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*

森下 洋平† 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†，

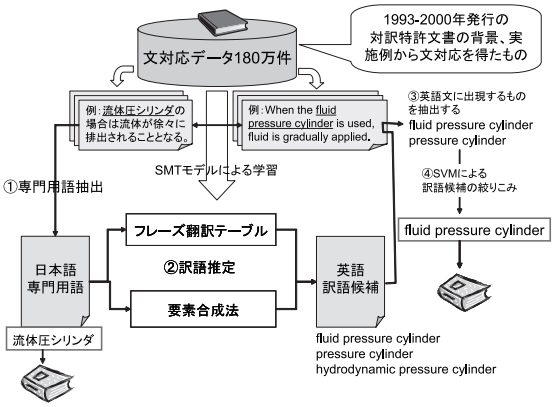

図 1: 複数の訳語推定手法を併用した対訳文からの対訳専門用語獲得の流れ

1 はじめに

[森下 08] では，対訳特許文からの専門用語対訳対獲得を目的として，NTCIR-7 の特許翻訳タスク [Fujii08] で配布された日英 180 万件の対訳特許文を用いて評価を行った．これらの手法では，フレーズベース統計的機械翻訳モデル [Koehn07]，要素合成法 [外池 07]，Support Vector Machines [Vapnik98] (SVM) 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の適合率を改善させている (図 1)．また，これらの手法は対訳特許文から抽出した日本語専門用語に対し訳語候補を作成し，トークン単位の評価を行う．そのため，ある専門用語対訳対を獲得する際に 1 対訳文しか考慮していない．しかし，実際は多くの専門用語は対訳特許文中に複数出現しているため，異なる訳が存在する．そのため，複数の文で獲得された専門用語対訳対に対し，専門用語間の同義，異義関係を見極めることが辞書作成に必要不可欠である．

そこで，本論文では対訳特許文から獲得された専門用語対訳対を用いて，同義専門用語集合の分析と同定を行うことを目的とする．評価実験においては，対訳特許文

*Analyzing and Identifying Sets of Synonymous Technical Terms in Acquisition of Bilingual Technical Term Lexicon from Parallel Patent Sentences

†Yohei Morishita, Takehito Utsuro, Mikio Yamamoto,Graduate School of Systems and Information Engineering,University of Tsukuba

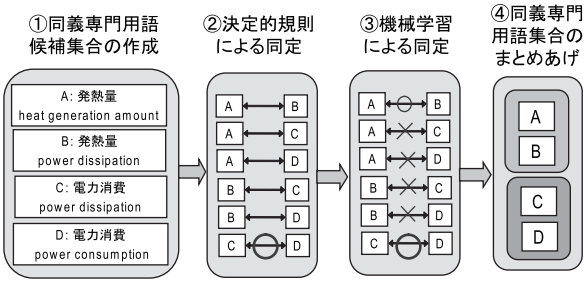

図 2: 同義専門用語集合同定の流れ

から得られた対訳対から同義専門用語候補集合を作成し，それらに対し同義集合を確実に同定するための，決定的規則を適用する．決定的規則により同義専門用語集合の 6%を同定した．さらに，決定的規則によって同定できない専門用語に対し，SVM を用いた同定を行った．その結果，決定的規則により同定できない同義専門用語集合に対して，適合率 93.2%, 再現率 23.1%で同定を行うことができた．決定的規則や SVM の素性として，文字列の類似度や共起語の類似度を用いた．

2 同義専門用語集合同定の流れ

図 2 に，同義専門用語集合同定の流れを示す．

(1) 180 万件から得られた対訳対の中から，同義専門用語候補集合を作成する．

(2) 同義専門用語候補集合に含まれる日英対訳対の全組み合わせを作成する．全組み合わせ中，15.6%が同義専門用語組，84.4%が異義専門用語組となった．それらに対し，同義集合を確実に選定するための決定的規則により同義専門用語および異義専門用語を同定する．決定的規則により，全組み合わせ中 0.9%の同義専門用語，2.2%の異議専門用語に対し，同定を行った．

(3) (2) で同定できない 96.9%の組み合わせに対して，機械学習による同義，異義の同定を行う．同義専門用語集合に対し，SVM による同義判定の結果，適合率 93.2%, 再現率 23.1%で同定を行った．また，異義判定の結果，適合率 93.7%, 再現率 74.9%で同定を行った．

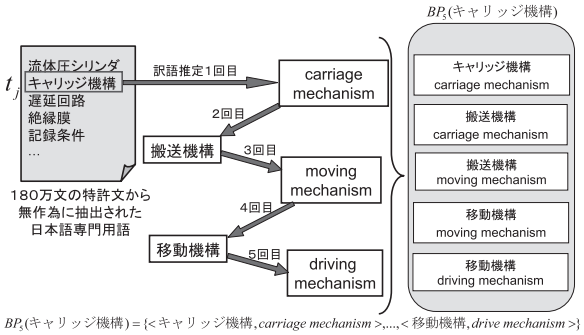

図 3: 同義専門用語候補集合の作成

表 2: 全同義専門用語候補集合に含まれる日英対訳対数

	合計	$BP_5$1 個あたりの平均
日英対訳対	1,921	38.4
語義 ID の種類数	257	5.1

(4) (2) および (3) の同定結果をもとに，同義専門用語集合を同定する．本論文では着手しておらず，今後の課題とする．

3 同義専門用語候補集合の作成

図 3 に，同義専門用語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日本語専門用語 t_J に対し，全対訳特許文 180 万件から得られた対訳専門用語[1] を用いて訳語推定を行い，英語専門用語を得る．
2. 1 で得られた英語専門用語に対し訳語推定を行い，日本語専門用語を得る．
3. 1，2 の処理を繰り返し，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 BP_k とする（本研究では，$k = 5$ とした）．
4. 日英対訳対が正しい対応でない場合，手動で除外する．

本論文では，50 個の t_J を用いて，50 個の BP_5 を作成した．単一の日英対訳対 $\langle t_J, t_E \rangle$ が 2 つ以上の異なる BP_5 に出現する場合，片方の BP_5 を廃棄し，再度作成した．また，BP_5 中に含まれる日英対訳対に対し，語義 ID を人手により付与し，日英対訳対が同義ならば同一の語義 ID を付与した．表 2 に，50 個の BP_5 中に含まれる日英対訳対および語義 ID の種類数を示す．

また，各 BP_5 に含まれる日英対訳対 $\langle t_J, t_E \rangle$ 間の同義，異議関係を判定するために，各 BP_5 の中に含まれる全組合わせ $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組を作成した．表 3 に，BP_5 中に含まれれる $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組数を示す．

[1]対訳特許文 180 万件中の頻度が 6 以上 1000 以下で，日英方向の訳語推定を行う場合は日英方向のフレーズ翻訳テーブルの順位が 1 位の対訳対，英日方向の訳語推定を行う場合，英日方向のフレーズ翻訳テーブルの順位が 1 位の対訳対を使用した．

表 3: $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組数

	各 BP_5 から作成した全組数 (割合)	$BP_5$1 個あたりの平均
合計	106,461(100%)	2129.2
同義組合計	16,615(15.6%)	332.3
異議組合計	89,846(84.4%)	1796.9

表 3 に示す全 $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組を評価対象とし，決定的規則による同義，異議組の同定を 5 節で行い，機械学習による同義，異議組の同定を 6 節で行う．

4 素性

決定的規則や機械学習に用いる素性として，表 1 に示す素性を用いた．素性は大きくわけて，文字列素性と共起語素性に分類される．

4.1 文字列素性

以下では，特に表 1 に示す $f5$,$f6$,$f7$ 素性について説明する．$f5$ 素性は，文字列の非共有箇所に対し要素合成法の同一訳が存在するか否かを求める．例えば,「アース電位」「接地電位」という対応の場合，非共有箇所である「アース」「接地」に対し，要素合成法で同一訳が得られれば素性の値は真となる．$f6$ および $f7$ の素性は，非包含箇所の文字列から，同義異義関係を求める．例えば，$f6$ で「プリンタ」「プリンター」という対応の場合，非包含箇所である「ー」から人手で作成した規則により同義と判定されれば，素性の値は真となる．また，人手による規則は評価事例を用いて作成した．

4.2 共起語素性

表 1 に示す $f8$ の素性は，共起語一致数を示す．$f8(t^i_J, t^j_J)$ の場合は，対訳特許文 180 万件内で，t^i_J, t^j_J と共起する日本語フレーズの中で，ϕ^2 尺度 [森下 08] の値が 0.00001 以上かつ上位 1000 位以内のものをそれぞれ求め，一致数を求めた．$f8(t^i_E, t^j_E)$ の場合も，同様に行った．

5 決定的規則による同定

表 1 に，決定的規則に用いる素性を示す．決定的規則として，同義の決定的規則と異義の決定的規則を定義した[2]．決定的規則が成り立った場合に，同義または異義を同定する．決定的規則により，全 $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組中，3.1%の同義組および異議組に対し同定を行い，残りの 96.9%に対し 6 節に示す機械学習による同定を行う．

5.1 同義の決定的規則

同義の決定的規則に用いる素性として，同義文字列素性と同義共起語素性の 2 つを定義した．2 つのうちいずれかの素性の値が真である場合，$\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ を

[2]本論文では，評価事例を用いて規則の作成を行った．今後は，評価事例以外の事例を用いて，規則を作成する

表 1: 決定的規則および SVM 学習に用いた素性

素性タイプ	素性名	定義
文字列素性	$f1$: 文字列が同一	$f1(t^i_X, t^j_X)$: t^i_X, t^j_X が同一ならば真となる
	$f2$: 編集距離類似度	$f2(t^i_X, t^j_X) = 1 - \frac{ED(t^i_X, t^j_X)}{max(\|t^i_X\|, \|t^j_X\|)}$: ED は t^i_X と t^j_X の間の編集距離，$\|X\|$ は X に含まれる文字数を示す
	$f3$: バイグラム類似度	$f3(t^i_X, t^j_X) = \frac{\|bigram(t^i_X) \cap bigram(t^j_X)\|}{max(\|t^i_X\|, \|t^j_X\|)+1}$: $bigram(X)$ は，X に含まれる文字単位のバイグラム
	$f4$: 同一の形態素 (単語) 数の割合	$f4(t^i_X, t^j_X) = \frac{\|morph(t^i_X) \cap morph(t^j_X)\|}{max(\|t^i_X\|, \|t^j_X\|)}$: $morph(X)$ は，X に含まれる形態素
	$f5$: 非共有箇所に対し要素合成法の同一訳が存在	$f5(t^i_X, t^j_X)$: t^i_X, t^j_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い，同一訳が存在する場合，素性の値は真となる．
	$f6$: 同義包含関係あり	$f6(t^i_X, t^j_X)$: t^i_X, t^j_X の一方がもう一方に包含されており，かつ非包含箇所が「ー」「s」「es」など，t^i_X, t^j_X が同義と判定できる文字列
	$f7$: 異義包含関係あり	$f7(t^i_X, t^j_X)$: t^i_X, t^j_X の一方がもう一方に包含されており，かつ非包含箇所が「flexible」「potential」など，t^i_X, t^j_X が異義と判定できる文字列
共起語素性	$f8$: 共起語一致数	$f8(t^i_X, t^j_X) = \|\ cooccur(t^i_X) \cap cooccur(t^j_X)\ \|$: $cooccur(X)$ は，対訳特許文 180 万件内で X と共起し，かつ ϕ^2 尺度の値が上位 1000 位以内の単語
翻訳素性	$f9$: 要素合成法の共通訳が存在	$f9(t^i_Z, t^j_Y)$:要素合成法により，t^i_Z を訳語推定し t^j_Y が得られる．または t^j_Y を訳語推定し t^i_Z が得られる
	$f10$: フレーズ翻訳テーブルの共通訳が存在	$f10(t^i_Z, t^j_Y)$:フレーズ翻訳テーブルにより，t^i_Z を訳語推定し t^j_Y が得られる．または t^j_Y を訳語推定し t^i_Z が得られる
決定的規則で用いる素性	$f11$: 同義文字列素性	$f11(t^i_J, t^j_J, t^i_E, t^j_E)$: $\{f1(t^i_J, t^j_J)$ が真 または $f5(t^i_J, t^j_J)$ が真 または $f6(t^i_J, t^j_J)$ が真 $\}$ かつ $\{\ f1(t^i_E, t^j_E)$ が真 または $f5(t^i_E, t^j_E)$ が真 または $f6(t^i_E, t^j_E)$ が真 $\}$
	$f12$: 同義共起語素性	$f12(t^i_J, t^j_J, t^i_E, t^j_E)$: $f8(t^i_J, t^j_J)$ の値が 800 以上かつ $f8(t^i_E, t^j_E)$ の値が 100 以上
	$f13$: 異議文字列素性	$f13(t^i_J, t^j_J, t^i_E, t^j_E)$: $f7(t^i_J, t^j_J)$ が真 または $f7(t^i_E, t^j_E)$ が真

$X \in \{J, E\}$, $(Z, Y) \in \{(J, E), (E, J)\}$

同義と判定する．同義決定的規則は，全組の 0.9%に適用された．また，同義組に対する同義同定の適合率は 100%(966/966), 再現率は 5.8%(966/16,615) となった．

以下では，同義決定的規則の各素性の性能を調べる．同義文字列素性の，同義組に対する同義同定の適合率は 100%(810/810), 再現率は 4.9%(810/16,615) となった．また，同義共起語素性の，同義組に対する同義同定の適合率は 100%(290/290), 再現率は 1.7%(290/16,615) となった．

5.2 異義の決定的規則

異義の決定的規則に用いる素性として，異義包含素性を定義した．異義包含素性の値が真である場合に，$\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ を異議と判定する．異義の決定的規則は，全組の 2.2%に適用された．また，異義組に対する異義同定の適合率は 100%(2391/2391), 再現率は 2.7%(2391/89,846) となった．

6 機械学習による判定

5 節で示した決定的規則による同定ができない 103,104 組の $\langle t^i_J, t^i_E \rangle$-$\langle t^j_J, t^j_E \rangle$ 組に対し，10 分割交差検定を用いて機械学習による判定を行う．SVM で評価に用いる素性は，訓練・評価事例以外を用いて表 1 の中から最適な組み合わせを決定した．また，信頼度の低いものを識別する手段として，分離平面から評価事例までの距離に下限を設定し，下限に満たない評価事例がある場合はそれらを除いた．下限値の調整の際には，訓練・評価事例以外の事例を用いた．各素性を用いた場合において，適合率，再現率，F 値がそれぞれ最大となる結果を表 4 に示す．

表 4 に，SVM による同義判定結果を示す．ベースラインは，t^i_J-t^j_J が同一または t^i_E-t^j_E が同一のものとした．ベースラインに対し，SVM で適合率が最大となるモデルとの比較を行ったところ，再現率が下がったものの，適合率が有意水準 1%で改善された (89.2%から 93.2%に改善)．

また，表 4 に，SVM による異議判定結果を示す．ベースラインは，t^i_J-t^j_J が同一でなく，かつ t^i_E-t^j_E が同一でないものとした．ベースラインに対し，SVM で適合率が最大となるモデルとの比較を行ったところ，再現率が下がったものの，適合率が有意水準 1%で改善された (88.8%から 93.7%に改善)．

表 5 に，SVM による改善例を示す．

表 4: 機械学習による同義・異義判定の性能評価 (%)

(1) 同義の判定

	適合率	再現率	F 値
ベースライン	89.2 (4742/5376)	33.6 (4742/15648)	48.8
SVM - 適合率が最大となる下限	**93.2** (3617/3879)	23.1 (3617/15649)	37.0
SVM - F 値が最大となる下限	71.0 (8922/12561)	57.0 (8922/15649)	**63.3**

(2) 異義の判定

	適合率	再現率	F 値
ベースライン	88.8 (86821/97728)	99.3 (86821/87455)	93.8
SVM - 適合率が最大となる下限	**93.7** (65488/69891)	74.9 (65488/87455)	83.2
SVM - F 値が最大となる下限	91.5 (85647/93625)	97.9 (85647/87455)	**94.6**

表 5: SVM による改善例

(1) 同義判定
ベースライン:t^i_J-t^j_J が同一または t^i_E-t^j_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

$\langle t^i_J, t^i_E\rangle$-$\langle t^j_J, t^j_E\rangle$	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	f8: 共起語一致数(log)
⟨アース電位‖ground potential⟩–⟨接地電位‖grounded potential⟩	同義	異義	同義	(t^i_J, t^j_J): 0.69 (t^i_E, t^j_E): 3.29

(b) ベースラインのみで同義と判定し不正解

$\langle t^i_J, t^i_E\rangle$-$\langle t^j_J, t^j_E\rangle$	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	f8: 共起語一致数(log)
⟨薬液‖liquid⟩–⟨冷却媒体‖liquid⟩	異義	同義	異義	(t^i_J, t^j_J) : 0 (t^i_E, t^j_E) : 0.61

(2) 異義判定
ベースライン:t^i_J-t^j_J が同一でない かつ t^i_E-t^j_E が同一でない
SVM:適合率が最大となる下限を用いたモデル

(a)SVM のみで異義と判定し正解

$\langle t^i_J, t^i_E\rangle$-$\langle t^j_J, t^j_E\rangle$	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	f2: 編集距離類似度	f3: バイグラム類似度	f4: 同一の形態素(単語)数の割合
⟨管状型部材‖tubular member⟩–⟨中空筒部材‖tubular member⟩	異義	同義	異義	(t^i_J, t^j_J): 0.5 (t^i_E, t^j_E): 1	(t^i_J, t^j_J) : 0.2 (t^i_E, t^j_E) : 1	(t^i_J, t^j_J): 0.33 (t^i_E, t^j_E): 1
⟨ファクシミリ送信‖data transmission⟩–⟨送信データ‖data transmission⟩	異義	同義	異義	(t^i_J, t^j_J): 0.1 (t^i_E, t^j_E): 1	(t^i_J, t^j_J) : 0.13 (t^i_E, t^j_E) : 1	(t^i_J, t^j_J): 0.5 (t^i_E, t^j_E): 1

(a) ベースラインのみで異義と判定し不正解

$\langle t^i_J, t^i_E\rangle$-$\langle t^j_J, t^j_E\rangle$	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	f2: 編集距離類似度	f3: バイグラム類似度	f4: 同一の形態素(単語)数の割合
⟨Pトランジスタ‖p transistor⟩–⟨P型トランジスタ‖p-type transistor⟩	同義	異義	同義	(t^i_J, t^j_J):0.89 (t^i_E, t^j_E):0.71	(t^i_J, t^j_J): 0.75 (t^i_E, t^j_E): 0.65	(t^i_J, t^j_J) : 0.67 (t^i_E, t^j_E) : 0.50

同義判定および異義判定にて，SVM で適合率が最大となるモデルとベースラインを比較した．例 (2) の場合は，t^i_E-t^j_E が同一なため，ベースラインで異義組である (⟨ 薬液,liquid⟩-⟨ 冷却媒体, liquid⟩) を同義と判定してしまった．一方，SVM では異義と判定することができた．SVM の素性である共起語一致数 ($f8$) は，t^i_E-t^j_E が一致していることから $f8(t^i_E, t^j_E)$ の値が高いものの，t^i_J-t^j_J の意味が異なるため値が低くなった．そのため，これらの素性が有効に働き，SVM は異義と判定できたと考えられる．例 (4) の場合は，t^i_J-t^j_J が異なり，かつ t^i_E-t^j_E が異なるためベースラインで同義組である (⟨ P トランジスタ,p transistor⟩-⟨P 型トランジスタ, p-type transistor⟩) を異義と判定してしまった．一方，SVM では同義と判定することができた．SVM の素性である編集距離類似度，バイグラム類似度，同一形態素 (単語) 数の割合が有効に働いたためと考えられる．

7 おわりに

本論文では，対訳特許文から獲得された専門用語対訳対を用いて，同義専門用語集合の分析と同定を行った．評価実験においては，まず対訳特許文から得られた対訳対から同義専門用語候補集合を作成し，それに対し人手で作成した決定的規則により同義専門用語を同定する．さらに，決定的規則によって同定できない専門用語に対し，SVM を用いた同定を行った．決定的規則や SVM の素性として，文字列の類似度や共起語の類似度を用いた．その結果，決定的規則により全体の約 6%の同義集合に対して，同定を行うことができた．また，決定的規則により同定できなかった同義専門用語集合に対して，機械学習により適合率 93.2%, 再現率 23.1%で同定を行うことができた．

同義専門用語集合同定の研究で，機械学習を用いたものに [Tsunakawa08] らの手法がある．本研究と，Tsunakawa らの研究で大きく異なる点として，Tsunakawa らの研究は既存の辞書である JST 辞書に含まれる対訳対に対して同義集合の同定を行うのに対し，本研究では対訳特許文から抽出した対訳対に対して同義集合の同定を行う点があげられる．本論文では，対訳特許文に含まれる共起語を用いることにより，同義判定の適合率を向上させた．

今後は，本論文で得られた同定結果をもとに，同義専門用語集合を同定し，分析を行う．また，SVM の素性や決定的規則の作成に，評価事例以外の事例を用いて評価を行う．

参考文献

[Fujii08] 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).

[Koehn07] Koehn, P., et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th ACL, Companion Volume*, pp. 177–180 (2007).

[森下 08] 森下洋平, 宇津呂武仁, 山本幹雄：対訳特許文書からの専門用語対訳辞書半自動獲得におけるフレーズテーブルと既存対訳辞書の併用, 情報処理学会研究報告, Vol. 2008, No. (2008–NL–187), pp. 91–98 (2008).

[外池 07] 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).

[Tsunakawa08] Tsunakawa, T. and Tsujii, J.: Bilingual Synonym Identification with Spelling Variations, *Proc. 3rd IJCNLP*, pp. 457–464 (2008).

[Vapnik98] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).