フレーズテーブルと要素合成法を用いた対訳特許文書からの専門用語対訳辞書生成*

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による方法, 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三手法を併用して, 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.

2 日英対訳特許文

本研究では, NTCIR-7 の特許翻訳タスク [5] 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. [Step1].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. [Step2]. 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. [Step3]. 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 [4] 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 120 万語である英辞郎を使用した.

3.2 要素合成法

単語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 [3] 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.

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 [1] を用いて, 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. 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.

3.4 統計的共起測定法

表 1: 日英名詞句が対訳文に出現する頻度

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

表 1 内で x は日本語名詞句, y は英語名詞句であり, また $freq(x,y)$ は日本語名詞句と英語名詞句が対訳文内に共起した文数である. 表 1 に示す文の数を, 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 [2]. ただし, 今回は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対象とした.

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

4 評価

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, 5000 文中 100 文以上に出現したものの

*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,

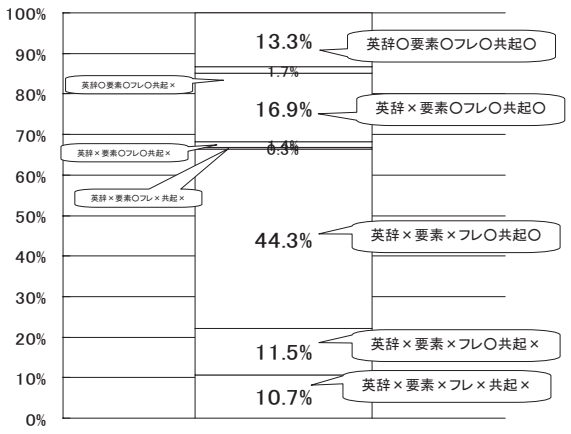


図 1: 英辞郎, 要素合成法, フレーズテーブル, 統計的共起測定法の訳語候補が英文中に存在した割合

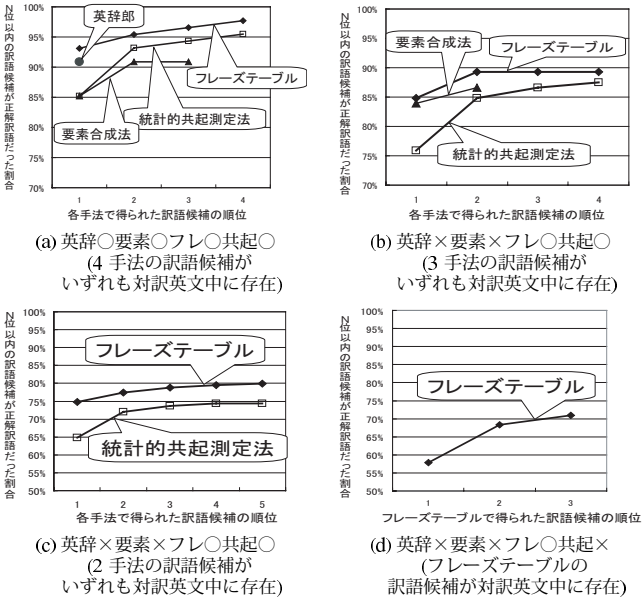
多くは専門用語ではないため, これらを取り除いた. 専門用語の抽出精度をランダムに選んだ 30 文の中から人手で求めた所, 適合率が 73.4%, 再現率が 89%となった. 得られた日本語名詞句 662 個（専門用語でないものも含む）を対象に, トークンベースでの評価を行った.

評価対象 165 件の特許文と, 全体訳文 1,798,571 件が属する特許文書を特許分類（セクション, A-H）ごとに分類した結果を表 2 に示す. 全体の割合では分類 G,H が多いが, 今回の評価対象では B セクション (処理操作, 運輸) の特許分類が多くなってしまっている.

表 2: 対訳文が属する文書分類

分類	名称	評価対象 対訳文数	割合	全対訳文数	割合
A.	生活必需品	1	0.6%	41,180	2.4%
B.	処理操作:運輸	140	84.8%	165,994	9.2%
C.	科学:冶金	5	3.0%	22,933	1.3%
D.	繊維：紙	0	0.0%	7,148	0.4%
E.	固定構造物	0	0.0%	5,906	0.3%
F.	機械工学：照明：加熱：武器：爆破	4	6.3%	113,604	6.3%
G.	物理学	8	4.8%	786,650	43.7%
H.	電気	7	4.2%	642,163	35.7%
	合計	165	100.0%	1,798,571	100.0%

662 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 1 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 89.0%,74.5%,33,5%,15,0%となった. 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 2 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 2(d)), 正解訳語が含まれる割合は



(a) 英辞〇要素〇フレ〇共起〇 (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×フレ〇共起〇 (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×フレ〇共起〇 (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×フレ〇共起× (フレーズテーブルの訳語候補が対訳英文中に存在)

図 2: 対訳英文中に存在する訳語候補が正解である割合

70%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 2(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.

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, 名詞句が属する特許文書分類も考慮した評価を行っていく.

謝辞: 本研究に関し共同で研究を行っている日本特許情報機構 奥直也氏, 大塩只明氏, 三橋朋晴氏に感謝する.

参考文献

[1] P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[2] 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.

[3] 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.

[4] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proc. MT summit XI*, pp. 475–482, 2007.

[5] 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.