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

森下洋平[†] 宇津呂武仁[†] 山本幹雄[†]

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，高適合率で対訳専門用語を獲得する手法を提案する．提案手法では，フレーズベース統計的機械翻訳モデルを用いて訳語推定を行う手法と，既存の対訳辞書を利用した要素合成法を併用するアプローチをとる．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で，訳語推定の適合率を改善することができた．また，Support Vector Machines (SVM) を用いて，フレーズベース統計的機械翻訳モデルにより生成した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した．その結果，フレーズテーブルから得た訳語候補の適合率を改善することができた．

Integrating Phrase Translation Table and a Bilingual Lexicon in Semi-Automatic Acquisition of Technical Term Translation Lexicon from Parallel Patent Documents

Yohei MORISHITA ,[†] Takehito UTSURO [†] and Mikio YAMAMOTO [†]

This paper presents an attempt at developing a technique of acquiring translation pairs of technical terms with sufficiently high precision from parallel patent documents. The approach taken in the proposed technique is based on integrating the phrase translation table of a state-of-the-art statistical phrase-based machine translation model, and compositional translation generation based on an existing bilingual lexicon for human use. Our evaluation results clearly show that the agreement between the two individual techniques definitely contribute to improving precision of translation candidates. We then apply the Support Vector Machines (SVMs) to the task of automatically validating translation candidates in the phrase translation table. Experimental evaluation results again show that the SVMs based approach to translation candidates validation can contribute to improving the precision of translation candidates in the phrase translation table.

1. はじめに

機械翻訳や人手による翻訳を行う場合，高い質を保つためには大規模で正確な対訳辞書が必要不可欠となる．人手によって対訳辞書を作成するのは，膨大な時間と労力を要するため，テキストから対訳辞書を自動作成する研究は，10 年以上にわたって行われてきた．対訳辞書に登録する訳語対を抽出する手法として，対訳文中の共起頻度を用いる手法[7]，コンパラブルコーパスから訳語対の獲得を行う手法[2]，既存の対訳辞書を用いた要素合成法[11]，検索エンジンから訳語が併記された文書を収集し，訳語対を獲得する手法[3] 等がある．

† 筑波大学大学院 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

しかし，これらの手法は，実用的な局面において，半自動的に専門用語対訳辞書を構築する際に利用するには，十分な性能であるとは言い難い．まず，対訳文から訳語対を獲得する手法と比較して，コンパラブルコーパスや，検索エンジンのスニペットから訳語対を獲得する手法は，十分な性能であるとは言い難い．また，対訳コーパスを用いたとしても，自動的に作成した対訳辞書は，半自動または手動で作成した辞書と比較して信頼性が低いのでそのままでは使えない．

以下，本論文では特許翻訳を例として考える．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るといえる．しかし，年々新しい技術開発が行

われ，新しい専門用語が産み出され，特許が申請されている．そのため，新しい専門用語を継続的に登録し対訳辞書を増強することが必要である．

継続的に対訳辞書を増強するために，対訳対自動獲得のアプローチが考えられる．しかし，このアプローチはあまり実用的ではない．自動で訳語対を獲得する手法が実用的でない理由は，多くの場合自動で獲得した訳語対を後から手動で選別する必要があり，効率が悪いからである．ただし，訳語対自動獲得手法の適合率が十分に高く，獲得した訳語対を手動で選別する時間を節約できるならば，訳語対の自動獲得手法を利用して，効率的に対訳辞書の増強を行うことも不可能ではない．以上のことから，実用的局面において訳語対自動獲得手法が活用できるためには，十分に高い適合率が必要不可欠である．ここで，適合率を重視するあまり，再現率が低くなり，その結果獲得できる訳語対の規模が小さくなるという懸念が考えられるかもしれない．しかし，対訳特許文書は大量に存在するため，適合率の高い訳語対のみをそこから獲得したとしても，十分な量の訳語対を獲得できる．そのため，適合率を重視して継続的に対訳辞書を増強することは可能である．

本論文では以上の議論にもとづいて，対訳特許文書から高適合率で専門用語訳語対を獲得する手法を提案する．本論文では，**フレーズベース統計的機械翻訳モデル**[5] 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**[11] を併用する．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) を用いて，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から生成した訳語候補の適合率を改善することができた．

2. 日英対訳特許文

本研究では，NTCIR-7 の特許翻訳タスク[1] 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 400 件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．なお，配布された文対応データは，以下の手順で得られたものである．

(1) 1993-2000 年発行の日本公開特許広報全文と米国特許全文を得る．
(2) 米国特許の中から日本に出願済みのものを優先権番号より得て，日米対訳特許文書を取得する．
(3) 日米特許で互いに対応関係にある部分（背景，実施例）を抽出し，文アラインメント[12] をつける．

表 1 は，以上の手法により求めた全対訳文の IPC 分類を示したものである．

表 1 対訳特許文全体の IPC 分類の分布

IPC 分類		文書数	%	文数	%
A.	生活必需品	1,606	3.5	41,180	2.4
B.	処理操作:運輸	5,948	12.8	165,994	9.2
C.	科学:冶金	1,606	3.5	22,933	1.3
D.	繊維:紙	331	0.7	7,148	0.4
E.	固定構造物	255	0.6	5,906	0.3
F.	機械工学:照明:加熱:武器:爆破	3,941	8.5	113,604	6.3
G.	物理学	16,533	35.7	786,650	43.7
H.	電気	16,127	34.8	642,163	35.7
	合計	46,347	100.0	1,798,571	100.0

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☆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[11] 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．図 1 に具体的な手順を示す．

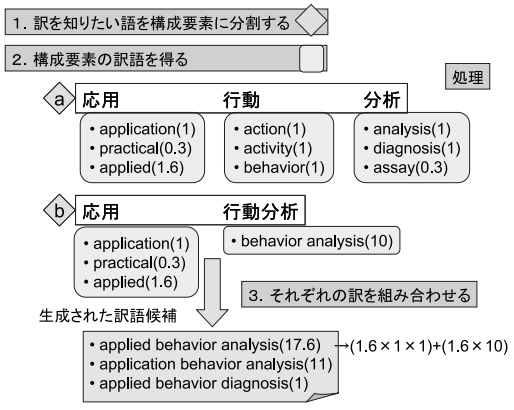


図 1 要素合成法による訳語推定

まず,「応用行動分析」を構成要素に分割し，その中に既存の対訳辞書に見出し語として登録されているものがあるかを検索する．図 1 中の，“a” および”b” が分割され，かつ見出し語が存在した構成要素となる．次に，それぞれの構成要素を英訳する．ここでは，それぞれの構成要素の訳に対しスコアが付与される．最後に，単語の連結ルールに従いそれぞれの訳語を再構成し，全体の訳語を得る．ここで構成要素に与えられるスコアは，対訳辞書に現れる構成要素の頻度などから求める．また，訳語全体のスコアは，構成要素訳のスコアの積となる．図 1 のように，複数の訳語候補が生成された場合，スコア

☆ http://www.eijiro.jp/

が高い順に順位づけられる．

また，構成要素の訳語を求め，スコアを付与するのに，要素合成法では 2 種類の辞書を用いる．既存の対訳辞書である英辞郎と，英辞郎から生成した部分対応対訳辞書である．部分対応対訳辞書とは，「複合語中の構成要素がどのように訳されるのが自然か」の情報を含む辞書である．例えば，「applied」は，一般的には「応用の」，「応用された」などに訳されるが，複合語としては

applied mathmatics ↔ 応用 数学

のように，「応用」と訳すことが多い．それにもかかわらず，既存の対訳辞書には applied ↔ 応用 という対応が登録されておらず，このような組合わせは数多く存在する．部分対応対訳辞書は，そのような複合語としての自然な訳を既存の対訳辞書から生成し，登録した辞書である．また，部分対応対訳辞書には二つの種類があり，前接語として自然な訳を登録する前方一致部分対応対訳辞書，後接語として自然な訳を登録する後方一致部分対応対訳辞書がある．

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[5] 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++[9] 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) 対称な単語対応表に矛盾しない，一貫したフレーズ対応をすべて求める．
(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[6]．

フレーズ対応のモデルパラメータは，フレーズの英日翻訳確率 $P(ja \mid en)$，英日方向の単語の翻訳確率 (IBM モデル) の積，日英翻訳確率 $P(en \mid ja)$，日英方向の単語の翻訳確率 (IBM モデル) の積，フレーズペナルティ(常に自然対数の底 e=2.718) の 5 つである．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，400 件の対訳文を無作為に抽出したものを評価用データとした☆．日英対訳文から辞書に登録すべき専門用語の日英対訳ペアを獲得するのに用いた方法を図 2 に示す．

(1) 全対訳文データ 180 万件中 400 件の日本語文を形態素解析し，日本語名詞句を得る．さらに，その中に含まれる専門用語 1,040 個を人手で抽出する☆☆．
(2) 得られた 1,040 個の日本語専門用語に対し，3 種類の訳語推定手法（図 2：A，B，C）それぞれで訳語推定を行い，英語訳語候補を得る．
(3) 得られた英語訳語候補の内，英文中に出現するものを抽出する．具体的には，$\langle S_J, S_E \rangle$ の日本語文 S_J 中の日本語専門用語 t_J に対して，3 種類の訳語推定手法を用いて英語訳語候補を生成し，英語文 S_E に出現する訳語候補を抽出する．
(4) 最後に，SVM を用いて英文中に出現した訳語候補の検証を行う．ここで，SVM に用いる素性は既存の対訳辞書を利用したもの，全対訳文から得られた統計量などを用いた．

表 2 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数および生成された全訳語候補数を示す

表 2 各訳語推定手法によって得られた訳語候補数（日本語専門用語 1,040 個が対象）

訳語推定手法	訳語候補が生成された日本語専門用語数	生成された訳語候補数 (日本語専門用語 1 語あたりの平均数)
英辞郎	175	177 (1.01)
要素合成法	450	465 (1.03)
フレーズテーブル	950	2851 (3.00)

また，図 3 に各訳語推定手法によって得られた訳語候補が英文中に出現した日本語専門用語の割合を示す．図 3 中の (a) は 1,040 個の全日本語専門用語を表し，その中で英辞郎によって訳を得られたものを E，要素合成法によって訳を得られたものを C，フレーズテーブルによって訳を得られたものを P とする．また，全ての手法により同一の訳語候補を生成できた日本語専門用語の集合を $E \cap P$，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生成できたが，英辞郎では訳語を生成できなかった日本語専門用語の集合を $(C \cap P) - E$，フレーズテーブルのみにより訳語候補を生成できた日本語専門用語の集合を $P - (C \cap P)$ と表す．

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 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．一方，英辞郎に

☆ 本研究の主たる目的は，対訳文から，対訳専門用語を半自動獲得することであり，そのような設定のもとではフレーズテーブル学習および訳語対獲得に同一の対訳文を用いることは極めて自然なことである．ただし，フレーズテーブル学習に用いた大規模な対訳文には含まれない対訳文を情報源として対訳専門用語獲得を行うという設定も十分に考えられる．

☆☆ 特許翻訳の実務機関等が本論文の手法を使用するという局面では，大規模な日本語専門用語辞書を用いることにより，対訳文から日本語専門用語の抽出を行う場合が多い．

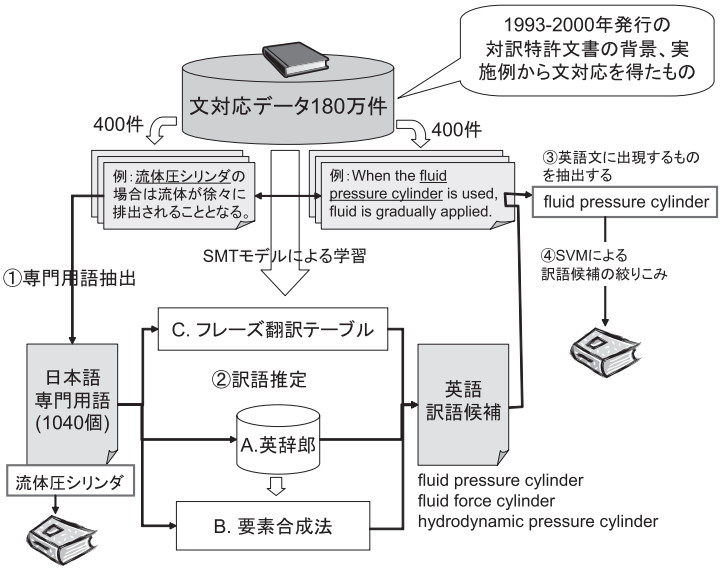


図 2 複数の訳語推定手法を併用した対訳文からの対訳専門用語獲得の流れ

よって訳語候補を得ることができた日本語専門用語に対して，英辞郎により出力した訳語は日本語専門用語 1 個につきほぼ 1 つしか存在しなかった☆．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．

4.2 評価結果

以下に示す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 3 の左半分に示す．

(a) 日本語専門用語全体 1,040 個
(b) 集合 $E \cap P$
(c) 集合 $(C \cap P) - E$
(d) 集合 $P - (C \cap P)$

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約 80%，適合率が約 87%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 1,040 個の 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．

5. SVM による訳語候補の検証

5.1 手　　法

本節では，Support Vector Machines[13] (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を検証して，信頼度の高いものと低いものに選別する．

SVM のツールとして，TinySVM☆☆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．1,040 個の日本語専門用語中，三手法のいずれかによって訳語を推定できたものは 954 個であり，それらの訳語候補の総数は 2,851 個であったため，訓練・評価事例は 2,851 個となる．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 1 つ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，以下の条件を満たすものを 1 つ選別した．

☆ 実際には，1,040 個の日本語専門用語中，英辞郎に見出し語が登録されているものは 321 個存在し，各日本語専門用語につき平均 2.31 個の見出し語が存在した．しかし，321 個の日本語専門用語中，英辞郎の見出し語が対訳文中の英語文側に出現したものは 175 個で，さらにそれらほぼ全ての日本語専門用語につき英文中に現れた訳語は一つのみだった．

☆☆ http://chasen.org/~taku/software/TinySVM/

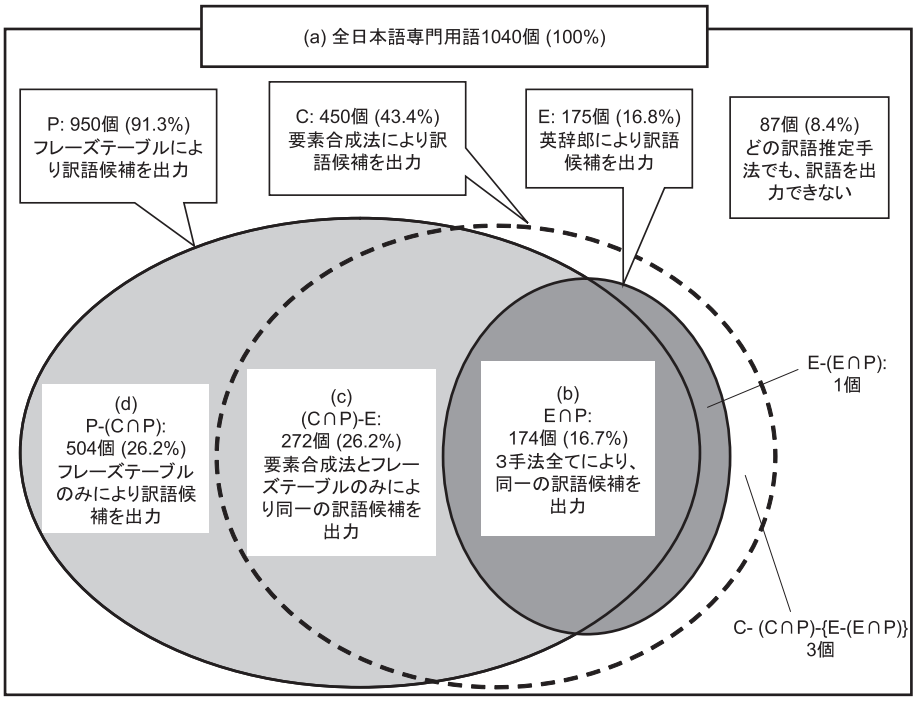


図 3 英辞郎，要素合成法，フレーズテーブルの訳語候補が英文中に存在した割合

表 3 スコア一位の訳語候補の再現率/適合率/F 値 (%)

(a) 1,040 個の全日本語専門用語を対象とした各訳語推定手法の性能

英辞郎	要素合成法	フレーズテーブル
16.3 (170/1040)	40.3 (419/1040)	79.3 (825/1040)
97.1 (170/175)	93.1 (419/450)	86.8 (825/950)
28.0	56.2	82.9

(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	フレーズテーブル	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ズテーブルのみによって同一の訳語候補を出力できた 272 個の日本語専門用語）

要素合成法	フレーズテーブル	要素合成法 1 位とフレーズテーブル 1 位の訳語候補が一致	SVM による検証
91.9 (250/272)	90.8 (247/272)	89.3 (243/272)	93.0 (253/272)
91.9 (250/272)	90.8 (247/272)	**93.1** (243/261)	93.0 (253/272)
91.9	90.8	**91.2**	**93.0**

(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 504 個の日本語専門用語)

フレーズテーブル	SVM による検証
81.5 (411/504)	57.5(290/504)
81.5 (411/504)	**90.1**(290/322)
81.5	70.2
	72.8(366/504)
	87.1(366/420)
	79.2

(1) SVM による判定結果が+のもの

(2) 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

表 3 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 97%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，272 個の日本語専門用語集合 (c) と，504 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．結果を表 3 中 “SVM による検証” に示す．

5.2 素　　性

表 4 に，SVM に用いた素性を示す．素性は大きくわけて，単言語素性と二言語素性から構成される．

単言語素性としては，日本語専門用語の構成要素数と，英語訳語候補の単語数を用いた．これらの素性を表 3 中

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	フレーズ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 $freq(t_E, t_J)$, $freq(t_E, \neg t_J)$, $freq(\neg t_E, t_J)$

の日本語専門用語集合 (c) および (d) において評価した結果，(c) に関してはこれらの素性を用いないほうが高い結果が得られた．そのため，本論文では (d) のみに対してこれらの素性を用いることにした．

一方，二言語素性としては，既存の対訳辞書である英辞郎を利用した素性と，対訳文から得られる統計量を利用した素性を用いた．英辞郎を利用した素性の一つは，要素合成法によって各訳語候補に付与されたスコアと順位である．この素性は，日本語専門用語集合 (c) の訳語候補のみに与えられる．(d) は要素合成法で訳語が生成できなかった集合である．そのため，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するか否かの情報を素性とした．例えば,「応用行動分析」「application behavior analysis」という日本語専門用語・英語訳語候補の組で,「分析」「analysis」という訳語対が英辞郎に登録されていれば，素性の値は真となる．

対訳文から得られる統計量を用いた素性の一つは，フレーズテーブルによって各訳語候補に付与されたスコアと順位である．また，もう一つの素性として，従来より統計的共起尺度にもとづいて訳語推定を行う手法でよく用いられた相互情報量，ϕ^2 尺度，dice 係数，対数尤度比[7] の考え方を利用する．各尺度の情報を流用するために，各尺度の値を求めるのに必要な，日英用語の共起頻度等の統計量を，素性の一つとして用いた．具体的には，日本語専門用語 t_J と英語訳語候補 t_E から，下に示す分割表から得られる共起頻度を素性とした．

	t_J	$\neg t_J$
t_E	$freq(t_E, t_J)$	$freq(t_E, \neg t_J)$
$\neg t_E$	$freq(\neg t_E, t_J)$	$freq(\neg t_E, \neg t_J)$

上に示す値から，ϕ^2 スコアを求め，素性とした場合も評価したが，素性に加えないほうが高い性能を得られた．そのため，ϕ^2 の値は計算せず，分割表に示す共起頻度の値を直接素性として用いた．

5.3 評 価 結 果

表 3 中の列”SVM による検証” の欄に結果を示す．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共通部分の F 値 (91.2%) をベースラインとした．SVM による検証結果の F 値は 93.0%となり，ベースラインの F 値 91.2%を改善することができた．ただし，両者の差は統計的には有意ではない．

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および F 値をベースラインとした．集合 (d) に対しては，フレーズテーブルのみが訳語候補を出力できている．(d) の適合率および F 値は 81.5%であり，(b) および (c) のものよりも低い．この場合，SVM を用いることにより，フレーズテーブルが出力した訳語候補のうちで，信頼性の低いものを識別する必要がある．そこで，ここでは，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．その結果，評価事例における適合率は 90.1%となった．一方，F 値と適合率を両立する下限での適合率は 87.1%となった．これらの結果とベースラインとの差は，有意水準 5%のもとで統計的に有意である．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適合率を改善することができた．

6. 日本語名詞句のタイプ単位評価のための分析

これまでに述べた評価では，ある日本語専門用語に対して訳語を獲得する際に，一つの対訳文のみしか考慮していない．しかし，本来は日本語専門用語が出現する全ての対訳文を考慮する必要がある．例えば，日本語専門用語が出現する対訳文に高い頻度で出現する英語訳語候補は，正解訳語である確率が高い可能性がある．また，場合によっては各対訳文の IPC 分類などを利用し，訳し分けを行う必要があるかもしれない．本節では，日本語専門用語の訳語を求める過程において，日本語専門用語が出現する全ての対訳文を考慮する，タイプ単位評価に必要な分析を述べる．

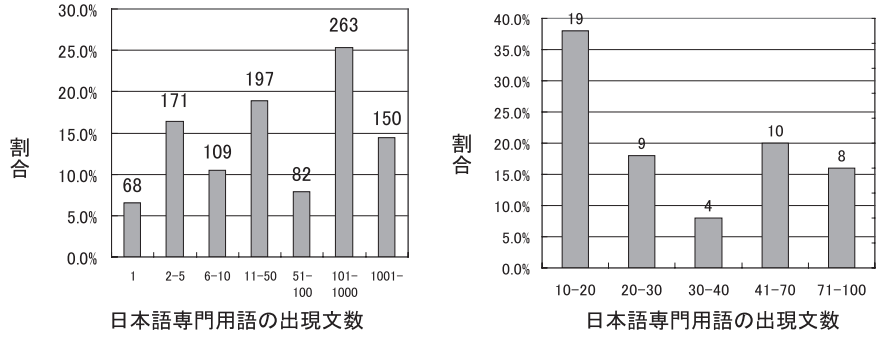
6.1 データセット

日本語専門用語 1,040 個の，全対訳文における出現頻度分布を図 4(a) に示す．また，日本語専門用語が出現する文の IPC 分類の種類数の分布を図 5(a) に示す．

以下に，タイプ単位評価の分析を行うのに必要なデータの作成手順を示す．

(1) 1,040 個の日本語専門用語中，対訳文 180 万件の出現数が 10 回以上 100 回以下のものから，無作為に 50 個抽出する．各日本語専門用語が出現した対訳文の総数は 1,862 件となり，日本語専門用語一語あたり平均 37.2 件に出現した．各日本語専門用語の出現頻度分布を図 4(b) に，出現する文の IPC 分類種類数の分布を図 5(b) に示す．

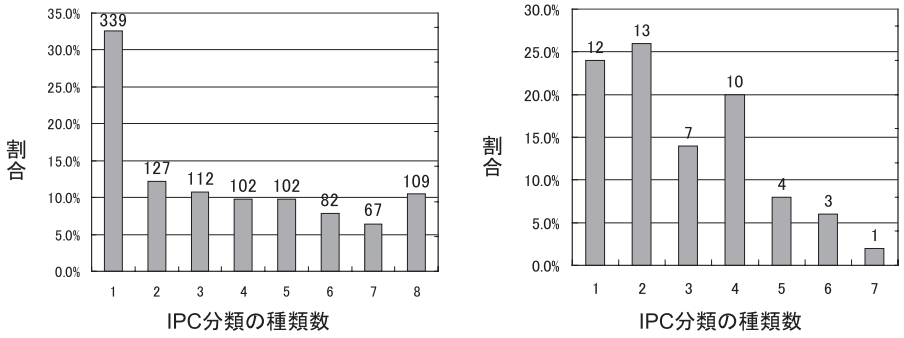
(2) 50 個の日本語専門用語に対し各訳語推定手法によ



(a) 日本語専門用語 1,040 個

(b) 正解訳語を付与した
日本語専門用語 50 個

図 4 日本語専門用語の出現頻度分布



(a) 日本語専門用語 1,040 個

(b) 正解訳語を付与した
日本語専門用語 50 個

図 5 日本語専門用語が出現した文の IPC 分類の種類数の分布

り英語訳語候補を求め，日本語専門用語が出現する対訳文それぞれに対して，英文中に出現する訳語候補を抽出する．各訳語推定手法によって得られた平均英語訳語候補数を表 5 に示す☆．また，各日本語専門用語タイプに対して，それぞれの訳語推定手法で求めた英語訳語候補の総種類数の分布を図 6 に示す．

(3) 各対訳文に対し日本語専門用語の正解訳を人手で求め☆☆，(2) で得た訳語候補が正解かを判定する．

6.2 訳語推定結果を用いた分析

ここでは，タイプ単位評価に必要な情報の分析を行う．各日本語専門用語タイプあたりの正解英語訳語の総種類数の分布を図 7 に示す．

また，日本語専門用語 t_J が出現する対訳文の集合において英語訳語候補 t_E が出現する回数と t_E の正解率の関係をプロットしたものを図 8 に示す．図 8 からわかるように，英語訳語候補 t_E の出現頻度が多いほど，正解である確率が高くなる．

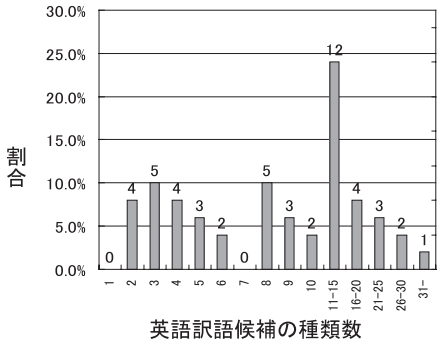


図 6 日本語専門用語 1 タイプあたりの
英語訳語候補の種類数の分布
(正解訳語付与済み日本語専門用語 50 個が対象)

7. 関 連 研 究

訳語対の自動獲得の手法で，統計的機械翻訳モデルにより学習されたフレーズテーブルを用いたものに，Itagaki[4] らの手法がある．本研究と，Itagaki らの研究で大きく異

☆ 日本語専門用語のトークン 1,862 個に対し，訳語推定手法のいずれかにより訳語候補を得られたものは 1,766 個，全ての手法により同一の訳語候補を得られたものは 297 個，要素合成法または英辞郎により訳語候補を生成できたが，フレーズテーブルではできなかったものが 11 個，フレーズテーブルのみによって訳語候補を生成できたものは 1,004 個，どの訳語推定手法を用いても英語訳語候補を得られなかったものは 93 個となった．

☆☆ 本論文では，日本語専門用語の訳は，全対訳文に対して一律ではなく，各対訳文ごとに異なる正解訳が存在すると仮定した．

表 5 各訳語推定手法によって得られた訳語候補数 (正解訳語を付与した日本語専門用語 50 個の全トークンが対象)

訳語推定手法	日本語専門用語数	平均訳語候補数 (マクロ平均)	(a) 日本語専門用語のトークン数	(b) 全英語訳語候補数	(b)/(a) 平均訳語候補数 (マイクロ平均)
英辞郎	13	1.000	307	307	1.000
要素合成法	32	1.005	762	764	1.003
フレーズテーブル	50	2.093	1754	4154	2.368

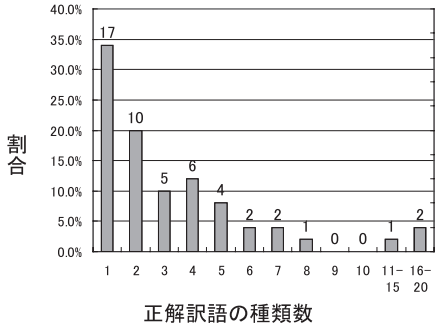


図 7 日本語専門用語 1 タイプあたりの正解英語訳語の種類数の分布 (正解訳語付与済み日本語専門用語 50 個が対象)

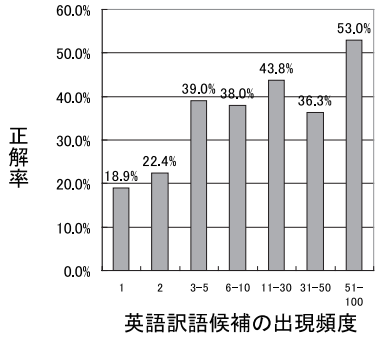


図 8 英語訳語候補の出現数と正解率の関係 (正解訳語付与済み日本語専門用語 50 個が対象)

なる点として，本研究はフレーズテーブルに加えて既存の対訳辞書を用いた要素合成法[11] を用い，それらを併用している点があげられる．本研究では，異なる二種類の翻訳資源を用いることにより，訳語推定の適合率を改善することができた．

Rosti[10]，Matusov[8] らの研究は，複数の手法を併用することにより，全体のシステムの性能を向上させる点が本研究と共通している．これらの研究では，文全体の翻訳において，複数の機械翻訳技術を併用し，翻訳結果を互いに補完しあう事が目的である．一方，本研究は対訳専門用語の半自動獲得を目的とし，再現率よりも適合率の向上を重視している．

8. 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，高適合率で対訳専門用語を獲得する手法を提案した．提案手法では，フレーズベース統計的機械翻訳モデルを用いて訳語推定を行う手法と，既存の対訳辞書を利用した要素合成法を併用するアプローチをとる．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で，訳語推定の適合率を改善することができた．また，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した．その結果，フレーズテーブルから得た訳語候補の適合率を改善することができた．

参 考 文 献

1) A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, 2008.
2) P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th and 36th*, pp. 414–420, 1998.
3) F. Huang, Y. Zhang, and S. Vogel. Mining key phrase translations from Web corpora. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 483–490, 2005.
4) M. Itagaki, T. Aikawa, and X. He. Automatic validation of terminology translation consistency with statistical method. In *Proc. MT summit XI*, pp. 269–274, 2007.
5) P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.
6) P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pp. 127–133, 2003.
7) Y. Matsumoto and T. Utsuro. Lexical knowledge acquisition. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 24, pp. 563–610. Marcel Dekker Inc., 2000.
8) E. Matusov, N. Ueffing, and H. Ney. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proc. 11th*, pp. 33–40, 2006.
9) F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. Vol. 29, No. 1, pp. 19–51, 2003.
10) A.-V. Rosti, S. Matsoukas, and R. Schwartz. Improved word-level system combination for machine translation. In *Proc. 45th ACL*, pp. 312–319, 2007.
11) M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proc. 2nd Intl. Workshop on Web as Corpus*, pp. 11–18, 2006.
12) M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proc. MT summit XI*, pp. 475–482, 2007.
13) V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.