日中パテントファミリーから抽出した対訳文を用いた専門用語の訳語推定*

董 麗娟† 龍 梓† 豊田 樹生† 宇津呂 武仁‡ 三橋 朋晴§ 山本 幹雄‡
筑波大学大学院 システム情報工学研究科† 筑波大学 システム情報系‡ 日本特許情報機構§

1 はじめに

近年，中国国内における特許出願は大幅な伸びを見せている．ここで，中国特許文書の翻訳は，特許文書の言語横断検索等のサービスにおいて不可欠のため，中国語の特許を日本語に翻訳する仕事が重要になっている．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不可欠となる．しかし，各国では，年々新しい技術開発が行われ，新しい専門用語が作られ，特許が申請されている．一方，人手によって，対訳辞書を作成するためには，膨大な時間と労力を要するため，自動もしくは半自動的に日中専門用語対訳辞書を構築する手法が必要である．これまでに，日英対訳特許文を情報源として，専門用語対訳対を自動獲得する手法の研究が行われてきた．文献 [3] では，NTCIR-7 特許翻訳タスク [1] において配布された日英 180 万件の対訳特許文を用いて，対訳特許文からの専門用語対訳対獲得を行った．この研究では，句に基づく統計的機械翻訳モデル [2] を用いることにより，対訳特許文から学習されたフレーズテーブルを用いることによって，専門用語対訳対獲得を行った．

ここで，上述の日英 180 万件の対訳特許文は，文献 [7] の手法により，日米パテントファミリーの対応特許文書中において,「背景」および「実施例」の部分の日英対訳文対を対応付けたものである．

そこで，本論文では，文献 [7] の手法を適用することによって，日中パテントファミリーから抽出した 360 万件の日中対訳特許文を言語資源として，句に基づく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を用いて，対訳専門用語を獲得する手法を提案する．具体的には，まず，専門用語対訳辞書獲得の情報源として用いる日中対訳文対に対して，句に基づく統計的機械翻訳モデルを適用することより，フレーズテーブルを学習する．次に，このフレーズテーブル，および，一組の日中対訳文を用いて日本語専門用語の中国語訳語推定を行う．本論文の評価実験においては，形態素単位の日本語文一文に対して，形態素単位の中国語文，および，文字単位 [5] の中国語文の 2 種類を用意した対訳文を対象とした．いずれの場合も，97%程度の適合率および F 値を達成することができた．

2 日中対訳特許文

本研究では，フレーズテーブルの訓練用データとして，約 360 万件の日中対訳特許文を用いた．なお，日中対訳特許文は，2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文に対して，以下の手順で得られたものである．

1. 文献 [7] の手法によって日中間で文対応を付ける．
2. スコア降順で上位の 360 万文対を抽出する．

3 統計的機械翻訳モデルのフレーズテーブル

句に基づく統計的機械翻訳モデルのツールキットである Moses [2] を用いて，2 節で述べた文対応データから，日中の句の組及び日中の句の組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) 文対応データに対する前処理として，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．

*Translation Estimation of Technical Terms using Parallel Sentences extracted from Japanese-Chinese Patent Families

†Lijuan Dong, Zi Long, Itsuki Toyota, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Mikio Yamamoto, Faculty of Engineering, Information and Systems, University of Tsukuba
§Tomoharu Mitsuhashi, Japan Patent Information Organization (JAPIO)

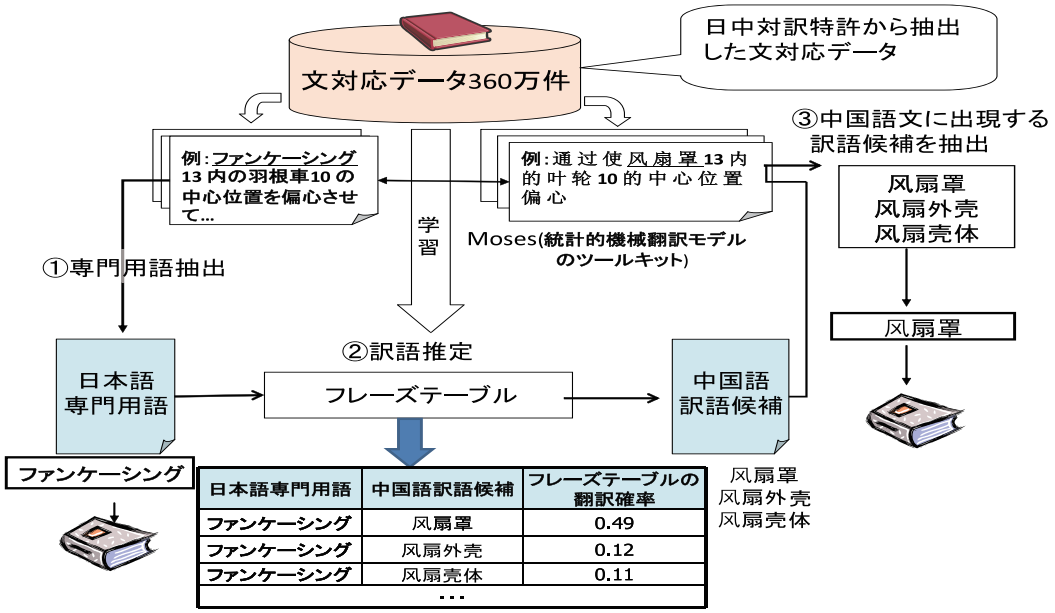

図 1: 対訳文およびフレーズテーブルを用いた対訳専門用語獲得の流れ

(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++ [4] を用いて，最尤な単語対応を得る．中日，日中の両方向で行う．

(3) 中日，日中両方向の単語対応から，ヒューリスティクスを用いて対称な単語対応を得る．

(4) 対称な単語対応を用いて，可能なすべての日中の句の組を作成する．

(5) 文対応データにおける日中の句の対応数に基づいて，各句の対応に翻訳確率等のパラメータを付与する．

本論文では，フレーズテーブルを用いて得られる中国語訳語候補のスコアとして，句対応の日中翻訳確率 $P(p_C \mid p_J)$ (ただし，p_C を中国語句，p_J を日本語句とする) を用いた．更に，日本語句の見出し語ごとに，中国語句をスコアの降順に順位付けした．

ここで，手順 (1) の対訳文は，形態素解析された形態素単位の日本語文一文に対して，Chinese Penn Treebank を用いた Stanford Word Segment [6] によって形態素解析された形態素単位の中国語文，および，文字単位[1]の中国語文の 2 種類を用意し，作成した．この 2 種類の対訳文に対して，独立に Moses を適用することより，形態素単位フレーズテーブルおよび文字単位フレーズテーブルをそれぞれ作成した．

[1]連続する数字とアルファベットは一個のトークンとして扱う．

表 1: スコア一位の訳語候補の再現率・適合率・F 値 (%)

	中国語側の区切り単位	
	形態素	文字
再現率	96.3 (497/516)	95.9 (495/516)
適合率	97.8 (497/508)	96.9 (495/511)
F 値	97.0	96.4

4 対訳文およびフレーズテーブルを用いた訳語推定

4.1 手順

全対訳文 360 万件から，無作為に抽出した 516 件を評価用対訳文とした．日中対訳文から辞書に登録すべき日中対訳専門用語を獲得するために用いた手順を図 1 に示す．

1. 全対訳文データ 360 万件中 516 件の日本語文を形態素解析し，日中対訳文 $\langle S_J, S_C \rangle$ 中の日本語文 S_J 中の日本語名詞句を得る．さらに，その中に含まれる専門用語 t_J を人手で抽出する

2. 得られた日本語専門用語 t_J に対し，統計的機械翻訳モデルのフレーズテーブルを用いて，訳語推定を行い，中国訳語候補を得る．

3. 得られた中国語訳語候補のうち，対訳文 $\langle S_J, S_C \rangle$ 中の中国語文 S_C に出現する訳語候補を抽出する．

表 2: 中国語訳語推定の成功例

	中国側の区切り単位:形態素			中国側の区切り単位:文字		
日本語専門用語	中国語訳語候補	フレーズテーブルの順位（日中対訳文の中国語側に存在）	フレーズテーブルの翻訳確率	中国語訳語候補	フレーズテーブルの順位（日中対訳文の中国語側に存在）	フレーズテーブルの翻訳確率
アンテナ装置	天线/装置	1	0.89	天/线/装/置	1	0.81
信号線駆動回路	信号/线/驱动/电路	1	0.9	信/号/线/驱/动/电/路	1	0.87
フォトダイオード	光电/二极管	1	0.83	光/电/二/极/管	1	0.81

各日本語専門用語に対し，フレーズテーブルにおける翻訳確率 1 位の訳語候補を生成し，評価対象とした．

4.2 評価対象日本語専門用語

以上の手順の後，全対訳文データ中の日本語文を形態素解析し，日本語名詞句を選択する際，以下に該当する日本語名詞句は評価対象外とした．

1. 語頭または語尾が不適切である日本語名詞句．具体的には,「上記，下記，当該，該，各」が語頭，または「等，毎」が語尾である日本語名詞句．
2. 訳語推定において用いる日中対訳文の日本語側において，他の日本語名詞句の部分文字列となる日本語名詞句．例えば,「三角波生成回路」という日本語名詞句の部分文字列である「生成回路」の部分が抽出された場合．
3. 語尾が記号である日本語名詞句．例えば,「メタンガス濃縮装置 M1」のように語尾が記号「M1」である日本語名詞句．

4.3 評価結果

表 1 に評価結果を示す．中国語側が形態素単位のフレーズテーブルを用いた場合，適合率は 97.8%，F 値は 97.0%となった．一方，中国語側が文字単位のフレーズテーブルを用いた場合，適合率は 96.9%，F 値は 96.4%となった．両者はほぼ同等の性能を達成したが，誤りの傾向は異なっている．

表 2 に，これらの両方の場合を対象として，中国語訳語推定の成功例を示す．中国語側が形態素単位の場合も文字単位の場合も，同一の訳語候補が出力され，参照用中国語訳語と同一となった．

表 3 に，中国語側が形態素単位のフレーズテーブルを用いた場合の中国語訳語推定の誤り例を示す．日本語専門用語が「動的/後退/接触/角」の場合，参照用中国語訳語**“动态/后退/接触/角”**は，中国語側が形態素単位の場合のフレーズテーブルに含まれてはいるが，中国語文の形態素解析結果**“在 /液浸部/ 12 /的 /动态 /后退/ 接触/ 角 θ /…”**において，**“角”**と**“ θ ”**が分割されなかったため，中国語文との照合が成功する訳語候補を抽出することができなかった．一方，日本語専門用語が「熱/圧着」の場合，参照用中国語訳語**“热压/接”**は，中国語側が形態素単位の場合のフレーズテーブルに含まれていた．しかし，中国語文の形態素解析結果**“…/热压 /接后/…”**において，**“接”**と**“后”**が分割されなかったため，中国語文との照合が成功する訳語候補としては，**“热压”**を出力してしまい，誤りとなった[2]．

表 4 に，中国語側が文字単位のフレーズテーブルを用いた場合の中国語訳語推定の誤り例を示す．これらの例においては，参照用中国語訳語はフレーズテーブルに含まれてはいるが，参照用中国語訳語の部分文字列の方が翻訳確率が高くなり，順位 1 位の訳語候補として出力された．例えば，日本語専門用語「複合/構造/物」の参照用中国語訳語**“复/合/结/构/物”**およびその部分文字列**“复/合/结/构”**が中国語文に出現する訳語候補として出力されたが，部分文字列**“复/合/结/构”**の方が**“复/合/结/构/物”**よりも翻訳確率が高くなり，順位 1 位の訳語候補として出力されてしまった．

5 関連研究

訳語対の自動獲得において，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として，文献 [3, 9] がある．文献 [3] においては，知識源として，句に基づく統計的機械翻訳モデルのフレーズテーブルおよび既存の対訳辞書を併用して，日英間の訳語推定を行った．一方，本研究では，日中の対訳特許文を対象として，句に基づく統計的機械翻訳モデルのフレーズテーブルを用いて，日中間の訳語推定を行っ

[2]中国語専門用語の形態素解析結果**“热压/接”**は誤りであり，正しくは**“热/压接”**と分割するべきであるが，フレーズテーブルにおける順位は**“热压/接”**の方が上位であるため，本論文の評価実験においては，**“热压/接”**を参照用中国語訳語とした．

表 3: 中国語側が形態素単位のフレーズテーブルを用いた場合の誤り例

日本語専門用語	中国語訳語候補	フレーズテーブルの順位（日中対訳文の中国語側に存在しないものも含む）	フレーズテーブルの翻訳確率	参照用中国語訳語（日中対訳文の中国語側の形態素解析誤りにより訳語候補とならず）	フレーズテーブルの順位（日中対訳文の中国語側に存在しないものも含む）	フレーズテーブルの翻訳確率
動的/後退/接触/角	訳語推定結果なし	—	—	动态/后退/接触角	1	0.67
熱/圧着	热压	2	0.24	热压/接	1	0.26

表 4: 中国語側が文字単位のフレーズテーブルを用いた場合の誤り例

日本語専門用語	中国語専門用語	フレーズテーブルの順位（日中対訳文の中国語側に存在）	フレーズテーブルの翻訳確率	参照用中国語訳語（日中対訳文の中国語側に存在）	フレーズテーブルの順位（日中対訳文の中国語側に存在）	フレーズテーブルの翻訳確率
複合/構造/物	复/合/结/构	1	0.27	复/合/结/构/物	2	0.1
剃刀/具	剃/刀	1	0.86	剃/刀/具	2	0.14
細胞/上澄み/液	胞/上/清/液	1	0.67	细/胞/上/清/液	2	0.33

た．一方，文献 [9] においては，日本語漢字と中国語簡体字への文字対象情報と既存の統計的機械翻訳モデルを用いて，日中特許対訳コーパスから対訳辞書を段階的に自動構築した．フレーズテーブルの利用において，句に基づくフレーズテーブルおよび階層的句に基づくフレーズテーブルの２種類のフレーズテーブルから，共通する訳語候補を抽出することより，訳語対を生成する．本研究と文献 [9] の間の最も大きな相違点として，本研究において，フレーズテーブルから得られた訳語候補のうち，日中対訳文対の中国語文に出現する訳語候補を抽出することより，訳語推定を行う点が挙げられる．

6　おわりに

本論文では，日中対訳特許文に対して，句に基づく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を用いて，専門用語の訳語推定する手法を提案した．提案手法では，句に基づく統計的機械翻訳モデルを用いることより，対訳特許文から学習されたフレーズテーブルを用いることによって，専門用語対訳対の獲得を行った．評価実験においては，97%程度の適合率およびF 値を達成した．今後は，中国語側の区切り単位として，形態素および文字の２種類の単位を併用し，Support Vector Machines (SVMs) [8] を用いることによって，フレーズテーブルから得られた訳語候補を検証し，性能を改善する方式について研究を進める．

謝辞

本研究においては，日本特許情報機構 (JAPIO) より提供して頂いた日中パテントファミリーのデータを利用させて頂いた．関係各位に感謝の意を表す．

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*, pp. 97–106, 2008.

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

[3] 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄. フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定. 電子情報通信学会論文誌, Vol. J93–D, No. 11, pp. 2525–2537, 2010.

[4] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[5] J. Sun and Y. Lepage. Statistical machine translation between unsegmented japanese and chinese texts. 言語処理学会第 19 回年次大会発表論文集, pp. 122–125, 2013.

[6] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pp. 168–171, 2005.

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

[8] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[9] K. Yasuda and E. Sumita. Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Computational Linguistics and Intelligent Text Processing*, Vol. 7817 of *LNCS*, pp. 276–284. Springer, 2013.