対訳特許文書からの専門用語対訳辞書生成: 統計的機械翻訳におけるフレーズテーブルと訳語推定手法の併用*

森下 洋平† 宇津呂 武仁‡ 山本 幹雄‡
筑波大学第三学群工学システム学類†，筑波大学大学院 システム情報工学研究科‡

1 はじめに

翻訳者にとって, 専門用語を訳す作業は大変労力を要する. 既存の辞書に登録されていない専門用語が文書に出現した場合, 対処法としてインターネットや他の専門文書などを参照する, などの方法が考えられるが, 正しい訳語を得るのに時間がかかることが多く翻訳者の作業効率は大幅に下がってしまう. 特に特許文書にはそのような専門用語が頻繁に出現し, 1ヵ月に訳 1 万語のペースで増加し続けている. そのため, 専門用語を特許文書から抽出し, 正しい訳を自動推定して, 翻訳辞書作成者を支援するシステムが求められている.

本研究では, 日本語特許文書およびその米国出願英文対訳特許文書の対を用いて訳語推定を行い, 翻訳辞書作成者を支援するアプローチをとる. 具体的には, 統計的機械翻訳モデルの学習によって得た**フレーズテーブル**を用いて訳語を推定する方法, 日英名詞句が対訳特許文書内で共起する頻度を用いた**統計的共起測定法**, および既存の対訳辞書を用いて単語の構成要素の訳語を取得し, それらを再構成して全体の訳語候補を得る**要素合成法**の 3 手法を併用して, 専門用語の対訳辞書に登録すべき訳語対の候補を自動生成する.

具体的には, NTCIR-7 の特許翻訳タスク [内山 07b] で配布された対訳特許文データの日本語文から日本語名詞句を抽出し, その中から人手で抜き出した専門用語を対象にし, 上に述べた 3 手法による訳語推定を行う. 次に, それぞれの手法で得た日本語専門用語の訳語候補が, 専門用語を抜き出した文と対訳となる英文中に出現した場合, それらの訳語候補は正解かを人手で判定し, 専門用語から正解訳語を得る性能を検証, 比較した. そこから, 3 手法を組み合わせ, 翻訳辞書作成者を支援する方法を考察する.

*Generating Technical Term Bilingual Lexicon from Parallel Patent Documents using a Phase Table and Compositional Translation Estimation

†Yohei Morishita, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
‡Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba

2 日英対訳特許文

本研究では, NTCIR-7 の特許翻訳タスクで配布された 1,798,571 件の文対応データを使用した. これらは, 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 [Utiyama07a] をつける.

3 訳語推定手法

3.1 英辞郎

専門用語が, 既存の辞書に登録されているか否かを調べるために, 既存の対訳辞書として, 収録語数約 129 万語である英辞郎[1] Ver.79 を使用した.

3.2 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. 図 1 に具体的な手順を示す.

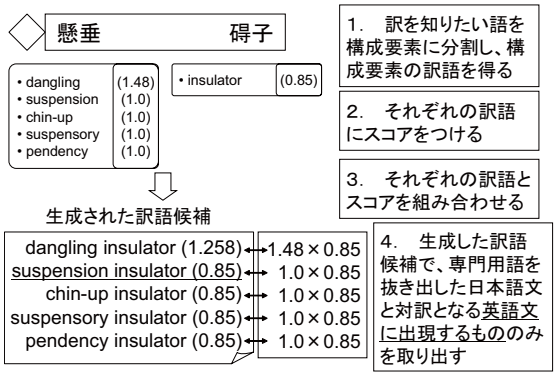


図 1: 要素合成法による訳語推定

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

スコアを求めるのに, 上で紹介した英辞郎の他に, 英辞郎から生成した, 前方, 後方一致部分対応対訳辞書を使用する. 前方, 後方一致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」の情報をスコアとして使える形に記述された辞書である.

3.3 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 [Koehn07] を用いて, 2 節で述べた文対応データから, 対応しやすいフレーズペア, 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++ [Och03] 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フレーズ対応を得る.
5. 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. reordering モデル（フレーズの並び替え確率）を計算, 生成する

作成したフレーズテーブルの形式を表 1 に示す.

表 1: フレーズテーブルの形式

日本語フレーズ	英語フレーズ	フレーズ対応のモデルパラメータ
その 結果	as a result	| 0.242637 | 0.0629573 | 0.229377 | | 0.000588454 | 2.718 |

フレーズ対応のモデルパラメータは 5 つとなり, 各種確率は, 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) となる. 今回は, 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.

表 2: 日英名詞句が対訳文に出現する頻度

	y	$\neg y$
x	$freq(x,y) = a$	$freq(x,\neg y) = b$
$\neg x$	$freq(\neg x,y) = c$	$freq(\neg x,\neg y) = d$

3.4 統計的共起測定法

表 2 内で x は日本語名詞句, y は英語名詞句であり, また $freq(x,y)$ は日本語名詞句 x と英語名詞句 y が対訳文内で共起した文数, $freq(x,\neg y)$ は日本語名詞句 x は出現したが, 英語名詞句 y は出現しなかった対訳文数, $freq(\neg x,y)$ は日本語名詞句 x は出現しなかったが, 英語名詞句 y は出現した対訳文数, $freq(\neg x,\neg y)$ はどちらも出現しなかった対訳文数である. a の値が高く, b,c の値が低いほど, 日本語名詞句 x と英語名詞句 y は対応しやすいペアだといえる. a,b,c,d それぞれのパラメータを, 全文対応データ 1,798,571 件より求めた.

これらの情報から, 以下の式を使って日本語名詞句と英語名詞句の対応のしやすさを数値化する [Matsumoto00].

$$\phi^2(x,y) = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

なお, 今回は a,b,c,d の値を求めるのに, フレーズテーブルの訳語推定結果を利用した. ただし, 英文中に出現するフレーズテーブルの訳語候補は多いため, フレーズテーブル中のスコアが 0.05 以上の訳語対のみを候補とした.

4 評価

4.1 日本語名詞句, 専門用語の抽出精度

1,798,571 件の対訳文中 165 件の日本語特許文に対し形態素解析等の処理を行うことによって日本語名詞句を抽出した. さらに日本語文から正しい日本語名詞句を抽出する性能, 専門用語を抽出する性能を求めた. 再現率, 適合率を求めるのに調べた文数はそれぞれ 30 文,800 文とした. 以下に結果を示す.

- 日本語名詞句

  再現率 = $\frac{\text{システムが抽出した正しい日本語名詞句数}}{\text{抽出すべき正しい日本語名詞句数}} = \frac{127}{138} = 92.0\%$

  適合率 = $\frac{\text{システムが抽出した正しい日本語名詞句数}}{\text{システムが抽出した日本語名詞句数}} = \frac{3541}{3737} = 94.8\%$

- 専門用語

  再現率 = $\frac{\text{システムが抽出した専門用語数}}{\text{抽出すべき専門用語数}} = \frac{76}{84} = 90.5\%$

  適合率 = $\frac{\text{システムが抽出した専門用語数}}{\text{システムが抽出した日本語名詞句数}} = \frac{2209}{3737} = 59.1\%$

表 3: 特許分類の分布（文書単位, 文単位, 専門用語単位）

分類	名称	文書単位		文単位				専門用語単位	
		全対訳文書数	割合	全対訳文数	割合	人手評価対象 対訳文数	割合	人手評価対象 専門用語数	割合
A.	生活必需品	1606	3.5%	41,180	2.4%	20	12.5%	49	12.3%
B.	処理操作:運輸	5948	12.8%	165,994	9.2%	20	12.5%	69	17.4%
C.	科学:冶金	1606	3.5%	22,933	1.3%	20	12.5%	44	11.1%
D.	繊維：紙	331	0.7%	7,148	0.4%	20	12.5%	52	13.1%
E.	固定構造物	255	0.6 %	5,906	0.3%	20	12.5%	39	9.8%
F.	機械工学：照明： 加熱：武器：爆破	3941	8.5%	113,604	6.3%	20	12.5%	41	10.3%
G.	物理学	16533	35.7%	786,650	43.7%	20	12.5%	43	10.8%
H.	電気	16127	34.8%	642,163	35.7%	20	12.5%	60	15.1%
	合計	46347	100.0%	1,798,571	100.0%	160	100.0%	397	100.0%

4.2 特許分類の分布

人手評価対象 160 件の特許文とそれに属する 397 個の専門用語, 全対訳文書 46,347 件とそれに属する対訳文 1,798,571 件を特許分類（セクション, A-H）ごとに分類した結果を表 3 に示す. 全体の割合は分類 G,H が多い結果となったが, 今回評価対象とした文対応データ 160 文は各分類に属する対訳文全体から 20 文づつ無作為に抽出したものとした.

4.3 評価対象データ

今回は, 自動で抽出した 699 個の日本語名詞句の中から, 人手で選別した 397 個の専門用語を対象に, トークン別, タイプ別の検証を行う. 397 個の専門用語と, 専門用語と英文中に存在した英語訳語候補のペアをトークンごと, タイプごとに分類したところ, 表 4 に示す結果が得られた.

表 4: 評価対象データのトークン数, タイプ数

	日本語専門用語	日本語専門用語と 英語訳語候補のペア
トークン数	397	
タイプ数	388	395

評価対象となった日本語専門用語と英語訳語候補ペアのトークン数, タイプ数はほぼ同じとなったので, 今回はトークン単位の検証のみ行う.

4.4 評価結果

397 個の日本語名詞句を対象に, 英辞郎, 要素合成法, フレーズテーブル, 統計的共起測定法によって求めた訳語候補が, 日本語名詞句を抜き出した日本語文と対訳となる英語文に出現した割合を図 2 に示す.

フレーズテーブル, 統計的共起測定法, 要素合成法, 英辞郎の順に, 訳語候補が英文中に現れる割合は 90.7%,84.6%,44.4%,15.6%となった.

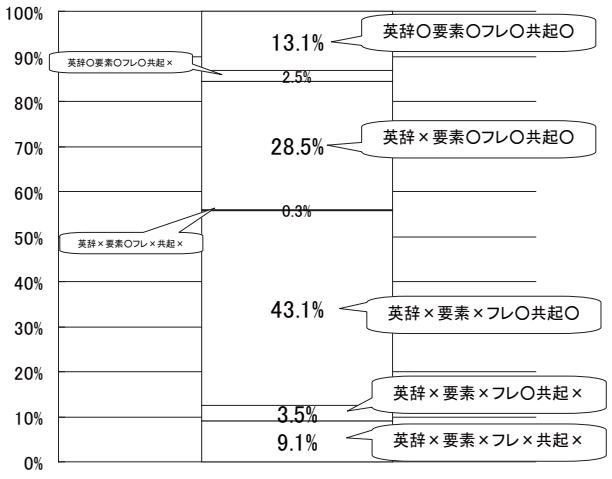
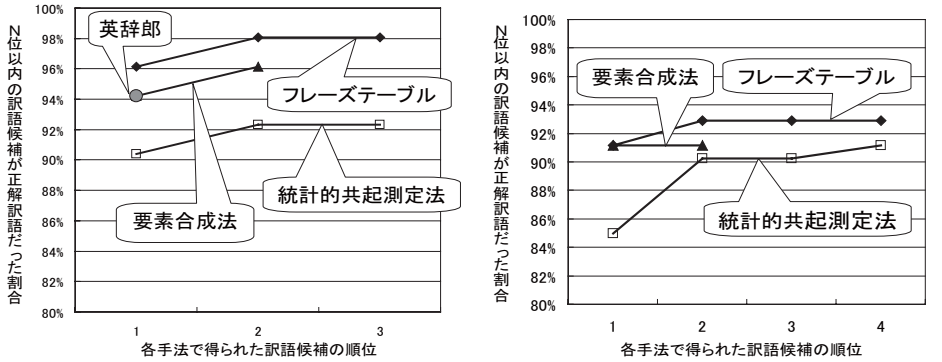


図 2: 英辞郎, 要素合成法, フレーズテーブル, 統計的共起測定法の訳語候補が英文中に存在した割合

各手法によって得られた訳語候補が対訳英文中に出現していた場合, 各手法のスコアで, 何位以内の訳語候補が正解だったかを人手で評価した結果を図 3 に示す.

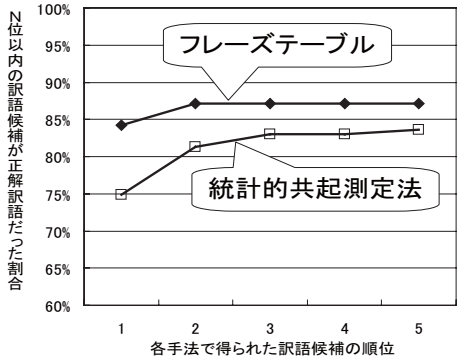
フレーズテーブルのみが対訳英文中に現れる訳語候補を出力した場合 (図 3(d)), 正解訳語が含まれる割合は 55%程度だが, 4 つの手法が対訳英文中に現れる訳語候補を出力した場合 (図 3(a)), フレーズテーブルの訳語候補に正解が含まれる割合は 95%以上と増加する. このように, 訳語候補を出力する手法が多くなるほど, 正解訳語を出力する割合が高くなることがわかる.

また, フレーズテーブルと統計的共起測定法を比較した場合, 全ての面においてフレーズテーブルの方が性能が高く, 2 つの手法を併用することによる利点がない結果となった. 双方の訳語候補が英文中に出現した場合 (図 3(c)), フレーズテーブルの訳語候補のみが英文中に出現した場合 (図 3(d)) と比べて, 訳語候補が正解訳語を含む割合が高くなる. しかし実際は統計的共起測定法を併用しなくてもフレーズテーブルのスコア 0.05 を境に図 3(c) と図 3(d) の切り分けが可能である.

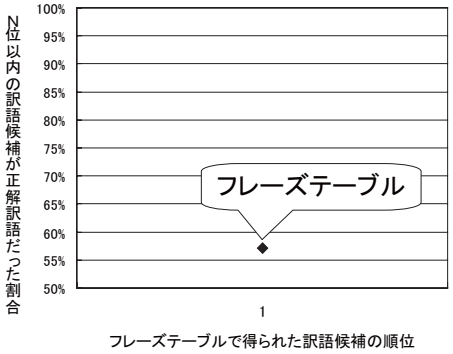


(a) 英辞○要素○フレ○共起○
(4 手法の訳語候補が
いずれも対訳英文中に存在)

(b) 英辞×要素×フレ○共起○
(3 手法の訳語候補が
いずれも対訳英文中に存在)



(c) 英辞×要素×フレ○共起○
(2 手法の訳語候補が
いずれも対訳英文中に存在)



(d) 英辞×要素×フレ○共起×
(フレーズテーブルの
訳語候補が対訳英文中に存在)

図 3: 対訳英文中に存在する訳語候補が正解である割合

4.5 まとめ

以上の結果から, 正解の専門用語訳語対が 95%以上の高い可能性で含まれているものを出力するシステムを希望する時は, 4 つの手法全てで英文中に訳語候補が出現したものを, 50%程度の割合でしか正解を含まない専門用語訳語対でもいいので, とにかく多くの日本語名詞句に対して訳語候補を得たい場合は, いずれかの手法で得た訳語候補で, 英文中出現したものを出力するなど, 使用する側のニーズにあったシステムを作ることが可能である.

5 おわりに

本稿では英辞郎, 要素合成法, フレーズテーブル, 統計的共起測定法の 4 つの手法を使い日本語名詞句から正解訳語を得る精度を検証, 比較した. 今後は文中から専門用語を抽出する精度を上げると共に, どの手法を利用しても正しい訳が得られない専門用語も考慮した評価を行っていく.

参考文献

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

[Matsumoto00] Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 24, pp. 563–610, Marcel Dekker Inc. (2000).

[Och03] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003).

[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).

[Utiyama07a] Utiyama, M. and Isahara, H.: A Japanese-English Patent Parallel Corpus, *Proc. MT summit XI*, pp. 475–482 (2007).

[内山 07b] 内山将夫, 山本幹雄, 藤井敦, 宇津呂武仁：特許情報を対象とした機械翻訳: — 共通基盤による評価タスクを目指して—, 情報処理学会研究報告, Vol. 2007, No. (2007–NL–180), pp. 133–138 (2007).