対訳特許文を用いた同義対訳専門用語の同定と収集*

梁 冰†　　宇津呂 武仁†　　山本 幹雄†
筑波大学大学院 システム情報工学研究科†

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，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 [3] では，NTCIR-7 の特許翻訳タスク [1] で配布された日英 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 [2] 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) [7] による機械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を改善している．

この手法の問題点の一つとして，ある日本語専門用語に対する英訳語を推定する際に，その日本語専門用語が出現する一つの対訳文に出現する英訳語のみを推定対象とする点が挙げられる．そのため，この手法では，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，同義・異義となる専門用語対訳対の集合を同定することができない．そこで本論文では，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英訳語の推定を行うことにより，同義となる専門用語対訳対を同定することを目的とする．

本論文において提案する手順においては，まず，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，同義となる専門用語対訳対の候補を生成する．生成した候補集合の中から同義判定する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 を適用することにより，同義集合・異義集合を同定する．評価実験においては，同義判定において 97.4% の適合率を達成した．

*Identifying and Collecting Bilingual Synonymous Technical Terms from Parallel Patent Sentences

†Bing Liang, Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba

2　日英対訳特許文

本論文では，NTCIR-7 の特許翻訳タスク [1] で配布された約 180 万対の日英文対応データを，フレーズテーブルの訓練用データとして使用した．この文対応データは，1993-2000 年発行の日本公開特許広報全文と米国特許全文を対象として，文献 [6] によって日英間で文対応を付けたものである．

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

本論文では，文献 [3] の場合と同様に，専門用語の訳語推定において，日英対訳特許文から学習したフレーズテーブルを用いる．フレーズテーブルにおいては，2 節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 [2] を適用することにより，日英の句の組，及び，日英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，共起単語表の作成などを行う．
2. 文対応データから最尤な単語対応を英日，日英の両方向において得る．
3. 英日，日英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，可能な全ての日英の句の組を作成し，各組に対して，「文単位の句対応制約」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みなす)．
5. 文対応データにおける日英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．

手順 4 について，以下に「文単位の句対応制約」の条件を示す．

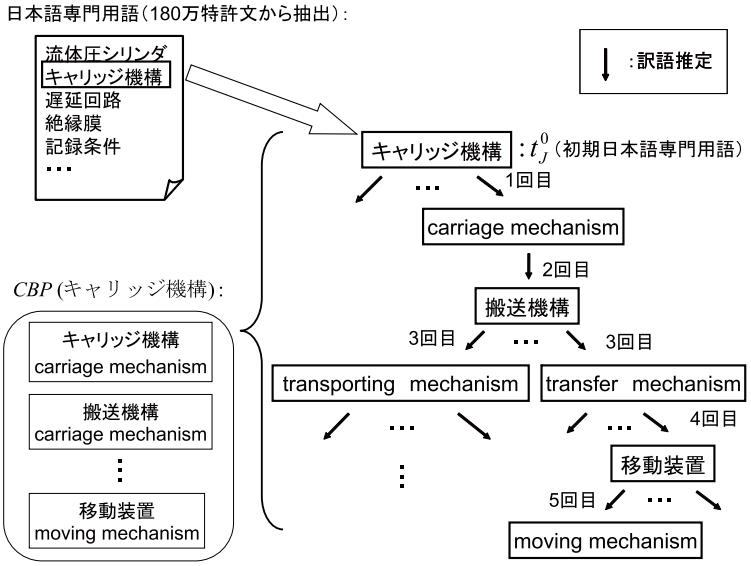

図 1: 専門用語対訳対同義候補集合の作成

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

	総要素数	134 個の集合の間の平均対数
同義候補集合 $\bigcup_{s_J} CBP(s_J)$	22,473	167.7
人手で同定した同義集合 $\bigcup_{s_{JE}} SBP(s_{JE})$	1,680	12.5

日本語文の形態素列中の形態素を文頭から順に $V_1, V_2, \ldots, V_n$，英文の単語列中の単語を文頭から順に $W_1, W_2, \ldots, W_m$ として，日本語句を $P_J(= V_p \cdots V_{p'})$ とし，英語句を $P_E(= W_q \cdots W_{q'})$ とする．ここで，日英句の組 $\langle P_J, P_E \rangle$ が含まれるある一つの対訳文対 $\langle T_J, T_E \rangle$ 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，「$p \leq i \leq p' \Leftrightarrow q \leq j \leq q'$ 」が成り立つ場合に，P_J と P_E は対訳文対 $\langle T_J, T_E \rangle$ において「文単位の句対応制約」に違反しない，と定義する．

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. 180 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 180 万件から学習されたフレーズテーブル[1] を用いて訳語推定を行い，英語訳語を得る．
2. 1 で得られた英語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)．

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，$|\, CBP(t_J^0)\, | \geq 10$)．

[1]ただし，日英方向の訳語推定を行う場合は，日英方向のフレーズテーブルの順位が一位となる英訳語を用い，英日方向の訳語推定を行う場合は，英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，対訳特許文 180 万件中の頻度が 6 以上 800 以下となる対訳対に限定した．なお，フレーズテーブルを用いた日英方向の訳語推定の精度は，91.9%である [3].

表 2: 専門用語対訳対の同義・異義集合同定のための素性

分類	素性名	定義 (ただし，$X \in \{J, E\}$，$(Y, Z) \in \{(J, E), (E, J)\}$)
対訳対 $\langle t_J, t_E \rangle$ の特性を規定する	f_1: 出現頻度	対訳特許文における $\langle t_J, t_E \rangle$ の出現頻度の自然対数.
	f_2: 英語訳語の順位	条件付き確率 $\mathrm{P}(t_E \mid t_J)$ の降順に t_E を順位付けしたときの t_E の順位.
	f_3: 日本語訳語の順位	条件付き確率 $\mathrm{P}(t_J \mid t_E)$ の降順に t_J を順位付けしたときの t_J の順位.
	f_4: 日本語文字数	t_J の文字数.
	f_5: 英語単語数	t_E の単語数.
	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 $\langle t_J, t_E \rangle$ と中心的対訳対 $\langle s_J, s_E \rangle$ 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: 英語用語が同一	$t_E = s_E$ 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}: 同一の形態素・単語数の割合	$f_{11}(t_X, s_X) = \frac{\lvert const(t_X) \cap const(s_X) \rvert}{\max(\lvert const(t_X) \rvert, \lvert const(s_X) \rvert)}$ ：$const(t)$ は t に含まれる形態素または単語の集合.
	f_{12}: 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_{13}: 英語語幹が同一	t_E と s_E の構成単語数が同一，かつ，対応する位置の単語の語幹が同一となる.
	f_{14}: 英語用語のハイフン・スペース	t_E と s_E は，ハイフンまたはスペースの有無のみが異なる.
	f_{15}: 非共有箇所に対し要素合成法の同一訳が存在	t_X, s_X で文字列が一致しない箇所 x_i, x_j に対して，要素合成法による訳語推定を行った場合に，同一訳が存在する.
	f_{16}: 要素合成法の共通訳が存在	要素合成法により，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.
	f_{17}: 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.

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，180 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** $s_{JE} = \langle s_J, s_E \rangle$ とする[2]．ここで，本論文では，対訳対が以下の条件を全て満たす場合に，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. 180 万対訳文における頻度が 500 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. 英語用語が一単語.

以上の手順に従って，合計 150 個の中心的対訳対を選定した．次に，中心的対訳対 s_{JE} 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 $CBP(s_J)$ を作成する．ただし，以下では，要素数の下限 (具体的には，$\mid CBP(s_J) \mid \geq 10$) を満たすもののみを対象とする．

以上の手順の結果，134 個の専門用語対訳対同義候補集合が作成された．表 1 に示すように，専門用語対訳対の総数は，22,473 個となった，なお，この過程において，訳語対応として正しくない対訳対は人手で除外した．

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_{JE} と同義となる対訳対の集合 $SBP(s_{JE})$，および，その他の対訳対の集合 $NSBP(s_{JE})$ に分割する．この結果，表 1 に示すように，中心的対訳対と同義となる対訳対の総数は 1,680 個となった．

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 $\langle t_J, t_E \rangle$ の特性を規定するもの，および，対訳対 $\langle t_J, t_E \rangle$ と中心的対訳対 $\langle s_J, s_E \rangle$ の間の関係を規定するものの 2 種類に分けられる．

[2]我々が行った先行研究 [4] においては，専門用語対訳対同義候補集合中の全ての同義組および異義組を同定するタスクを設定した．一方，本論文では，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点が異なる．本論文においては，要素技術の性能の限界を解明することを主目的として，問題の本質を特定するためのタスク設定を採用した．

表 3: 同義判定の性能評価 (%)

手法		適合率	再現率	F 値
ベースライン		67.0	54.3	60.8
SVM	適合率最大	**97.4**	31.0	44.9
	F 値最大	73.3	62.9	**65.7**

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_E と s_E が同一
SVM:	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

$\langle t_J, t_E \rangle - \langle s_J, s_E \rangle$	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
〈保持回路，holding circuit〉− 〈ホールド回路，hold circuit〉	同義	異義	同義

(b) ベースラインのみで同義と判定し不正解

$\langle t_J, t_E \rangle - \langle s_J, s_E \rangle$	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
〈配電器，distributor〉− 〈分配器，distributor〉	異義	同義	異義

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，SVM による同義・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，本論文で調整の対象としたパラメータは，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_E と s_E 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，97.4%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は，ベースラインの F 値を上回る F 値を達成した．

表 4 に，SVM による同義判定の改善例を示す．「(a) SVM のみで同義と判定し正解」の例においては，「英語語幹が同一」，「非共有箇所に対し要素合成法の同一訳が存在」，「フレーズテーブルの共通訳が存在」等の素性の効果によって，SVM のみが同義と判定できたと考えられる．一方，「(b) ベースラインのみで同義と判定し不正解」の例においては，「日本語訳語の順位」や「対訳特許文における出現頻度の自然対数」等の素性の効果によって，SVM のみが異義と判定できたと考えられる．

7 おわりに

本論文では，対訳特許文を用いて，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，同義・異義判定のための素性を用いて，SVM によって同義・異義判定を行う手法を提案した．評価実験においては，97%以上の適合率を達成した．今後は，SVM によって高適合率で同義と判定した専門用語対訳対を自動的に中心的対訳対として選定し，専門用語対訳対同義候補集合の作成と SVM の適用を再帰的に行う枠組みを開発し，再現率を改善する方式について研究を進める．

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] 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.

[4] 森下洋平, 宇津呂武仁, 山本幹雄. 対訳特許文からの対訳専門用語獲得における同義専門用語集合の分析と同定. 言語処理学会第 16 回年次大会論文集, pp. 206–209, 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] 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.