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*

龍 梓† 董 麗娟† 豊田 樹生† 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 [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 [1,9] 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 [3] 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) [8] 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，本論文では，文献 [3] 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ことを目的とする．

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 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 [1] 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．なお，学習に用いられた対訳文は，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 [6] 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 [5][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 2 つの対訳文に対して，独立に Moses [2] 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．

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．

*Identifying Bilingual Synonymous Technical Terms from Parallel Sentences extracted from Japanese-Chinese Patent Families

†Zi Long, Lijuan Dong, Itsuki Toyota, Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba

‡Tomoharu Mitsuhashi, Japan Patent Information Organization (JAPIO)

[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．

[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．

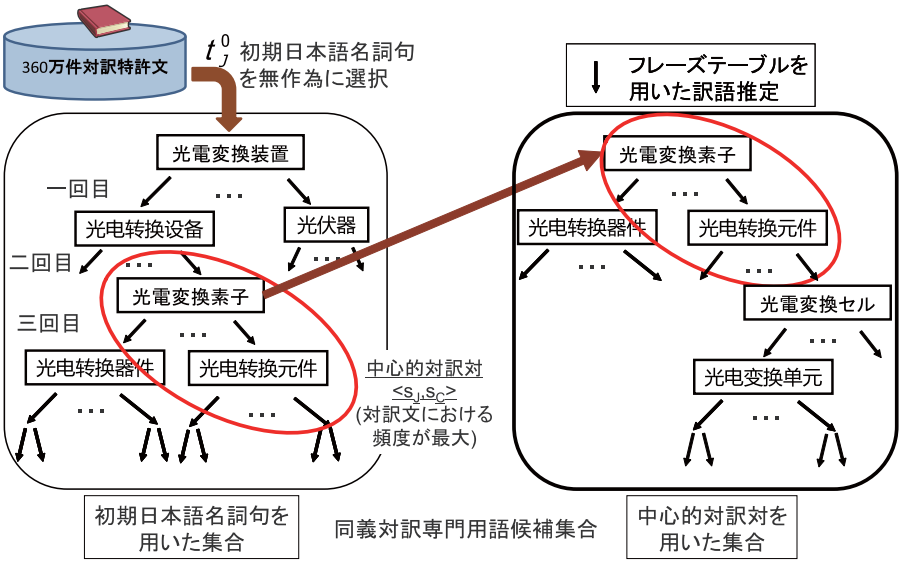


図 1: 専門用語対訳対同義候補集合の作成

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 $\bigcup_{s_J} CBP(s_J)$	形態素単位の集合のみが含む	14,161	28,948	124.2	253.9
	文字単位の集合と共通	14,787		129.7	
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が含む	180	2,604	1.6	22.8
	文字単位の集合と共通	2,424		21.3	

(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が含む	8,816	22,563	77.3	197.92
	形態素単位の集合と共通	13,747		120.6	
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が含む	309	2,496	2.7	21.9
	形態素単位の集合と共通	2,187		19.2	

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，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. $\langle t_J, t_C \rangle$ の頻度が 3,000 未満.

本論文では，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，本論文では，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，$|\, CBP(t_J^0)\, | \geq 10$).

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** $s_{JC} = \langle s_J, s_C \rangle$ とする[4]．ここで，本論文では，対訳対が以下の条件を全て満たす

[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] 本論文では，文献 [3] 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 2: 専門用語対訳対の同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 $\langle t_J, t_C \rangle$ の特性を規定する	f_1: 出現頻度	対訳特許文における $\langle t_J, t_C \rangle$ の出現頻度の二進対数.
	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同一訳の割合	$f_{15}(t_X, s_X) = \frac{\lvert trans(t_X) \cap trans(s_X) \rvert}{\max(\lvert trans(t_X) \rvert, \lvert trans(s_X) \rvert)}$ ：$trans(t)$ 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}: フレーズテーブルの共通訳が存在	フレーズテーブルにより，t_Y を訳語推定し s_Z が得られる．または s_Z を訳語推定し t_Y が得られる.

場合に，その対訳対は「一般語の対訳対」であるというヒューリスティクスを用いた.

1. 360 万対訳文における頻度が 3,000 以上.
2. 日本語用語が以下のいずれかを満たす.
   (a) 漢字または平仮名を含む場合は，二文字以下.
   (b) カタカナ語の場合は，複合語でない.
3. 中国語用語が 3 文字以下，または形態素数が 2 以下.

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.

表 1 では，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) における共通対訳対数の方が表 1(b) よりも多くなっている.

5 機械学習を用いた同義判定

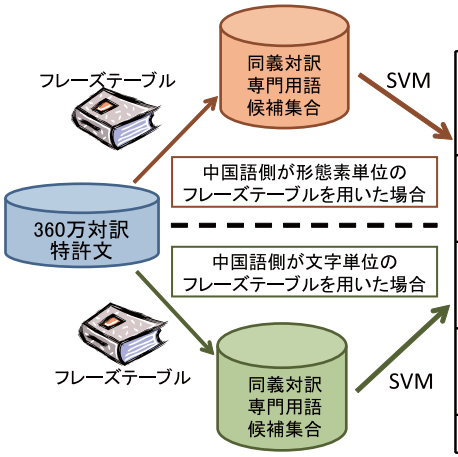
5.1 適用手順

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，文献 [3] における交差検定手順により 2 種類のパラメータ (SVM のソフトマージンを制約するパラメータ，および，分離平面から評価用事例までの距離の下限) に対して，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．さらに,「「中国語文を形態素単位に分割」,「中国語文を文字単位に分割」の両方が一致して同義と判定する場合のみ同義と判定する」という判定手法を導入し，パラメータの調整において，同義判定の適合率を最大化[5] する調整を行った.

5.2 同義・異義判定のための素性

同義対訳専門用語の同定に用いた素性を表 2 に示す．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,

[5]ただし,「再現率が 25%以上」という条件を付けて，パラメータの調整を行った.



対訳対 <t_J,t_C>	中心的対訳対 <s_J,s_C>	SVMの同義・異義判定		「形態素単位∩文字単位」による同義・異義判定	人手による同義・異義判定
		「形態素単位」での同義・異義判定	「文字単位」での同義・異義判定		
<印刷/回路/基板, 印刷/电路/基板 or 印/刷/电/路/基/板>	<プリント/回路/基板, 印刷/电路/板 or 印/刷/电/路/板>	同義	同義	同義	同義
<電磁/駆動/装置, 电池/驱动/装置 or 电/池/驱/动/装/置>	<電磁/駆動/装置, 电磁/驱动/装置 or 电/磁/驱/动/装/置>	異義	同義	異義	異義
<流体/導管, 流体/传导 or 流/体/传/导>	<流体/導管, 流体/管道 or 流/体/管/道>	同義	異義	異義	異義
… …					

図 2: 「中国語側が形態素単位」および「中国語側が文字単位」のフレーズテーブルを用いた同義・異義判定の例

表 3: 同義対訳専門用語同定の評価結果 (%)

手法			適合率	再現率	F 値
形態素単位	ベースライン		69.1	40.0	50.7
	SVM	適合率最大	**84.3**	24.5	38.0
		F 値最大	68.6	54.4	**60.7**
文字単位	ベースライン		71.5	39.4	50.8
	SVM	適合率最大	**86.6**	25.4	39.3
		F 値最大	70.0	53.3	**60.6**
形態素単位 ∩ 文字単位	ベースライン		77.3	33.1	46.3
	適合率最大		**90.0**	25.1	39.2

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．

5.3 評価結果

同義対訳専門用語同定の評価結果を表 3 に，判定結果の例を図 2 に，それぞれ示す．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」という条件を用いた．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 80.1%の適合率を達成し，「文字単位」では 86.6%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」，「形態素単位」とも，ベースラインを上回る F 値を達成した．「形態素単位」，「文字単位」の両方が一致して同義と判定する場合のみ同義と出力する「形態素単位 ∩ 文字単位」の手法では，適合率を最大化する調整を行うことにより，「形態素単位」・「文字単位」の単独判定の評価結果を上回る適合率を達成した．

6 おわりに

本論文では，対訳特許文を用いて，日中同義対訳専門用語の同定と収集を行う手法を提案した．特に，中国語文に対して，形態素単位と文字単位の 2 通りに分割を行い，SVM によってそれぞれ独立に同義・異義判定を行った後，同義判定結果の一致する同義関係の評価を行い，適合率 90%，再現率 25%を達成した．今後は，中国語側において，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を併用し，単一の SVM によって同義・異義判定を行うことにより，適合率と再現率の両方を改善する方式について研究を進める．

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, 2014.

[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, 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] J. Sun and Y. Lepage. 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.

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

[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.