対訳特許文書からの専門用語対訳辞書生成: 機械学習によるフレーズテーブルと既存対訳辞書の統合*

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†

1 はじめに

本論文では，対訳特許文書から専門用語訳語対を獲得する手法を提案する [森下 08]．本論文では，**フレーズベース統計的機械翻訳モデル**により学習されるフレーズテーブルと，既存の対訳辞書を用いる **要素合成法** [外池 07] を併用する (図 1)．評価実験においては，まずフレーズテーブルにより生成された訳語候補と要素合成法により生成された訳語候補をそれぞれ単独で評価し，また，翻訳資源が異なるこれら二手法によって生成された訳語候補が一致する場合について，その訳語候補を評価する．さらに，Support Vector Machines (SVM) [Vapnik98] を用いた機械学習により，得られた訳語候補の検証を行う．SVM の素性は，既存の対訳辞書を利用したものや，全対訳文から得られる統計量などを用いた．その結果，SVM を用いることによりフレーズテーブルを用いた訳語候補推定の性能を改善することができた．

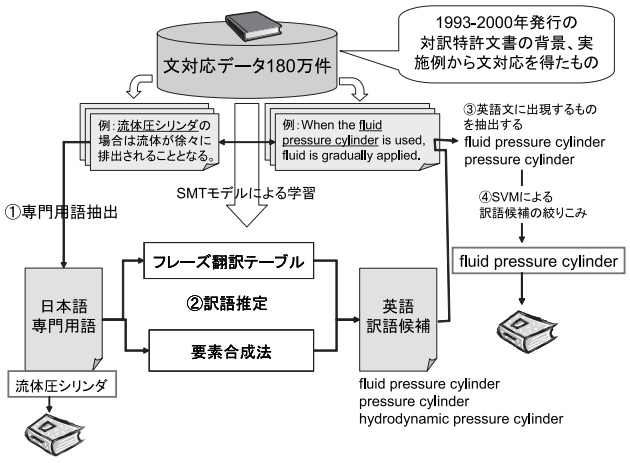

図 1: 複数の訳語推定手法を併用した対訳文からの対訳専門用語獲得の流れ

2 日英対訳特許文

本研究では，NTCIR-7 の特許翻訳タスク [Fujii08] で配布された 1,798,571 件の文対応データを，フレーズテーブルの学習データとして，またその中の日本語文から抜き出した専門用語を評価対象データとして使用した．

3 訳語推定手法

3.1 既存の対訳辞書を用いた手法

3.1.1 英辞郎

既存の対訳辞書を用いる手法として収録語数約 129 万語である英辞郎[1]Ver.79 を使用した．

3.1.2 要素合成法

名詞句を構成要素に分解し，既存の対訳辞書（英辞郎）を用いて構成要素ごとに訳語を求め，それらを再構成して全体の訳を得る要素合成法 [外池 07] を用いる．要素合成法によって，対象日本語名詞句の訳語候補と，それらが対応するスコアを求める．また，複数の訳語候補が生成された場合，スコアが高い順に順位づけられる．

3.2 統計的機械翻訳モデルのフレーズテーブル

フレーズベースの統計的機械翻訳モデルのツールキットである Moses を用いて，2 節で述べた文対応データから，フレーズペアおよびフレーズペアが対応する確率を示したフレーズテーブルを作成する．以下に Moses がフレーズテーブルを作成する過程を示す．

(1) 文対応データの前処理として，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)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である GIZA++を用いて，最尤な単語対応を得る．英日，日英の両方向で行う．(3) 日英両方向の単語対応から，対称な単語対応をヒューリスティックスを用いて得る．(4) 対称な単語アラインメントを用いて,「フレーズ対応の一貫性制約」という条件を定義し，この条件を満たすかどうかを調べる．一貫性制約に違反しない例として，”加速度依存性” と”acceleration dependency” の例を図 2 に示す. 以下に「フレーズ対応の一貫性制約」の条件を示す.

日本語フレーズを $Pj(= Aj \cdots Cj)$ とし，英語フレーズを $Pe(= Ae \cdots Ce)$ とする．ただし，$Aj \cdots Cj$ は日本

*Generating Technical Term Bilingual Lexicon from Parallel Patent Documents: Integrating Phrase Translation Table and a Bilingual Lexicon by Machine Learning

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

[1]http://www.eijiro.jp/

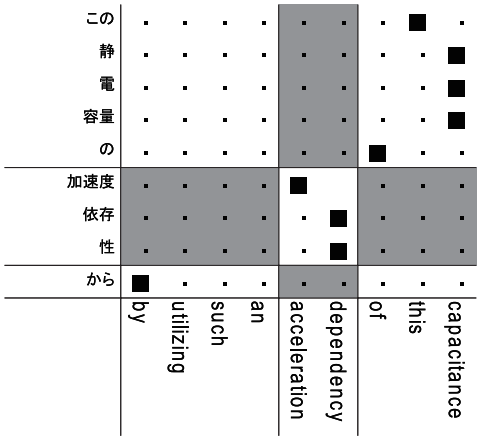

図 2: 一貫したフレーズ対応の例

語フレーズの形態素列，$Ae \cdots Ce$ は英語フレーズの単語列である．ここで，対訳文中に含まれるあらゆる単語対応 (Bj, Be) について，「Bj が Pj に含まれている $\Leftrightarrow$ Be が Pe に含まれている」が成り立つ場合に，Pj と Pe は一貫性制約に違反しない，と定義する．

一貫性制約に違反するフレーズ対応は正しくない場合が多いが，そうでない場合もある．よって，これらの情報を訳語候補の絞りこみや SVM の素性として使用し，適合率の向上に貢献するかを検証する．

(5) 学習データからフレーズ対応の数を数えてフレーズ翻訳確率を付与する [Koehn03]．本論文では，フレーズテーブルのスコアとして，フレーズの日英翻訳確率 $P(en \mid ja)$ を用いた．さらに，日本語フレーズの見出し語ごとに，英語フレーズをスコアの高い順に順位づけした．

4 各訳語推定手法単独の出力およびその共通部分の性能評価

4.1 評価手順

全対訳文 180 万件から，IPC 分類が均等になるように，無作為に抽出した対訳文を評価用データとした．

表 1 の左半分に，各訳語推定手法にて，訳語候補が生成された日本語専門用語の割合および日本語専門用語 1 語あたりの訳語候補平均数を示す．要素合成法およびフレーズテーブルにより生成された訳語候補は，スコアもしくは確率値が高い順に順位付けされている．そこで，表 1 の右半分および表 2 と表 3 に示すように，各訳語推定手法の順位が 1 位の訳語候補を評価した．例外的に二つ以上訳語が存在したものについては，評価において全ての訳語候補を順位 1 位として扱った．

表 1: 各訳語推定手法の比較

訳語推定手法	訳語候補が英文中に存在した日本語専門用語の割合 (%)	日本語専門用語 1 語あたりの訳語候補平均数	日本語専門用語全体におけるスコア 1 位の適合率 (%)
英辞郎	16.8	1.01	97.1
要素合成法	43.4	1.03	93.1
フレーズテーブル			
1. フレーズ対応の一貫性制約なし	91.3	3.00	86.8
2. フレーズ対応の一貫性制約あり	74.6	1.06	91.9

以下の日本語専門用語集合に対する各訳語推定手法単独での性能を，表 1 の右半分および表 2 と表 3 に示す．

(a) 日本語専門用語全体 (1,040 個)

(b) 集合 $E \cap P$（全ての手法により同一の訳語候補を出力できた 174 個の日本語専門用語）

(c) 集合 $(C \cap P) - E$（フレーズテーブルと要素合成法により同一の訳語候補を出力できたが，英辞郎では訳語を生成できなかった 487 個の日本語専門用語）

(d) 集合 $P - (C \cap P)$（フレーズテーブルのみにより訳語候補を生成できた 875 個の日本語専門用語）

また，表 3 に示すように，フレーズテーブルに対しフレーズ対応の一貫性制約による絞りこみを行ったものを別に評価した．

4.2 評価結果

日本語専門用語集合全体 (a) に対しては，英辞郎および要素合成法の再現率は低いが適合率は 90%を超えた．一方，フレーズテーブルは再現率が 79.2%，適合率が 86.8%となった．本論文では，対訳専門用語の半自動獲得において，再現率よりも適合率を重視する立場をとる．そこで，本論文では，フレーズテーブルによって得られた訳語候補のうち信頼度の高いものを選別することを目的とする．ベースラインを集合 (a) に対するフレーズテーブルの適合率 (約 87%) に設定し，これより高い適合率を目指す．

三手法全てにより，同一の訳語候補を得られた日本語専門用語集合 (b) と，要素合成法およびフレーズテーブルにより，同一の訳語候補を得られた日本語専門用語集合 (c) に対する F 値は 90%を超え，ベースラインより高いものとなった．集合 (b) および (c) では，既存の対訳辞書およびフレーズテーブルという，異なる性質を持つ翻訳資源を併用することにより，共通の訳語を得た場合に適合率を改善する結果が得られた．また，集合 (b) と (c) を合わせると，全日本語専門用語の 43%に対し，約 95% の適合率を実現している．したがって，以上のような手法により，対訳専門用語の半自動獲得において高い適合率を実現するという，本論文の目的を達成できることが分かる．

また，集合 (c)(d) に対し，フレーズテーブルにフレーズ対応の一貫性制約による絞込みを行うと，再現率が下がるものの適合率を大きく改善する結果が得られた．

表 2: スコア一位の訳語候補の再現率/適合率/F 値 (%) (フレーズ対応の一貫性制約による絞りこみなし)

(b) 集合 $E \cap P$ を対象とした性能（三手法全てによって同一の訳語候補を出力できた 174 個の日本語専門用語）

英辞郎	要素合成法	フレーステーブル	三手法のスコア 1 位が一致
97.7 (170/174)	97.1 (169/174)	96.0 (167/174)	96.0 (167/174)
97.7 (170/174)	97.1 (169/174)	96.0 (167/174)	**98.8** (167/169)
97.7	97.1	96.0	**97.4**

(c) 集合 $(C \cap P) - E$ を対象とした性能（要素合成法とフレーステーブルのみによって同一の訳語候補を出力できた 487 個の日本語専門用語）

要素合成法	フレーステーブル	要素合成法 1 位とフレーステーブル 1 位の訳語候補が一致
96.5 (470/487)	93.4 (455/487)	92.4 (450/487)
96.5 (470/487)	93.4 (455/487)	97.2 (450/463)
96.5	93.4	94.7

(d) 集合 $P - (C \cap P)$ を対象とした性能 (フレーステーブルのみによって訳語候補を出力できた 875 個の日本語専門用語)

フレーステーブル
83.8 (733/875)
83.8 (733/875)
83.8

表 3: スコア一位の訳語候補の再現率/適合率/F 値 (%) (フレーズ対応の一貫性制約による絞りこみあり)

(b) 集合 $E \cap P$ を対象とした性能（三手法全てによって同一の訳語候補を出力できた 174 個の日本語専門用語）

フレーステーブル	三手法のスコア 1 位が一致
86.2 (150/174)	85.6 (149/174)
97.4 (150/154)	98.0 (149/152)
91.5	91.4

(c) 集合 $(C \cap P) - E$ を対象とした性能（要素合成法とフレーステーブルのみによって同一の訳語候補を出力できた 487 個の日本語専門用語）

フレーステーブル	要素合成法 1 位とフレーステーブル 1 位の訳語候補が一致
94.3 (457/487)	92.6 (451/487)
96.2 (459/477)	**98.3** (451/487)
95.2	95.3

(d) 集合 $P - (C \cap P)$ を対象とした性能 (フレーステーブルのみによって訳語候補を出力できた 875 個の日本語専門用語)

フレーステーブル
70.3 (615/875)
88.6 (615/694)
78.4

5 SVM による訳語候補の検証

5.1 手法

本節では，SVM を用いて，三種類の訳語推定手法によって得られた訳語候補の検証を行う．

SVM のツールとして，TinySVM[2]を用いた．また，訓練および評価事例を $\langle t_J, t_E, c \rangle$ と記述する．ここで，t_J は日本語専門用語，t_E は少なくとも一つの手法で生成された英語訳語候補，c は t_E が t_J の正解訳か否かを示す．t_E が正解の場合，$c = +$ となり，そうでない場合 $c = -$ となる．カーネル関数として，線形カーネルと二次多項式カーネルを比較し，より高い性能が得られた二次多項式カーネルを採用した．評価時においては，事例 $\langle t_J, t_E, c \rangle$ のうち日本語側に日本語専門用語 x_J を持つ事例 $\langle x_J, t_E, c \rangle$ を集めて，クラス c の判定を行い，十分に信頼できる事例があれば，それを選別する方式で評価を行った．本論文では，SVM の分離平面から，評価事例までの距離を信頼度とし，x_J を共有する事例の中で分離平面からの距離が最長のものを選択した．

表 2 中の 174 個日本語専門用語集合 (b) では，3 手法によって得た訳語候補の共通部分の適合率が 98%を超えており，これ以上の適合率の上昇が見込めない．その為，487 個の日本語専門用語集合 (c) と，875 個の日本語専門用語集合 (d) から得られた訓練，評価事例を用意した．これら 2 種類の集合から得られた訓練，評価事例は別々に扱い，それぞれに対して 10 分割交差検定を行った．

5.2 素性

表 4 に，SVM に用いた素性を示す．

表 4: SVM 学習に用いた素性

素性タイプ	素性
単言語素性 (集合 (d) が対象)	日本語専門用語の形態素数
	英語訳語候補の単語数
二言語素性— 英辞郎を利用	要素合成法により出力された訳語候補のスコアと順位 (集合 (c) が対象)
	日本語専門用語・英語訳語候補の構成要素の対応が少くとも一つ英辞郎に存在 (集合 (d) が対象)
二言語素性 — 対訳文から得られる統計量を利用	フレーステーブルに含まれる訳語候補の日英翻訳確率と順位
	分割表の頻度 $freq(t_E, t_J)$, $freq(t_E, \neg t_J)$, $freq(\neg t_E, t_J)$
	フレーズ対応の一貫性制約の違反

5.3 評価結果

フレーズ対応の一貫性制約を用いるにあたって，以下の三つの条件で SVM による絞りこみを行った．

(1) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いない．(2) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いる．(3) フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外する．SVM の素性に違反の有無素性を用いない．

表 5 に結果を示す．訳語候補のうち，信頼性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，適合率が最高となる下限を用いた結果を表枠内の上に，F 値が最高となる下限を用いた結果を下に示す．ここで，表 2，表 3，表 5 に示す太字の数字は，各日本語専門用語集合 (b)(c)(d) 各々において全ての結果を比較した場合に，最高となる適合率と F 値である．

日本語専門用語集合 (c) の評価においては，要素合成法およびフレーズテーブルにより得られた訳語候補の

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

表 5: SVM のスコア 1 位の訳語候補の再現率/適合率/F 値 (%)

(c) 集合 $(C \cap P) - E$ を対象とした性能（要素合成法とフレーズテーブルのみによって同一の訳語候補を出力できた 487 個の日本語専門用語）

		全ての素性の中から，最適なものを使用		
英辞郎ベースの素性のみを使用	統計ベースの素性のみを使用	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いない．	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いる．	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外する．SVM の素性に違反の有無素性を用いない．
87.5(426/487) 96.4(426/442) 91.7	80.7(393/487) 96.3(393/408) 87.8	92.4(450/487) 96.2(450/468) 94.2	67.1(327/487) 97.0(327/337) 79.4	91.0(443/487) 97.4(443/455) 94.1
87.5(426/487) 87.5(426/487) 87.5	94.9(462/487) 94.9(462/487) 94.9	95.7(466/487) 95.7(466/487) 95.7	96.1(468/487) 96.1(468/487) **96.1**	93.2(454/487) 95.6(454/475) 94.4

(d) 集合 $P - (C \cap P)$ を対象とした性能 (フレーズテーブルのみによって訳語候補を出力できた 875 個の日本語専門用語)

	全ての素性の中から，最適なものを使用		
統計ベースの素性のみを使用	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いない．	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外しない．SVM の素性に違反の有無素性を用いる．	フレーズ対応の一貫性制約に違反する訳語候補をあらかじめ除外する．SVM の素性に違反の有無素性を用いない．
25.3(221/875) 91.7(221/241) 39.6	44.0(385/875) **92.5**(385/416) 59.6	24.2(212/875) 92.2(212/230) 38.4	26.3(230/875) 88.5(230/260) 40.5
81.6(714/875) 81.6(714/875) 81.6	82.4(721/875) 82.4(721/875) 82.4	81.9(717/875) 81.9(717/875) 81.9	68.6(600/875) 86.5(600/694) 76.5

共通部分の適合率 (98.3%) および F 値 (95.3%) をベースラインとした．その結果，評価事例における適合率は 97.4%となり，ベースラインよりも劣ってしまった．一方，F 値は 96.1%となり改善されたものの，ベースラインとの差は統計的に有意でない．また，日本語専門用語集合 (d) の評価においては，フレーズテーブルの適合率 (88.6%) および F 値 (83.8%) をベースラインとした．その結果，評価事例における適合率は 92.5%となり，ベースラインを大きく改善した．これらの差は有意水準 1%のもとで統計的に有意である．一方，F 値は 82.4%となりベースラインよりも劣ってしまった．

以上の結果から，SVM を用いた訳語候補の検証により，対訳専門用語の半自動獲得において F 値または適合率を改善することができた．表 6 に，SVM による改善例を示す．集合 (c) に対しては，SVM とベースラインで各々の F 値が最も高くなる結果を比較した．また，集合 (d) に対しては，SVM とベースラインで各々の適合率が最も高くなる結果を比較した．

表 6: SVM による改善例

(c) 集合 $(C \cap P) - E$ に対し，SVM のみで訳語候補を出力し正解した日本語専門用語

日本語専門用語	正解英語訳語	英語訳語候補	フレーズテーブルのスコア	フレーズテーブルの順位	要素合成法のスコア	要素合成法の順位	$freq(t_j,t_k)$	$freq(t_j,-t_k)$	$freq(-t_j,t_k)$	一貫性制約に違反
連結板	connecting plate	connecting plate	0.26		3.28	1	99	173	81	する
		a connecting plate	0.02				21	251	16	する
		connecting plate member	0.02	1			8	264	0	しない
		plate	0.01				223	49	40626	する
		and a connecting plate	5.45E−03				5	267	1	する
		a connecting plate member	5.45E−03	2			2	270	0	しない
護岸壁	shore protection wall	shore protection wall	1		4.17	1	3	0	0	する

(d) 集合 $P - (C \cap P)$ に対し，SVM のみで訳語候補を出力し正解した日本語専門用語

日本語専門用語	正解英語訳語	英語訳語候補	フレーズテーブルのスコア	フレーズテーブルの順位	要素合成法の部分訳	日本語専門用語形態素数	英語訳語候補単語数	$freq(t_j,t_k)$	$freq(t_j,-t_k)$	$freq(-t_j,t_k)$	一貫性制約に違反
シート状包装体	sheetlike wrapping body	sheetlike wrapping body	1		あり	4	3	7	0	0	する
遠心脱水装置	spin extractor	spin extractor	1		あり	3	2	4	0	0	する

(d) 集合 $P - (C \cap P)$ に対し，ベースラインのみで訳語候補を出力し不正解の日本語専門用語

日本語専門用語	正解英語訳語	英語訳語候補	フレーズテーブルのスコア	フレーズテーブルの順位	要素合成法の部分訳	日本語専門用語形態素数	英語訳語候補単語数	$freq(t_j,t_k)$	$freq(t_j,-t_k)$	$freq(-t_j,t_k)$	一貫性制約に違反
電解イオン水供給ライン	electrolytic ionic water discharge line	line	0.25	1	あり	5	1	1	4	69019	しない
		discharge line	0.25	1	あり	5	2	1	4	25	しない
		first discharge line	0.25	1	あり	5	3	1	4	4	しない

6 おわりに

本論文では，対訳特許文に対して，複数の訳語推定手法を併用し，対訳専門用語獲得の性能を改善する手法を提案した．翻訳知識源が異なる 2 種類の訳語推定手法を用いることと，SVM を用いて，フレーズテーブルから得た訳語候補を検証し，信頼度の低いものを排除することで，適合率または F 値を改善することができた．

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).

[Koehn03] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. HLT-NAACL*, pp. 127–133 (2003).

[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).

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