日中パテントファミリーを用いた同義対訳専門用語同定手法の評価*

龍 梓† 董 麗娟† 宇津呂 武仁† 三橋 朋晴‡ 山本 幹雄†
筑波大学大学院 システム情報工学研究科† 日本特許情報機構‡

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 [1,5,9] では，パテントファミリーを情報源として，対訳特許文から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，ある専門用語の訳語を獲得することはできるが，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，先行研究 [3] では，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，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) [8] を適用することにより，日中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，本論文では特に，文献 [3] における素性の組み合わせを改善し，評価実験を通して，最適な性能を達成する素性の組み合わせを示した．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 [6] 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わ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．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，文献 [7] の手法によって日中間で文を対応付け，スコア降順で上位の 360 万文対を抽出したものである．

3 分類器学習を用いた同義対訳専門用語の同定

本論文では，文献 [4] と同様の手法を用いて，表 1 に示す専門用語対訳対同義候補集合[1]を生成し，SVM を用いて同義対訳専門用語の同定を行った．同義対訳専門用語の同定に用いた素性としては，表 2 に示すように，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[2]．このうち，文献 [3] における素性の組み合わせを改善する形で新たに導入された 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．

4 評価

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[3] 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

*Evaluating a Method of Identifying Bilingual Synonymous Technical Terms using Japanese-Chinese Patent Families

†Zi Long, Lijuan Dong, Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Tomoharu Mitsuhashi, Japan Patent Information Organization (JAPIO)

[1]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．

[2]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二種類の素性を用いる．

[3]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．

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	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	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	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	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 3: 同義対訳専門用語同定の評価結果 (%)

(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わせ: $f_{1\sim6} + f_{9\sim16}$)	適合率最大	**89.0**	23.9	37.7
SVM (文献 [6] 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わせ: $f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$)	適合率最大	**90.4**	25.5	40.4
SVM (文献 [6] の素性)	適合率最大	74.4	36.7	49.2
	F 値最大	72.7	53.7	61.8

「SVM(全素性)」欄) には 86.5%，適合率最大となる素性の組合わせ ($f_{1\sim6} + f_{9\sim16}$) を用いた場合 (表 3「SVM(適合率最大となる素性の組合わせ)」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わせ ($f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$) 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み合わ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[4].

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 (%)

(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

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わせのうち，二種類の素性 (一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二つの素性) 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，重要な素性の一つであることが分かる．この f_{15} および f_{16} は，文献 [2] における素性の組合わ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 [6] における素性と同等の素性の組合わせを設計[5] し，同様な訓練，調整，評価の手順を適用して性能評価を行った結果を表 3「SVM(文献 [6] の素性)」欄に示す．この結果から，提案手法に

[4]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．

[5]素性の具体的な説明は文献 [4] に参照する．

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，$X \in \{J, C\}$，$(Y, Z) \in \{(J, C), (C, J)\}$)
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	$f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max(\lvert t_X \rvert, \lvert s_X \rvert)}$ ：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二種類を区別して用いる).

よって，先行研究 [6] における素性と同等の素性の組合わ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 5 に示す.

表 5 (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 5 (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 6 に示す.

表 6(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 6(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}:

表 5: 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 $\langle s_J, s_C \rangle$	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 6: 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 $\langle s_J, s_C \rangle$	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

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．

5 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 [2] 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) より提供して頂いっ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る．

参考文献

[1] 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.

[2] 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.

[3] 龍梓, 董麗娟, 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同義対訳専門用語の同定. 言語処理学会第 20 回年次大会発表論文集, pp. 955–958, 2014.

[4] 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, 2015.

[5] 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.

[6] T. Tsunakawa and J. Tsujii. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pp. 457–464, 2008.

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

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

[9] K. Yasuda and E. Sumita. 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.