用いて、日英専門用語対訳対獲得を行った。文献 [1][3] では、日中パテントファミリーを情報源として、日中対訳特許文から日中専門用語対訳対を獲得する手法を提案している。しかし、これらの手法では、ある日本語専門用語の中国語訳語を獲得することはできるが、日中専門用語対訳対の集合における同義・異義の関係を同定することはできない。

一方、文献 [8] では、日英パテントファミリーの対訳特許文から、句に基づく統計的機械翻訳モデルのフレーズテーブルを用いて専門用語を収集し、Support Vector Machines（SVMs）[7] を適用することにより、日英専門用語対訳対の同義・異義関係の判定を行っている。そこで、本稿では、文献 [8] と同様に、日中パテントファミリーを情報源とし、ある日本語専門用語が出現する複数の対訳文を入力として中国語訳語の推定を行うことにより、同義となる日中専門用語対訳対を同定する [9]。

2 日中対訳特許文

本稿では、フレーズテーブルの訓練用データとして、約 360 万件の日中対訳特許文を用いた。この日中対訳特許文は、2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文に対して、以下の手順によって得られたものである。

1. 文献 [6] の手法によって日中間で文対応を付ける。

2. スコア降順で上位の 360 万文対を抽出する。

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

本研究では、文献 [3] の場合と同様に、専門用語の訳語推定において、日中対訳特許文から学習したフレーズテーブルを用いる。フレーズテーブルは、2. 節で述べた日中対訳特許文に対して、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[2] を適用することにより、日中の句の組、及び、日中の句が対応する確率を推定し記述する。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. 対訳文に対する前処理として、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を行う。
2. 対訳文から中日、日中の両方向において最尤な単語対応を得る。
3. 中日、日中両方向の単語対応から、ヒューリスティックスを用いて対称な単語対応を得る。
4. 対称な単語対応を用いて、可能な全ての日中の句対応の組を作成する。
5. 対訳文における日中の句の対応の数を集計し、各句の対応に翻訳確率を付与する。

本稿では、手順 1 の対訳文は、形態素解析された形態素単位の日本語文一文に対して、Chinese Penn Treebank を用いた Stanford Word Segment[5] によって形態素解析された形態素単位の中国語文、及び、文字単位 [10][1] の中国語文の二種類を用意し、作成されたものである。このような 2 つの対訳文に対して、独立に Moses を適用することにより、形態素単位フレーズテーブルおよび文字単位フレーズテーブルをそれぞれ作成した。その際、日本語フレーズの形態素数の上限、中国語側形態素単位フレーズテーブルの中国語形態素数の上限、および、中国語側文字単位フレーズテーブルの中国語文字数の上限を、いずれも 15 とした。

1 連続する数字とアルファベットは一個のトークンとして扱う。

4 フレーズテーブルを用いた専門用語対訳対の同義集合の生成

4.1 専門用語対訳対同義候補集合の作成

図 1 に、専門用語対訳対同義候補集合作成の流れを示す。

1. 360 万文の特許文から無作為に抽出した初期日本語専門用語 t_J^0 に対し、全対訳特許文 360 万件から学習されたフレーズテーブル [2] を用いて訳語推定を行い、中国語訳語を得る。
2. 1 で得られた中国専門用語に対して訳語推定を行い、日本語訳語を得る。
3. 1、2 の手順を繰り返し、k 回訳語推定を行うことにより得られた対訳専門用語を集めた集合を CBP（t_J^0）とする（本論文では、k=6 とした）。

なお、手順 3 においては、以下の条件の全てを満た

2 ただし、日中方向の訳語推定を行う場合は、日中方向のフレーズテーブルの順位が一位となる中国語訳語を用い、中日方向の訳語推定を行う場合は、中日方向のフレーズテーブルの順位が一位となる日本語訳語を用いた。また、形態素単位フレーズテーブルと文字単位フレーズテーブルは、それぞれ独立に用いて、訳語推定を行う。なお、フレーズテーブルを用いた日中方向の訳語推定の精度、「形態素単位」では 97.8% で、「文字単位」では 95.9% である。

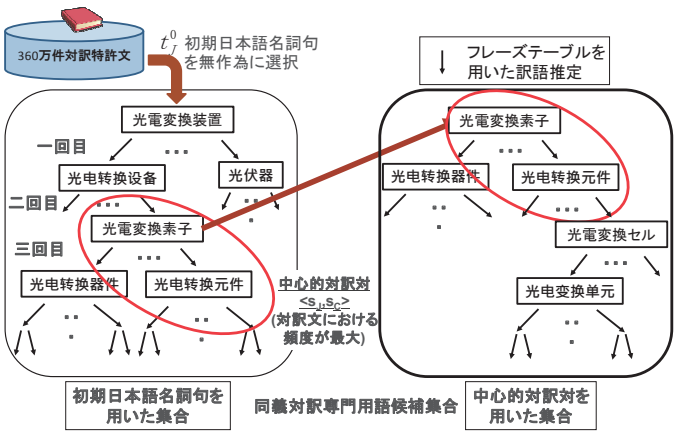

図 1　専門用語対訳対の訓練・評価用同義・異義集合の作成

す対訳対 〈t_J,t_C〉（ただし、t_J, t_C はそれぞれ日本語専門用語、及び中国語専門用語）のみ残し、その他の組を枝刈りする。

1. t_J, t_C のいずれの頻度も 12,500 未満。
2. t_J, t_C のいずれの頻度も 700 未満、又は、長さの下限[3] を満す。
3. t_J, t_C いずれも語頭及び語尾が機能語、数字、句読点でない（これらはいずれも、フレーズ自動抽出時に自動生成されたものであり、専門用語の語頭・語尾としては不適切なものである）。
4. t_J, t_C の頻度が 3,000 未満。

本稿では、以上の手順に従って、4,000 個の初期日本語名詞句を用いて、専門用語対訳対の同義候補集合 CBP(t_J^0) を作成した。なお、本稿では、専門用語対訳対同義候補集合 CBP(t_J^0) に対して、要素数の下限を設定した（具体的には、|CBP(t_J^0)| ≥ 10|）。

4.2 中心的対訳対を用いた参照用同義集合の作成

次に、前節で作成した同義候補集合 CBP(t_J^0) 中の専門用語対訳対の中から、「一般語の対訳対」を除いて、360 万対訳文中の頻度が最大となる対訳対を選定し、**中心的対訳対** s_{JC}= 〈s_J,s_C〉 とする[4]。

以上の手順に従って、合計 114 個の中心的対訳対を選定した。次に、中心的対訳対 s_{JC} のうちの日本語専門用語 s_J を用いて、前節の手順によって専門用語対訳対同義候補集合 CBP(s_J) を作成した。作成された同義候補集合中の対訳対数を表 1 に示す。なお、以上の過程においては、訳語対応として正しくない対訳対を人手で除外した。

最後に、人手によって、同義候補集合 CBP(s_J) を、中心的対訳対 s_{JC} と同義となる対訳対の集合 SBP(s_{JC})、および、その他の対訳対の集合 NSBP(s_{JC}) に分割した。ここで、表 1 においては、中国語側が形態素単位のフレーズテーブルを用いた場合の同義候補集合、及び、中国語側が文字単位のフレーズテーブルを用いた場合の同義候補集合の両方に共通に含まれる専門用語対訳対を示している。ただし、中国語側の形態素解析誤りが原因で、同一の文字列に対する形態素分割のパターンが 2 通り以上出現する場合があるため、表 1（a）における共通専門用語対訳対数の方が表 1（b）よりも多くなっている。

5 同義・異義判定のための素性

同義専門用語対訳対の同定に用いた素性を表 2 に示す。素性は大きく、対訳対 〈t_J,t_C〉 の特性を規定するも分類器学習を用いた同義対訳専門用語の同定の、および、対訳対 〈t_J,t_C〉 と中心的対訳対 〈s_J,t_C〉 の間の関係を規定するものの 2 種類に分けられる。

3 t_J が (i) 連続する漢字長が 3 以上、(ii) 漢字数が 4 以上、(iii) 文字数が 6 以上、かつ、形態素数が 2 以上、(iv) 一形態素の場合は 10 文字以上、のいずれかを満たし、かつ、t_C が (i) 文字数が 4 以上、(ii) 形態素数が 2 以上の場合は 3 文字以上、のいずれかを満たす。

4 本稿では、文献 [8] 同様、専門用語対訳対同義候補集合中において中心的対訳対を選定し、中心的対訳対との間でのみ同義・異義を識別するという、より簡単化したタスクを設定する。

表 1　作成された専門用語対訳対同義候補集合中の対訳対数

(a) 中国語側が形態素単位のフレーズテーブルを用いた場合

		総要素数		114個の集合の間の平均対数	
同義候補集合 $U_{s_J} CBP(s_J)$	形態素単位の集合のみに含まれる	12,640	24,621	110.9	216.0
	文字単位の集合と共通	11,981		105.1	
人手で同定した候補集合 $U_{s_{JC}} SBP(s_{JC})$	形態素単位の集合のみに含まれる	228	2,473	2.0	21.7
	文字単位の集合と共通	2,245		19.7	

(b) 中国語側が文字単位のフレーズテーブルを用いた場合

		総要素数		114個の集合の間の平均対数	
同義候補集合 $U_{s_J} CBP(s_J)$	文字単位の集合のみに含まれる	6,358	17,478	55.8	153.3
	形態素単位の集合と共通	11,120		97.5	
人手で同定した候補集合 $U_{s_{JC}} SBP(s_{JC})$	文字単位の集合のみに含まれる	287	2,318	2.5	20.3
	形態素単位の集合と共通	2,031		17.8	

表２　専門用語対訳対の同義・異義同定のための素性

分類	素性名	定義 (ただし，X ∈ {J, C}，(Y,Z) ∈ {(J, C), (C, J)})
対訳対 $\langle t_J, t_C \rangle$ の特性を規定	f_1 ： 共起頻度	対訳特許文における$\langle t_J, t_C \rangle$の共起頻度の二進対数.
	f_2 ： 中国語訳語の順位	条件付き確率P(t_C | t_J)の降順にt_Cを順位付けしたときのt_Cの順位の二進対数.
	f_3 ： 日本語訳語の順位	条件付き確率P(t_J | t_C) の降順にt_J を順位付けしたときのt_Jの順位の二進対数.
	f_4 ： 日本語文字数	t_Jの文字数
	f_5 ： 中国語文字数	t_Cの文字数.
	f_6 ： 訳語推定における繰り返しの回数	s_Jから訳語推定を開始し，訳語としてt_Yを生成した直後にt_Yからt_Z を訳語推定した場合の，s_J からt_Z までの繰り返し訳語生成回数.
対訳対 $\langle t_J, t_C \rangle$ と中心的対訳対 $\langle s_J, s_C \rangle$ の間の関係を規定する	f_7 ： 日本語用語が同一	$t_J = s_J$ならば，1 となる.
	f_8 ： 中国語用語が同一	$t_C = s_C$ならば，1 となる.
	f_9 ： 編集距離類似度	$f_9(t_X, s_X)=1-\frac{ED(t_X, s_X)}{\max(\|t_X\|,\|s_X\|)}$: ED はt_X とs_X の間の編集距離，$\|t\|$はt に含まれる文字数を表す.
	f_{10} ： バイグラム類似度	$f_{10}(t_X, s_X)=\frac{\|bigram(t_X)\cap bigram(s_X)\|}{\max(\|t_X\|,\|s_X\|)-1}$: $bigram(t)$ は，t に含まれる文字単位のバイグラムの集合.
	f_{11} ： 日本語用語の同一形態素の割合	$f_{11}(t_J, s_J)=\frac{\|const(t_J)\cap const(s_J)\|}{\max(\|const(t_J)\|,\|const(s_J)\|)}$: $const(t)$ は日本語用語t に含まれる形態素単語の集合.
	f_{12} ： 中国語用語の同一文字の割合	$f_{12}(t_C, s_C)=\frac{\|const(t_C)\cap const(s_C)\|}{\max(\|const(t_C)\|,\|const(s_C)\|)}$: $const(t)$ は中国語用語t に含まれる文字の集合.
	f_{13} ： 日本語用語の文字列の包含関係もしくは異表記	t_J とs_J は，以下のいずれかの関係を満たす．(i) 構成要素の差分は接尾辞のみ，(ii) 構成文字列の差分は，長音「ー」のみ，(iii) 構成文字列の差分は，送り仮名の違いのみ.
	f_{14} ： 中国語用語の文字列の包含関係	t_C とs_C の構成要素の差分は語頭・語尾でない「的」のみ.
	f_{15} ： フレーズテーブルの共通訳の割合	$f_{15}(t_X, s_X)=\frac{\|trans(t_X)\cap trans(s_X)\|}{\max(\|trans(t_X)\|,\|trans(s_X)\|)}$: $trans(t)$ は，フレーズテーブルから得られる用語t のすべての訳語の集合.
	f_{16} ： 全非共有箇所に対しフレーズテーブルにおける共通訳の割合	t_X とs_Xの間で文字列が一致しない箇所$x_t{}^1, \ldots, x_t{}^m$, $x_s{}^1, \ldots, x_s{}^n$に対して，$x_t{}^i (i = 1, \ldots, m)$ と $x_s{}^j (j = 1, \ldots, n)$ の1 対1 対応に対して，フレーズテーブルから得られる訳語の集合$trans(x_t{}^i)$および $trans(x_s{}^j)$ 中の共通訳の割合を求め，その共通訳の割合の積$(i = 1, \ldots, m, j = 1, \ldots, n)$ が最大となる 1 対1 対応において，共通訳の割合の積を素性値とする.
	f_{17} ： フレーズテーブルの訳語関係が存在	フレーズテーブル中にt_Y とs_Z の訳語関係が存在する．（$\langle t_J, s_C \rangle$ または$\langle s_J, t_C \rangle$のどちらか一方のみの訳語関係が存在することを表す素性，および，$\langle t_J, s_C \rangle$ と$\langle s_J, t_C \rangle$の両方の訳語関係が存在することを表す素性の二種類を区別して用いる）.

6 機械学習を用いた同義・異義判定

6.1 適用手順

前節で示した素性を用いて、中国語側が形態素単位の場合の同義候補集合、および、中国語側が文字単位の場合の同義候補集合に対して、それぞれ独立に SVM を適用し、同義・異義判定の評価を行った。4.2 節において作成した専門用語対訳対同義候補集合 CBP(s_J) を全参照用事例として、8 割を用いて SVM の訓練を行い、残りのうちの 1 割を用いて 2 種類のパラメータの調整を行い、最後の 1 割を評価用事例とした。以上の手順を 10 通り繰り返し、その平均値を算出し同義判定の性能評価を行った。2 種類のパラメータの調整においては、同義判定の適合率を最大化する場合、および、同義判定の F 値を最大化する場合の 2 通りの調整を行った。なお、本稿で調整の対象としたパラメータは、分離平面から評価用事例までの距離の下限である。

6.2 評価結果

表 3 に、同義判定における性能の評価結果を示す。ベースラインとしては、「t_J と s_J が同一、または、t_C と s_C が同一の場合に、対訳対 〈t_J,t_C〉 と中心的対訳対 〈s_J,s_C〉 と同義である」という規則を用いた。同義判定の適合率を最大化する調整を行った場合は、「形態素単位」では 86.5% の適合率を達成し、「文字単位」では 89.0% の適合率を達成した。一方、同義判定の F 値が最大化する調整を行った場合、「文字単位」も「形態素単位」も、ベースラインを上回る F 値を達成した。

次に、ベースラインによる同義判定の結果を、SVM によって改善する例を表 4 に示す。

表３　同義判定の性能評価 (%)

手法			適合率	再現率	F 値
形態素単位	ベースライン		71.4	40.0	51.3
	SVM	適合率最大	**86.5**	26.5	40.5
		F 値最大	64.3	64.1	**64.2**
文字単位	ベースライン		74.0	40.1	52.0
	SVM	適合率最大	**89.0**	26.1	40.4
		F 値最大	63.5	65.3	**64.4**

表４　同義判定における SVM による改善例

ベースライン：	t_J と s_J が同一，または，t_C と s_C が同一の場合に，対訳対 $\langle t_J, t_C \rangle$ は中心的対訳対 $\langle s_J, s_C \rangle$ と同義である
SVM：	中国語側が形態素単位のフレーズテーブルを用いた場合，適合率が最大となる下限を用いたモデル

(a) SVM のみで同義と判定し正解

中心的対訳対 $<s_J, s_C>$	専門用語対訳対 $<t_J, t_C>$	人手による同義・異義判定	ベースラインによる判定	SVM による判定
<グラス転移温度，**玻璃化转变温度**>	<グラス転移点，**玻璃态转化温度**>	同義	異義	同義

(b) SVM のみで異義と判定し正解

中心的対訳対 $<s_J, s_C>$	専門用語対訳対 $<t_J, t_C>$	人手による同義・異義判定	ベースラインによる判定	SVM による判定
<集電装置，**集电器**>	<コレクト，**集电器**>	異義	同義	異義

表５　同義判定における提案手法の誤り例

(a) 提案手法により同義と判定し不正解

中心的対訳対 $<s_J, s_C>$	専門用語対訳対 $<t_J, t_C>$	日本語側		中国語側		素性 f_{17}（両方の訳語関係が存在）	素性 f_{17}（片方の訳語関係のみが存在）	人手による同義・異義判定	提案手法による判定
		素性 f_9	素性 f_{10}	素性 f_9	素性 f_{10}				
<断熱体，**绝热体**>	<インシュレータ，**绝缘件**>	0	0	0.33	0	1	1	異義	同義

(b) 提案手法により異義と判定し不正解

中心的対訳対 $<s_J, s_C>$	専門用語対訳対 $<t_J, t_C>$	日本語側		中国語側		素性 f_{17}（両方の訳語関係が存在）	素性 f_{17}（片方の訳語関係のみが存在）	人手による同義・異義判定	提案手法による判定
		素性 f_9	素性 f_{10}	素性 f_9	素性 f_{10}				
<成膜室，**成膜室**>	<成膜チャンバー，**膜成形室**>	0.29	0.17	0.5	0	0	1	同義	異義

表４(a)「SVM のみで同義と判定し正解」の例においては、専門用語対訳対と中心的対訳対の日本語表記および中国語表記の両方とも異なる場合（$t_J \neq s_J$、$t_C \neq s_C$）、ベースラインでは異義であると判定されたが、提案手法では、「f_{17}: フレーズテーブルの訳語関係が存在」（フレーズテーブルにおいて「グラス転移温度」の訳語として“**玻璃态转化温度**”が存在し、「グラス転移点」の訳語として“**玻璃化转变温度**”が存在しており。f_{17}（$\langle t_J, t_C \rangle$, $\langle s_J, s_C \rangle$）=1）となる素性の効果によって、同義と判定できた。

一方、表４（b）「SVM のみで異義と判定し正解」の例においては、専門用語対訳対の中国語表記と中心的対訳対の中国語表記が同一のため（t_C=s_C）、ベースラインでは同義であると判定されたが、提案手法では、日本語用語 t_J「集電装置」および s_J「コレクト」の文字列の間で、素性「f_9：編集距離類似度」および素性「f_{10}：バイグラム類似度」のいずれも値が０となった（f_9（$\langle t_J, t_C \rangle$, $\langle s_J, s_C \rangle$）=0、および、f_{10}（$\langle t_J, t_C \rangle$, $\langle s_J, s_C \rangle$）=0）。提案手法では、これらの素性の効果によって異義と判定できた。

最後に、提案手法による誤り例を表５に示す。

表５（a）「提案手法により同義と判定し不正解」の例では、素性「f_{17}：フレーズテーブルの訳語関係が存在」において、フレーズテーブル中に誤った対訳対〈断熱体、**绝缘件**〉、および〈インシュレータ、**绝热体**〉、が含まれることが原因で、「f_{17}：フレーズテーブル中に $\langle t_J, s_C \rangle$、$\langle s_J, t_C \rangle$ 両方の訳語関係が存在」および「f_{17}: フレーズテーブル中に $\langle t_J, s_C \rangle$ または $\langle s_J, t_C \rangle$ の片方の訳語関係のみが存在」の両方の値が１となってしまい、最終的に誤って同義と判定されてしまった。この場合、フレーズテーブル中の対訳対の正誤判定を行う分類器の訓練・適用過程を導入することによって、素性 f_{17} の判定精度を高めることにより誤りを改善できると考えられる。

一方、表５（b）「提案手法により異義と判定し不正解」の例では、素性「f_{17}: フレーズテーブルの訳語関係が存在」において、対訳対〈成膜チャンバー , 成膜室〉、のみがフレーズテーブルに含まれることから、「f_{17}: フレーズテーブル中に $\langle t_J, s_C \rangle$ または $\langle s_J, t_C \rangle$ の片方の訳語関係のみが存在」の値は１となるものの「f_{17}: フ

レーズテーブル中に〈t_J,s_C〉,〈s_J,t_C〉両方の訳語関係が存在」の値が 0 となっている。また、中国語文字列“成膜”と“膜成形”は実際同義関係にあるにも関わらず、文字列が逆順となっていることが原因でバイグラム類似度が 0 となっている。主としてこれらが原因となって、最終的に誤って異義と判定されてしまった。この場合、文字列の順序の異なりを反映しない文字列類似度に相当する素性を導入することによって、誤りが改善できると考えられる。

7 おわりに

本稿では、専門用語対訳対の獲得というタスクにおける同義語同定問題を解決する手法を提案した。提案手法では、対訳特許文および句に基づく統計的機械翻訳モデルのフレーズテーブルを用いて専門用語対訳対を自動収集し、それに対して、SVM を適用することにより、専門用語対訳対間の同義・異義関係の判定を行った。日中パテントファミリーから抽出した 360 万対の日中対訳文に対して提案手法を適用し、同義関係にある日中対訳専門用語の同定において、再現率が 25% 以上という条件のもとで、約 90% の適合率を達成した。今後の課題として、再現率を改善するため、文献 [11] で提案された、人手の介在を併用する半自動的な同義対訳専門用語の同定の枠組を開発することが重要であると考えられる。

謝辞

本研究においては、日本特許情報機構（Japio）より提供して頂いた日中パテントファミリーのデータを利用させて頂いた。関係各位に感謝の意を表す。

参考文献

[1] Dong, L., Long, Z., Utsuro, T., Mitsuhashi, T., and Yamamoto, M.: Collecting Bilingual Technical Terms from Japanese-Chinese Patent Families by SVM, in Proc. PACLING, pp. 71–79, 2015.

[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] Yasuda, K. and Sumita, E.: Building a Bilingual Dictionary from a Japanese-Chinese Patent Corpus, in Computational Linguistics and Intelligent Text Processing, Vol.7817 of LNCS, pp. 276–284, Springer, 2013.

[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] H. Tseng, P. Chang, G. Andrew, D.Jurafsky, and C. Manning. A conditional random field word segmenter for SIGHAN bake off 2005. In Proc. 4th SIGHAN Workshop on Chinese Language Processing, pp. 168–171, 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.

[8] B. Liang, T. Utsuro and M. Yamamoto. Identifying Bilingual Synonymous Technical Terms from Phrase Tables and Parallel Patent Sentences, in Procedia - Social and Behavioral Sciences, Vol. 27, No. 4, pp. 50-60, 2011

[9] 龍梓，董麗娟，宇津呂武仁，三橋朋晴，山本幹雄．日中対訳文を用いた同義対訳専門用語の同定手法．情報処理学会論文誌，Vol.56, No.3, pp.960-971, 2015.

[10] Sun, J. and Lepage, Y.: Statistical Machine Translation between Unsegmented Japanese and Chinese Texts, 言語処理学会第 19 回年次大会発表論文集，pp. 122–125, 2013.

[11] B. Liang, T. Utsuro, and M. Yamamoto. Semi-automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In Proc. 25th PACLIC, pp. 196–205, 2011.