情報処理学会論文誌 Vol.56 No.3 960–971 (Mar. 2015)

日中対訳文を用いた同義対訳専門用語の同定手法

龍 梓[1] 董 麗娟[1] 宇津呂 武仁[2,a)] 三橋 朋晴[3] 山本 幹雄[2]

受付日 2014年6月30日, 採録日 2014年11月10日

概要：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．その中でも有力な手法の1つとして，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とSVMを用いて，専門用語対訳対獲得を行う手法があげられる．しかし，この手法では，ある日本語専門用語の訳語推定の際に，その日本語専門用語が出現する1つの対訳文に出現する訳語のみを推定対象としていた．したがって，他の対訳文に出現している同義の専門用語対訳対とはまったく独立に訳語推定が行われており，本来同義関係にある複数の専門用語対訳対の間の関係を同定できない，という問題点があった．そこで，本論文では，ある日本語専門用語およびその同義語候補が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．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，専門用語対訳対の同義・異義関係の判定を行う．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．

キーワード：訳語対獲得，同義語，フレーズテーブル，専門用語

Identifying Bilingual Synonymous Technical Terms from Japanese-Chinese Parallel Sentences

Long Zi[1] Dong Lijuan[1] Takehito Utsuro[2,a)] Tomoharu Mitsuhashi[3]
Mikio Yamamoto[2]

Received: June 30, 2014, Accepted: November 10, 2014

Abstract: In the process of translating patent documents, a bilingual lexicon of technical terms is inevitable knowledge source. It is important to develop techniques of acquiring technical term translation equivalent pairs automatically from parallel patent documents. We take an approach of utilizing the phrase table of a state-of-the-art phrase-based statistical machine translation model. Especially, in the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considers situations where a technical term or its synonym candidate is observed in many parallel patent sentences and is translated into many translation equivalent and studies the issue of identifying synonymous translation equivalent pairs. First, we collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, we apply the Support Vector Machines (SVMs) to the task of identifying bilingual synonymous technical terms. Finally, we achieve the performance of over 90% precision with the condition of more than or equal to 25% recall.

Keywords: bilingual lexicon acquisition, synonyms, phrase tables, technical terms

[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, Ibaraki 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, Ibaraki 305–8573, Japan
[3] 日本特許情報機構
Japan Patent Information Organization (JAPIO), Koto, Tokyo 135–0016, Japan
[a)] utsuro@iit.tsukuba.ac.jp

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等といったサービスにおいて不可欠である．機械翻訳[*1]においても，また，翻訳者による翻訳においても，大規模で正確な対訳辞書は，高い訳質を保証するための必要な情報源である．しかし，人手によって対訳辞書を作成し，継続的に収録語数を増やし辞書を維持・管理していく作業は膨大な時間と労力を要する．そこで，自然言語処理分野においては，多様なテキストデータを情報源として，対訳辞書を自動もしくは半自動的に作成する技術に関する研究が行われてきた．

これまでの研究を大まかに分類すると，初期の研究としては，二言語間で文間の対応が付けられた文対応対訳コーパスを情報源として，対訳文中の共起頻度を用いる手法 [1] がよく研究された．また，文対応対訳コーパスよりも利用可能性の高いコンパラブルコーパスを情報源とする手法（たとえば，文献 [2], [3], [4]）も，初期の時期から近年に至るまでよく研究されている．さらに，近年では，多言語文書を収集する情報源として，ウェブ上の多様な言語・分野・ジャンルの文書を利用する研究も数多く進められている．たとえば，複合語である専門用語の構成要素の訳語を連結して訳語の候補を生成する要素合成法，および，ウェブから収集した目的言語の専門分野コーパスを用いて，生成された訳語候補を検証する手法 [5]，検索エンジン等を利用して訳語が併記された文書を収集し，訳語対を獲得する手法訳語対を獲得する手法 [6], [7]，多言語文書である Wikipedia を情報源とする手法 [8] 等がある．

それらの研究の中で，特に，専門用語対訳辞書の作成においては，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 [9] では，NTCIR-7 の特許翻訳タスク [10] において配布された日英 180 万件の対訳特許文を情報源として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 [11] を用いることにより日英対訳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）[12] による分類器学習を用いることによって，専門用語対訳対獲得における適合率 91.9%，再現率 70%を達成した．また，文献 [13] では，文献 [9] と同様の考え方に基づき，日中パテントファミリーから抽出した 360 万件の日中対訳特許文を言語資源としてフレーズテーブルを学習し，そこ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，対訳特許文を情報源として専門用語対訳対獲得を行うこれらの手法 [9], [13] では，ある日本語専門用語の訳語推定の際に，その日本語専門用語が出現する 1 つの対訳文に出現する訳語のみを推定対象としていた．したがって，他の対訳文に出現している同義の専門用語対訳対とはまったく独立に訳語推定が行われており，本来同義関係にある複数の専門用語対訳対の間の関係を同定できない，という問題点があった．そこで，本論文では，ある日本語専門用語およびその同義語候補が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．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，対訳関係にある日中専門用語の対（「日中対訳専門用語」と呼ぶ）$\langle t_J, t_C \rangle$（ただし，t_J，t_C はそれぞれ日本語専門用語，および中国語専門用語）を多数収集する．そして，2 組の日中対訳専門用語 $\langle t_J, t_C \rangle$ および $\langle t'_J, t'_C \rangle$ の間で以下の同義関係を定義し，この同義関係の判定を行うタスクに対して SVM を適用するというアプローチをとる．

$$\langle t_J, t_C \rangle \text{ と } \langle t'_J, t'_C \rangle \text{ が同義である} \quad \longleftrightarrow \quad \begin{array}{c} t_J \text{ と } t'_J, \text{ および,} \\ t_C \text{ と } t'_C \text{ の組が} \\ \text{それぞれ同義である.} \end{array}$$

本論文の実際の手順においては，まず，ある日本語専門用語を種として，同義関係にある専門用語対訳対の候補を生成・収集する．生成・収集した候補集合の中から同義判定を行うための**中心的対訳対**を選び，中心的対訳対のうちの日本語専門用語に対して，専門用語対訳対同義候補集合を再生成する．再生成した候補集合に対して SVM 分類器を適用することにより，同義集合・異義集合を同定する．

本論文では特に，日英パテントファミリーから抽出した日英対訳特許文を対象として日英の同義対訳専門用語の同定を行った手法 [14] に対して，日中パテントファミリーから抽出した日中対訳特許文を対象として日中の同義対訳専門用語の同定を行い，その有効性を実証した．具体的には，文献 [14] において日英を対象として提案された素性に対して，日中を対象とする場合の素性を定義しなおすとともに，文献 [14] における素性の組合せを改善する形で導入された新たな素性として「フレーズテーブルにおける共通訳の割合」を提案した．さらに，評価実験を通して，最適な性能を達成する素性の組合せを示した．その結果として，新素性である「フレーズテーブルにおける共通訳の割合」が性能に大きな影響を持つ重要な素性であることを示した[*2]．実際に，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

[*1] ここでの「機械翻訳」とは，狭義には，「人手によって作成された規則を用いる古典的機械翻訳」のことを指すが，6 章末尾において簡単に触れるように，近年の統計的機械翻訳の研究においては，コーパスから獲得した語彙知識を有効に活用して翻訳性能およびカバレージを改善する方式も提案されている．

[*2] ここで，新素性「フレーズテーブルにおける共通訳の割合」は言語対に対して独立な素性であるので，本論文で対象とした日中間の同義対訳専門用語同定タスクだけでなく，文献 [14] における対象である日英間の同義対訳専門用語同定タスクにおいても効果的である可能性は十分にあると考えられる．

条件のもとで，約 90%の適合率を達成した．さらに，比較対象として，日英同義対訳専門用語の同定を対象とした先行研究 [15] における素性と同等の素性のもとで日中同義対訳専門用語の同定を行った評価結果との比較を行い，本論文で提案する素性の組合せによって大幅に性能が改善されることを示した．

2. 日中対訳特許文

本論文では，フレーズテーブルの訓練用データとして約 360 万対の日中対訳特許文を使用した．なお，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 年中国特許全文を対象として，以下の手順で得られたものである．

(1) 2004–2012 年発行の日本公開特許広報全文と 2005–2010 年中国特許全文を対象として，総計 312,492 件の日中パテントファミリー全文を得る．
(2) 日中パテントファミリーに対して，文献 [16] の手法によって日中間で文対応を付ける[*3]．
(3) 抽出された約 2,400 万件の日中対訳文のうち，スコア降順上位の 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を用いた訳語推定

本論文では，専門用語の訳語推定において，約 360 万件の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．

3.1 フレーズテーブルの作成

フレーズテーブルにおいては，2 章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 [11]（バージョン 0.91）を適用することにより，日中の句の組，および，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) 文対応データ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，共起単語表の作成を行う．
(2) IBM モデルにより文対応データから単語対応を生成するツール GIZA++ [17] を用いて，中日，日中の両方向に対して最尤な単語対応を得る[*4]．
(3) 日中，中日両方向の単語対応から，（パラメータ `alignment` を `grow-diag-final-and` とする）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，可能なすべての日中の句の組を作成し，各組に対して，「文単位の句対応制約」[*5]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見なす）．
(5) 文対応データにおける日中の句の対応数に基づいて，各句の対応に翻訳確率等のパラメータを付与する．

ここで，手順 (1) の対訳文は，MeCab [*6] によって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 [18] によって形態素解析された形態素単位の中国語文，および，文字単位[*7]の中国語文の 2 種類を用意し，作成されたものである．このような 2 種類の対訳文に対して，独立に Moses を適用することより，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．

3.2 1 組の日中対訳文およびフレーズテーブルを用いた訳語推定

本論文では，フレーズテーブルおよび日中対訳特許文を用いて，専門用語の訳語推定を行う．訳語推定手法において，1 つの日中対訳文を対象として，その日中対訳文に出現する用語の訳語対（用語対訳対）を推定する．

日中方向 日本語用語 t_J に対して，t_J が出現する日中対訳文のうちの 1 組 $\langle S_J, S_C \rangle$ およびフレーズテーブルを用い，以下の条件をすべて満たす訳語候補 t_C を推定する[*8]．

(1) t_J の訳語として，フレーズテーブルに存在する．
(2) (1) を満たす訳語であり，かつ，与えられた対訳文 $\langle S_J, S_C \rangle$ の中国語文 S_C に出現する．
(3) (1)，(2) を満たす訳語の中で，フレーズテーブルにおける翻訳確率 $P(t_C \mid t_J)$ が最大である[*9]．

中日方向 中国語用語 t_C に対して，t_C の出現する日中対訳文のうちの 1 組 $\langle S_J, S_C \rangle$ およびフレーズテーブルを用い，以下の条件をすべて満たす訳語候補 t_J を推定する．

[*3] 文対応付けにおいては，約 170,000 見出し語の中日対訳辞書を用いた．
[*4] 各 IBM モデルの繰返し回数は，デフォルトのものをそのまま用いた．
[*5] 日本語文の形態素列中の形態素を文頭から順に $V_1, V_2, \ldots, V_n$，中国語の形態素列中の形態素を文頭から順に $W_1, W_2, \ldots, W_m$ として，日本語句を $P_J (= V_p \cdots V_{p'})$ とし，中国語句を $P_C (= W_q \cdots W_{q'})$ とする．ここで，日中句の組 $\langle P_J, P_C \rangle$ が含まれる ある 1 つの対訳文対 $\langle S_J, S_C \rangle$ 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，「$p \le i \le p' \Leftrightarrow q \le j \le q'$」が成り立つ場合に，$P_J$ と P_C は対訳文対 $\langle S_J, S_C \rangle$ において「文単位の句対応制約」に違反しない，と定義する．
[*6] http://mecab.sourceforge.net
[*7] ただし，連続する数字とアルファベットは，それぞれ 1 個のトークンとして扱う．
[*8] ここでは，Moses を用いたデコーディングによって訳語を推定したのではなく，フレーズテーブル中の訳語を直接参照して訳語推定を行う．フレーズテーブルを用いた日中方向の訳語推定の精度は，「中国語側が形態素単位」の場合では 97.8%，「中国語側が文字単位の場合」では 95.9%である [13]．
[*9] ただし，本研究において，翻訳確率が最大となる訳語候補が複数存在する場合には，いずれの訳語候補も信頼性が低いと見なし，不採用とする．中日方向の場合も同様である．

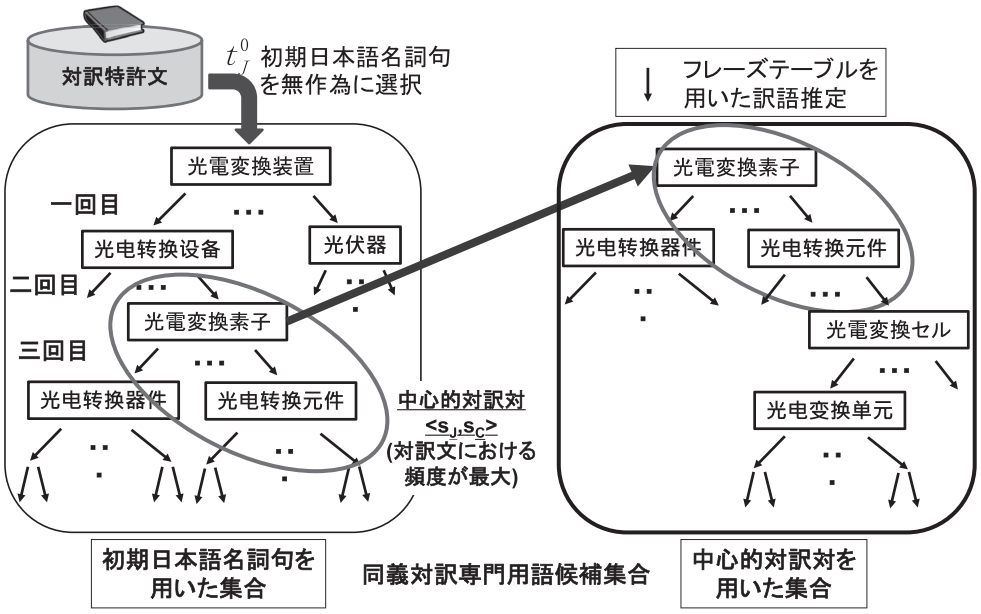


図 1 専門用語対訳対の訓練・評価用同義・異義集合の作成
Fig. 1 Developing a reference set of bilingual synonymous technical terms.

(1) t_C の訳語として，フレーズテーブルに存在する．
(2) (1) を満たす訳語であり，かつ，与えられた対訳文 $\langle S_J, S_C \rangle$ の日本語文 S_J に出現する．
(3) (1)，(2) を満たす訳語の中で，フレーズテーブルにおける翻訳確率 $P(t_J \mid t_C)$ が最大である．

4. 専門用語対訳対の訓練・評価用同義・異義集合の作成

4.1 作成手順

以下では，**図 1** に沿って，対訳特許文およびフレーズテーブルを用いて，専門用語対訳対の訓練・評価用同義・異義集合を作成する流れを示す．以下の手順においては，訓練・評価用の同義・異義専門用語対訳対の候補を生成するための種として，**中心的対訳対**と呼ぶ対訳対を選定し，この中心的対訳対を初期対訳対として，そこから訓練・評価用の同義・異義専門用語対訳対の候補を生成する．ただし，中心的対訳対を選定するにあたっては，無作為に抽出した初期日本語名詞句を用いることにより，まず，同義・異義専門用語対訳対の初期候補集合を生成し，この初期候補集合中の要素のうち一定の基準を満たす対訳対を中心的対訳対として選定するという方式を用いる[*10]．

(1) まず，日中対訳特許文から無作為に初期日本語名詞句 t_J^0 を抽出する．
(2) そして，初期日本語名詞句 t_J^0 に対して，本節末に示す「反復手続き：専門用語対訳対同義候補集合の生成」を適用することにより，図 1 左半分の過程を行い初期候補集合 $CBP(t_J^0)$ を生成する．この過程においては，対訳特許文，および，フレーズテーブルを用いるが[*11]，中国語側が形態素単位のフレーズテーブル，および中国語側が文字単位のフレーズテーブルをそれぞれ独立に用いて，初期候補集合 $CBP(t_J^0)$ を生成する．ここで，集合 $CBP(t_J^0)$ の要素数が m_0 以上である（すなわち，$|CBP(t_J^0)| \geq m_0$）ならば，以降の手続きを続ける．
(3) 次に，t_J^0 を初期日本語名詞句として作成した同義候補集合 $CBP(t_J^0)$ の要素の中から，4.2.2 項の手順に従い，**中心的対訳対** $s_{JC} = \langle s_J, s_C \rangle$ を選定する．ただし，4.2.2 項の手順においては，初期候補集合 $CBP(t_J^0)$ 中に「専門用語の対訳対」が存在しない場合は，その初期候補集合 $CBP(t_J^0)$ は，その時点で棄却される．
(4) 次に，中心的対訳対 $s_{JC} = \langle s_J, s_C \rangle$ の日本語用語 s_J に対して，本節末に示す「反復手続き：専門用語対訳対同義候補集合の生成」を適用することにより，図 1 の右半分の過程を行い，中国語側が形態素単位のフレーズテーブル，および中国側が文字単位のフレーズ

[*10] 本論文の主たる主張は，専門用語対訳対の同義候補集合を生成した後，分類器学習手法によって同義対訳専門用語を同定する手法を提案することである．本節においては，中心的対訳対を用いて同義候補集合を作成する手順を示しているが，これはあくまで経験的な知見を示したものに過ぎない．特に，専門用語対訳対の同義候補集合において十分な数の正例・負例を含むためには，中心的対訳対の候補集合を作成する段階において十分に大きく，かつ，一定以下の大きさに制限された集合を作成する必要があるという経験的知見を得ており，この知見に基づき，本論文で述べる手順を用いる．

[*11] この「反復手続き：専門用語対訳対同義候補集合の生成」においては，訳語推定回数を 6 回としている．これは，予備実験において，6 回の訳語推定において，同義対訳専門用語として獲得することが望ましい専門用語対訳対のうちの大半が生成され，しかも，同義関係にある専門用語対訳対以外の候補の生成数が最少となったためである．なお，これらの反復手続きにおいては，同一の日本語用語，および，中国語用語を検出することにより，それらの重複生成は回避されている．

テーブルをそれぞれ独立に用いて，専門用語対訳対の同義候補集合 $CBP(s_J)$ を生成する．

(5) 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（訓練・評価用同義対訳専門用語集合）$SBP(s_{JC})$，および，その他の対訳対の集合（訓練・評価用異義対訳専門用語集合）$NSBP(s_{JC})$ に分割する．

反復手続き：専門用語対訳対の同義候補集合の生成

ステップ 1 入力の日本語用語 t'_J に対して，日中対訳特許文の中から t'_J が出現する対訳文をすべて収集する．そして，収集した各対訳文に対して，フレーズテーブルを参照して 3.2 節で述べた手法を適用することにより，t'_J の中国語訳語を推定する[*12]．そして，推定された中国語訳語を t^i_C（$i = 1,\ldots,n_1$，ただし，n_1 は推定された中国語訳語の数）として，すべての対訳対 $\langle t'_J, t^i_C\rangle$ を収集し，専門用語対訳対の同義候補集合の初期集合 $CBP(t'_J)$ を作成する．

ステップ 2 同様に，すべての中国語用語 $t_C \in CBP(t'_J)$ に対して，対訳特許文の中から t_C を含む対訳文をすべて収集し，t_C の日本語訳語を推定する．そして，推定された日本語訳語を t^j_J（$j = 1,\ldots,n_2$，ただし，n_2 は推定された日本語訳語の数）として，すべての対訳対 $\langle t^j_J, t_C\rangle$ を $CBP(t'_J)$ に追加する．

ステップ 3 同様に，すべての日本語用語 $t_J \in CBP(t'_J)$ に対して，対訳特許文の中から t_J を含む対訳文をすべて収集し，t_J の中国語訳語を推定する．そして，推定された中国語訳語を t^k_C（$k = 1,\ldots,n_3$，ただし，n_3 は推定された中国訳語の数）として，すべての対訳対 $\langle t_J, t^k_C\rangle$ を $CBP(t'_J)$ に追加する．

ステップ 4 「ステップ 2」の処理を実行する．

ステップ 5 「ステップ 3」の処理を実行する．

ステップ 6 「ステップ 2」の処理を実行する．

ただし，「ステップ 1」から「ステップ 6」までのすべてのステップにおいて収集対象とする対訳対を選定する際には，4.2.1 項の手順に従う．

4.2 作成手順における詳細設定

本節では，本論文において，前節の作成手順に沿って実際に専門用語対訳対の訓練・評価用同義・異義集合の作成を行う際にどのような具体的な設定を用いたのかの詳細を説明する．

まず，情報源として用いた対訳特許文は，2 章で述べた 360 万対訳特許文である．また，前節の手順 (2) における初期候補集合 $CBP(t^0_J)$ の要素数の下限値 m_0 は 10 とする．その他，「反復手続き：専門用語対訳対の同義候補集合の生成」において，収集対象とする対訳対を選定する手順，および，手順 (3) において中心的対訳対を選定する手順の詳細は以下の各項のとおりである．

4.2.1 「反復手続き：専門用語対訳対の同義候補集合の生成」における収集対象対訳対の選定手順

本項では，前節の「反復手続き：専門用語対訳対の同義候補集合の生成」において，収集対象とする対訳対を選定する手順を述べる．

4.1 節の「反復手続き：専門用語対訳対の同義候補集合の生成」の「ステップ 1」から「ステップ 6」までのすべてのステップにおいて，「一般語の対訳対」をできるだけ除去し，「専門用語の対訳対」をできるだけ多く残すための予備調査を行ったところ，以下の条件をすべて満たす対訳対の 7 割以上が「専門用語の対訳対」であったのに対して，以下の条件のうち 1 つしか満たさない対訳対のうち「専門用語の対訳対」であるものは 2 割以下であった．また，その大半においては，対訳となる日中用語の一方のみの頻度が 12,500 以上となっており，「専門用語の対訳対」としても相対的な適切さの度合いが低い対訳対であった．以上の予備調査の結果をふまえて，「ステップ 1」から「ステップ 6」までのすべてのステップにおいて，以下の条件をすべて満たす対訳対 $\langle t_J, t_C\rangle$（ただし，t_J，t_C はそれぞれ日本語専門用語，および中国語専門用語）のみを収集対象として残し，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，または，長さの下限[*13]を満す．

(3) t_J，t_C いずれも語頭および語尾が機能語，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) $\langle t_J, t_C\rangle$ の頻度が 3,000 未満．

4.2.2 中心的対訳対の選定手順

本項では，4.1 節の手順 (3) において，t^0_J を初期日本語名詞句として作成した同義候補集合 $CBP(t^0_J)$ の要素の中から中心的対訳対を選定する手順を述べる．

本論文において，「一般語の対訳対」と「専門用語の対訳対」を区別するための予備調査を行ったところ，以下で述べる条件を 1 つも満たさない場合に「専門用語の対訳対」となる割合は十分に低いが，以下で述べる条件を少なくとも 1 つ満たす場合には一定の割合で「専門用語の対訳対」となることが分かった．そこで，以下で述べる条件を少なくとも 1 つ満たす場合に，その対訳対は「一般語の対訳対」

[*12] この手順を経ることにより，各対訳文からは，フレーズテーブルに含まれる対訳対のうち，誤りに相当する部分文字列等が削除され，対訳対として適切である可能性の高いものだけが得られる可能性が高くなる．

[*13] 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) 1 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．

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
Table 1 Numbers of bilingual technical terms: Candidates and reference of synonyms.

(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 $\bigcup_{s_J} CBP(s_J)$	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	
人手で同定した同義集合 $\bigcup_{s_{JC}} SBP(s_{JC})$	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 $\bigcup_{s_J} CBP(s_J)$	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	
人手で同定した同義集合 $\bigcup_{s_{JC}} SBP(s_{JC})$	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	

でないというヒューリスティクスを用いた．

(a) 日本語用語が以下のいずれかを満たす．
  (i) 漢字または平仮名を含む場合は，3 文字以上．
  (ii) カタカナ語の場合は，複合語である．

(b) 中国語用語が 4 文字以上，または形態素数が 3 以上．

そして，初期候補集合 $CBP(t_J^0)$ の要素のうち，このヒューリスティクスによって一般語の対訳対であると判定された対訳対を除去し，残った対訳対を対象として，360 万対訳文中の共起頻度が最大となる対訳対を人手で確認し，専門用語の対訳対として適切であると判定された場合は，その対訳対を「中心的対訳対」とする．それ以外の場合は，初期候補集合 $CBP(t_J^0)$ 中のすべての対訳対のうち，最も適切な対訳対を「中心的対訳対」とする．ただし，初期候補集合 $CBP(t_J^0)$ 中に「専門用語の対訳対」が存在しない場合は，その初期候補集合 $CBP(t_J^0)$ は，その時点で棄却する．

4.3 作成結果

4.1 節の手順に従い，専門用語対訳対の訓練・評価用同義・異義集合の作成を行った結果を以下に示す．まず，手順 (1) において無作為に選択した 4,000 個の初期日本語名詞句に対して，手順 (2) に従い初期候補集合 $CBP(t_J^0)$ を生成した結果，およそ 500 個の集合が要素数の下限を満たした．さらに，これらの初期候補集合に対して，手順 (3) に従って中心的対訳対の選定を行った結果，合計 114 個の中心的対訳対が選定された．この後，手順 (4) に従うことにより，114 個の専門用語対訳対同義候補集合が生成された．それらの同義候補集合における専門用語対訳対の総数および平均対訳対数を**表 1** に示す[*14]．最後に，手順 (5) に従って，中心的対訳対と同義となる対訳対の選定を行った結果，表 1 に示すように，114 個の専門用語対訳対の候補集合において，中心的対訳対と同義となる対訳対の総数および平均数は，「中国語側が形態素単位のフレーズテーブル」を用いた場合では，2,473 個および 21.7 個となり，「中国語側が文字単位のフレーズテーブル」を用いた場合では，2,318 個および 20.3 個となった．

5. 分類器学習を用いた同義対訳専門用語の同定

本章では，SVM を用いて同義対訳専用語を同定する手法について述べる．

5.1 適用手順

まず，114 個の専門用語対訳対同義候補集合 $CBP(s_J)$ の和集合を全事例集合 CBP とし，互いに素な部分集合 CBP_i（$i=1,\ldots,10$）に 10 分割する[*15]．本論文では，TinySVM[*16]を利用して，評価実験を行った．カーネル関数としては，一次多項式カーネルおよび二次多項式カーネルを評価し，大きな性能差が観測されなかったため，評価実験においては一次多項式カーネルを用いた．また，SVM の分離平面から評価事例までの距離を信頼度とし，正例（中心的対訳対と同義）判定において信頼度の下限を設定した．$CBP_1,\ldots,CBP_{10}$ の 10 個の部分集合のうち，8 個を訓練用事例集合として SVM の訓練を行い，残りのうちの 1 個を調整用事例集合とし，最後の 1 個を評価用事例集合とした．調整用事例集合を用いたパラメータの調整においては，分離平面から評価用事例までの距離の下限のパラ

[*14] 表 1 では，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) において「文字単位の集合と共通」となる対訳対数が，表 1 (b) において「形態素単位の集合と共通」となる対訳対数よりも多くなっている．

[*15] 各 CBP_i（$i=1,\ldots,10$）における正例（中心的対訳対と同義）・負例（中心的対訳対と異義）の数が，各 CBP_i（$i=1,\ldots,10$）の間で均等になるように，中心的対訳対の集合を分割した．

[*16] http://chasen.org/~taku/software/TinySVM/

表 2 専門用語対訳対の同義・異義同定のための素性（提案手法）
Table 2 Features for identifying bilingual synonymous technical terms (proposed method).

分類	素性名	定義（ただし，$X \in \{J,C\}$，$(Y,Z) \in \{(J,C),(C,J)\}$）
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： 中国訳語の順位	条件付き確率 $\mathrm{P}(t_C \mid t_J)$ 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 $\mathrm{P}(t_J \mid t_C)$ 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	$f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max(\lvert t_X \rvert, \lvert s_X \rvert)}$ ：ED は t_X と s_X の間の編集距離，$\lvert t \rvert$ は t に含まれる文字数を表す．
	f_{10}： バイグラム類似度	$f_{10}(t_X, s_X) = \frac{\lvert bigram(t_X) \cap bigram(s_X) \rvert}{\max(\lvert t_X \rvert, \lvert s_X \rvert) - 1}$ ：$bigram(t)$ 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	$f_{11}(t_J, s_J) = \frac{\lvert const(t_J) \cap const(s_J) \rvert}{\max(\lvert const(t_J) \rvert, \lvert const(s_J) \rvert)}$ ：$const(t)$ 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数の割合	$f_{12}(t_C, s_C) = \frac{\lvert const(t_C) \cap const(s_C) \rvert}{\max(\lvert const(t_C) \rvert, \lvert const(s_C) \rvert)}$ ：$const(t)$ 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	$f_{15}(t_X, s_X) = \frac{\lvert trans(t_X) \cap trans(s_X) \rvert}{\max(\lvert trans(t_X) \rvert, \lvert trans(s_X) \rvert)}$ ：$trans(t)$ 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 2 種類を区別して用いる）．

メータの調整を行った[*17]．以上の訓練，調整，評価の手順を 10 通り繰り返し，その評価結果のマイクロ平均を算出し，同義判定の性能評価を行った．

5.2 同義・異義判定のための素性

表 2 に示すように，同義対訳専門用語の同定に用いた素性は，大きく，対訳対 $\langle t_J, t_C \rangle$ の特性を規定するもの，および，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ものの 2 種類に分けられる．以下にその詳細を述べる．

[*17] SVM のソフトマージンを制約するパラメータについても調整を行い性能への影響を評価したが，SVM light（http://svmlight.joachims.org/）におけるデフォルト値（訓練事例の素性値の二乗和の平均値の逆数）と比べて大きな変化が観測されなかったため，SVM のソフトマージンを制約するパラメータはデフォルト値に固定して評価を行った．

5.2.1 対訳対の特性を規定する素性

対訳対の特性を規定する素性としては，対訳特許文における対訳対の共起頻度の素性（f_1），訳語の翻訳確率における順位の素性（f_2，f_3），用語の文字列長・単語長の素性（f_4，f_5），訳語推定において，中心的対訳対の日本語用語 s_J から t_C または t_J を生成するまでの繰返し訳語生成回数の素性（f_6）を用いる．

5.2.2 対訳対と中心的対訳対の間の関係を規定する素性

対訳対と中心的対訳対の間の関係を規定する素性としては，用語表記の同一性の素性（f_7，f_8），用語文字列の編集距離類似度の素性（f_9），用語文字列の 2 グラム類似度の素性（f_{10}），日本語用語間における同一形態素数の素性（f_{11}），中国語用語間における同一文字数の割合（f_{12}）日本語用語文字列包含関係・異表記の素性（f_{13}），中国語用語文字列包含関係の素性（f_{14}），フレーズテーブルから得られる訳語のうち共通なものの割合の素性（f_{15}），文字列

表 3 同義対訳専門用語同定の評価結果（%）
Table 3 Evaluation results of identifying bilingual synonymous technical terms (%).

(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

手法（素性・分離平面からの距離下限調整の基準）		適合率	再現率	F 値
ベースライン		71.4	40.0	51.3
SVM（全素性）	適合率最大	**86.5**	26.5	40.5
	F 値最大	64.3	64.1	**64.2**
SVM（適合率最大となる素性の組合せ：$f_{1\sim6} + f_{9\sim16}$）	適合率最大	**89.0**	23.9	37.7
SVM（文献 [15] の素性）	適合率最大	72.6	26.1	38.4
	F 値最大	71.0	54.7	61.5

(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

手法（素性・分離平面からの距離下限調整の基準）		適合率	再現率	F 値
ベースライン		74.0	40.1	52.0
SVM（全素性）	適合率最大	**89.0**	26.1	40.4
	F 値最大	63.5	65.3	**64.4**
SVM（適合率最大となる素性の組合せ：$f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$）	適合率最大	**90.4**	25.5	40.4
SVM（文献 [15] の素性）	適合率最大	74.4	36.7	49.2
	F 値最大	72.7	53.7	61.8

の非共有箇所のみに対してフレーズテーブルから得られる訳語のうち共通なものの割合の素性（f_{16}），フレーズテーブルにおいて s_J と t_C または s_C と t_J の間に訳語関係が存在するか否かの素性（f_{17}）を用いる[*18]．このうち，f_{15} および f_{16} は，フレーズテーブルにおいてどの程度の割合で共通の訳語を持つかという情報と，単言語において同義関係にある度合いとの間の相関に着目した素性であり，次節の評価結果において示すように，性能に大きな影響を持つ重要な素性である．

5.3 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 $\langle t_J, t_C \rangle$ は中心的対訳対 $\langle s_J, s_C \rangle$ と同義である」

という規則を用いた．まず，分離平面からの距離下限のパラメータに対して，同義判定の適合率を最大化する調整[*19]を行った．「中国語側が形態素単位」の場合，全素性を用いた場合（表 3「SVM（全素性）」欄）には 86.5%，適合率最大となる素性の組合せ（$f_{1\sim6} + f_{9\sim16}$）を用いた場合（表 3「SVM（適合率最大となる素性の組合せ）」欄）には 89.0%の適合率を達成した．一方，「中国語側が文字単位」の場合，全素性を用いた場合には 89.0%，適合率最大となる素性の組合せ（$f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$）を用いた場合には 90.4%の適合率を達成した．ただし，「中国語側が形態素単位」の場合，および，「中国側が文字単位」の場合，いずれにおいても，全素性を用いた場合と適合率最大となる素性の組合せを用いた場合との間で適合率の差には有意差（有意水準 5%）はない．次に，全素性を用いて，分離平面からの距離下限のパラメータに対して，同義判定の F 値を最大化する調整を行ったところ，「中国語側が形態素単位」の場合 64.2%の F 値を，「中国語側が文字単位」の場合 64.4%の F 値を，それぞれ達成した[*20]．

性能に大きな影響を持つ素性を同定するために，適合率最大の場合との間で有意差（有意水準 5%）のない適合率となる素性の組合せのうち，2 種類の素性（1 つの素性で日中二言語の情報を記述するもの，もしくは，同種類の情報を記述する日本語素性および中国語素性の 2 つの素性）からなる場合の性能を**表 4** に示す．この結果から，f_{15} および f_{16} のように，単言語の各専門用語またはその断片の間にフレーズテーブルにおける共通の訳語が存在するか否かを記述する素性が，重要な素性の 1 つであることが分かる．この f_{15} および f_{16} は，文献 [14] における素性の組合せを改善する形で新たに導入された「フレーズテーブルにおける共通訳の割合」の考え方に基づく素性であるが，本節の評価結果より，この新素性が性能に大きな影響を持つ重要な素性であることが示された．

また，比較対象として，日英同義対訳専門用語の同定を対象とした先行研究 [15] における素性と同等の素性の組合

表 4 「適合率最大の場合」との間で有意差（有意水準 5%）のない適合率となる 2 種類の素性情報の組とその評価結果（%）
Table 4 Pairs of features having no significant difference (5% significance level) with maximum precision features and their evaluation results (%).

(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

素性	適合率	再現率	F 値
f_{15}（日中）+ f_{16}（日中）	85.6	25.4	39.2
f_9（日中）+ f_{16}（日中）	86.8	24.9	38.7
f_{13}（日）+ f_{14}（中）+ f_{16}（日中）	86.8	24.8	38.6

(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

素性	適合率	再現率	F 値
f_9（日中）+ f_{15}（日中）	87.4	25.4	39.3

[*18] ただし，素性 f_9，f_{10}，f_{15}，f_{16} においては，それぞれ日本語側素性および中国語側素性の 2 種類の素性を用いる．
[*19] ただし，再現率が 25%以上となるという条件のもとで，パラメータの調整を行った．
[*20] その他，「中国語側が形態素単位」と「中国語側が文字単位」の間で判定結果の AND 条件をとった場合の適合率の評価も行ったが，「中国語側が形態素単位」単独，および，「中国語側が文字単位」単独の場合の適合率を有意に改善することはできなかった．

表 5 専門用語対訳対の同義・異義同定のための素性（文献 [15] の手法）
Table 5 Features for identifying bilingual synonymous technical terms (previous method [15]).

分類	素性名	定義
基本素性	h_{1J}，h_{1C}： 第 1 文字の一致	日中各単言語において中心的対訳対の専門用語との間で第 1 文字が一致するか否か．
	h_{2J}，h_{2C}： 編集距離類似度	f_9 と同じ．
	h_{3J}，h_{3C}： バイグラム類似度	f_{10} と同じ．
	h_{4J}，h_{4C}： 部分文字列の一致数	日中各単言語において中心的対訳対の専門用語との間で部分文字列が一致する回数（文献 [15] では，中心的対訳対の専門用語の部分文字列との間で既知の同義関係が成り立つ回数もあわせて数えているが，本論文では，利用可能な既知の同義関係の情報がないため，部分文字列の一致回数のみ素性として用いる）．
	h_{5J}，h_{5C}： フレーズテーブルの訳語関係が存在	f_{17} と同じ（文献 [15] では，フレーズテーブルの代りに訓練用の対訳辞書を用いている）．
	h_6： 中国語用語の文字列の包含関係	f_{14} と同じ（文献 [15] では，英語における頭字語を扱うための素性として用いているが，本論文では，中国語における「的」のための素性に置き換える）．
	h_7： 日本語用語の文字列の包含関係もしくは異表記	f_{13} と同じ（文献 [15] では，片仮名語の異表記のみを扱うための素性として用いているが，本論文では，片仮名語の長音「ー」のほか，接尾辞および送り仮名の異表記を扱うための素性に置き換える）．
複合素性	$h_{1J} \wedge h_{1C}$	—
	$\sqrt{h_{2J} \cdot h_{2C}}$	—
	$\sqrt{h_{3J} \cdot h_{3C}}$	—
	$h_{5J} \wedge h_{5C}$	—
	$h_6 \cdot h_{2J}$	—
	$h_7 \cdot h_{2C}$	—

せを**表 5** のように設計し，5.1 節に示した訓練，調整，評価の手順をそのまま適用して性能評価を行った結果を表 3「SVM（文献 [15] の素性）」欄に示す．この結果から，提案手法によって，先行研究 [15] における素性と同等の素性の組合せの性能を大幅に改善することが分かる．

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 6** に示す．

表 6 (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 6 (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 7** に示す．

表 7 (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 7 (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 $\langle t_J, s_C\rangle$，$\langle s_J, t_C\rangle$ 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**” は実際は同義関係にあるにもかかわらず，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．

表 6 同義判定における SVM による改善例
Table 6 Examples of improvement in identifying bilingual synonymous technical terms by SVM.

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 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	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 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	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 7 同義判定における提案手法の誤り例
Table 7 Examples of errors in identifying bilingual synonymous technical terms by the proposed method.

(a) 提案手法により同義と判定し不正解

中心的対訳対 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	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 $\langle s_J, s_C \rangle$	専門用語対訳対 $\langle t_J, t_C \rangle$	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

6. 関連研究

テキストから二言語対訳辞書を獲得する一連の研究の中で，文献 [15] においては，専門用語対訳対の同義判定手法を提案しており，また，手法として分類器学習を適用している．したがって，手法の点においても，また，分類器学習で用いている素性の点においても，本論文の手法と密接に関連している．しかし，文献 [15] においては，同義判定の対象とする専門用語対訳対の収集を手動で行っており，手法の適用範囲が限定される点が短所である．一方，本論文の手法においては，毎年公開される対訳特許文書を情報源として，同義判定の対象とする専門用語対訳対を収集しており，中心的対訳対の選定過程を除けば，その他の全過程がほぼ自動化されている．したがって，本論文の手法においては，文献 [15] と比較した場合の重要な長所として，手法の適用範囲を大幅に拡大できている点をあげることができる．また，前節で示したように，分類器学習において用いる素性の性能比較の点においても，提案手法の素性によって，文献 [15] で用いられた素性の性能を大幅に改善することが実現できている．

その他，本論文で対象とした同義関係の同定に関連して，いい換え知識や翻訳知識を獲得するとともに，それらの知識を利用することにより統計的機械翻訳の翻訳性能やカバレージが改善できることが報告されている [7], [19], [20]．具体的には，訳語が併記されたウェブ文書を情報源として獲得された訳語対を利用することにより統計的機械翻訳の翻訳性能を改善した事例 [7]，および，対訳コーパスを情報源として獲得したいい換え知識を利用することにより，統計的機械翻訳の翻訳性能およびカバレージを改善した事例 [19], [20] が報告されている．これらの成果をふまえると，本論文において同定された同義関係についても，それらを有効に利用することにより，統計的機械翻訳の翻訳性能の改善が期待できる．

7. おわりに

本論文では，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 [14] 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．

参考文献

[1] 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).

[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] Bouamor, D., Semmar, N. and Zweigenbaum, P.: Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora, *Proc. 51st ACL*, pp.759–764 (2013).

[4] Morin, E. and Hazem, A.: Looking at Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction, *Proc. 52nd ACL*, pp.1284–1293 (2014).

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

[7] Lin, D., Zhao, S., Van Durme, B. and Paşca, M.: Mining Parenthetical Translations from the Web by Word Alignment, *Proc. 46th ACL: HLT*, pp.994–1002 (2008).

[8] Erdmann, M., Nakayama, K., Hara, T. and Nishio, S.: Improving the Extraction of Bilingual Terminology from Wikipedia, *ACM Trans. Multimedia Computing, Communications and Applications*, Vol.5, No.4, pp.31:1–31:17 (2009).

[9] 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).

[10] 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).

[11] Koehn, P. et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th ACL, Companion Volume*, pp.177–180 (2007).

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

[13] 董　麗娟，龍　梓，豊田樹生，宇津呂武仁，三橋朋晴，山本幹雄：日中パテントファミリーから抽出した対訳文を用いた専門用語の訳語推定，言語処理学会第 20 回年次大会発表論文集，pp.368–371 (2014).

[14] Liang, B., Utsuro, T. and Yamamoto, M.: Semi-Automatic Identification of Bilingual Synonymous Technical Terms from Phrase Tables and Parallel Patent Sentences, *Proc. 25th PACLIC*, pp.196–205 (2011).

[15] Tsunakawa, T. and Tsujii, J.: Bilingual Synonym Identification with Spelling Variations, *Proc. 3rd IJCNLP*, pp.457–464 (2008).

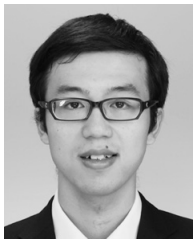
[16] Utiyama, M. and Isahara, H.: A Japanese-English Patent Parallel Corpus, *Proc. MT Summit XI*, pp.475–482 (2007).

[17] Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).

[18] Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C.: A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005, *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pp.168–171 (2005).

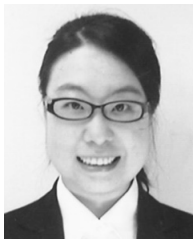
[19] Callison-Burch, C., Koehn, P. and Osborne, M.: Improved Statistical Machine Translation Using Paraphrases, *Proc. HLT-NAACL*, pp.17–24 (2006).

[20] He, W., Wu, H., Wang, H. and Liu, T.: Improve SMT Quality with Automatically Extracted Paraphrase Rules, *Proc. 50th ACL*, pp.979–987 (2012).



龍 梓　（学生会員）

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．現在，筑波大学大学院システム情報工学研究科博士前期課程在学中．機械翻訳の研究に従事．



董 麗娟

2012 年大連理工大学ソフトウェア工学学院卒業．現在，筑波大学大学院システム情報工学研究科博士前期課程在学中．機械翻訳の研究に従事．

宇津呂 武仁　（正会員）

1994 年京都大学大学院工学研究科博士課程修了．博士（工学）．1994 年奈良先端科学技術大学院大学情報科学研究科助手，2000 年豊橋技術科学大学情報工学系講師，2003 年京都大学情報学研究科講師．2006 年筑波大学大学院システム情報工学研究科助教授，准教授を経て 2012 年筑波大学システム情報系教授，現在に至る．自然言語処理，ウェブマイニングの研究に従事．電子情報通信学会，言語処理学会，人工知能学会，ACL 各会員．

三橋 朋晴

1988 年より財団法人日本特許情報機構職員．特許情報検索システムの開発に従事．2007 年より同財団内特許情報研究所にて機械翻訳用辞書の研究開発に従事．

山本 幹雄 （正会員）

1986 年豊橋技術科学大学大学院修士課程修了．同年（株）沖テクノシステムズラボラトリ研究開発員．1988 年豊橋技術科学大学情報工学系教務職員．1991 年同助手．1995 年筑波大学電子・情報工学系講師．1998 年同助教授．2008 年筑波大学大学院システム情報工学研究科教授．博士（工学）．自然言語処理の研究に従事．電子情報通信学会，言語処理学会，人工知能学会，ACL 各会員．