SVMを用いた 日中パテントファミリーからの対訳専門用語収集*

董 麗娟† 龍 梓† 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† 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であるため，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必要である．ここで，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．しかし，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．

そこで，本論文では，文献 [6] 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により，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を用いて日本語専門用語の中国語訳語推定を行う．最後に，獲得した日中対訳専門用語に対して，複数の対訳文から得られる情報を素性として，Support Vector Machines (SVMs) [7] を適用する．結果として，提案手法により，再現率 60%以上という条件のもとで，90%程度の適合率を達成した．

2 日中対訳特許文

本論文では，フレーズテーブルの訓練用データとして約 360 万対の日中対訳特許文を使用した. この日中対訳特許文は，2004-2012 年発行の日本公開特許広報全文と 2005-2010 年中国特許全文を対象として，文献 [6] の手法によって日中間で文を対応付け，スコア降順で上位の 360 万文対を抽出したものである．

3 句に基づく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．フレーズテーブルを作成する際の準備として，日本語文に対しては，MeCab[1] による形態素解析を行い，一形態素を単語の単位とする．一方，中国語文に対しては，Chinese Penn Treebank を用いた Stanford Word Segment [5] によって形態素解析を行い，一形態素を単語の単位とする．以上の準備を行った日中対訳文に対して，Moses を適用することにより，フレーズテーブルを作成する．その際，日本語句および中国語句の形態素数の上限をいずれも 15 とする．

4 一組の日中対訳文およびフレーズテーブルを用いた訳語推定

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の概要を図 1 に示す．日中対訳特許文およびフレーズテーブルを用いて，専門用語の訳語推定を行う．訳語推定手法において，一つの日中対訳文を対象として，その日中対訳文に出現する用語の訳語対 (用語対訳対) を推定する．具体的に，まず，全対訳文データ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 中に含まれる専門用語を t_J とする．次に，得られた日本語専門用語 t_J に対し，フレーズフレーズに存在し，かつ，得られた対訳文 $\langle S_J, S_C \rangle$ の中国語文 S_C に出現する訳語候補を抽出する．ここで，各日本語専門用語に対し，フレーズ

*Collecting Japanese-Chinese Translation of Technical Terms from Patent Families by SVM

†Lijuan Dong, Zi Long, Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba

‡Tomoharu Mitsuhashi, Japan Patent Information Organization (JAPIO)

[1] http://mecab.sourceforge.net

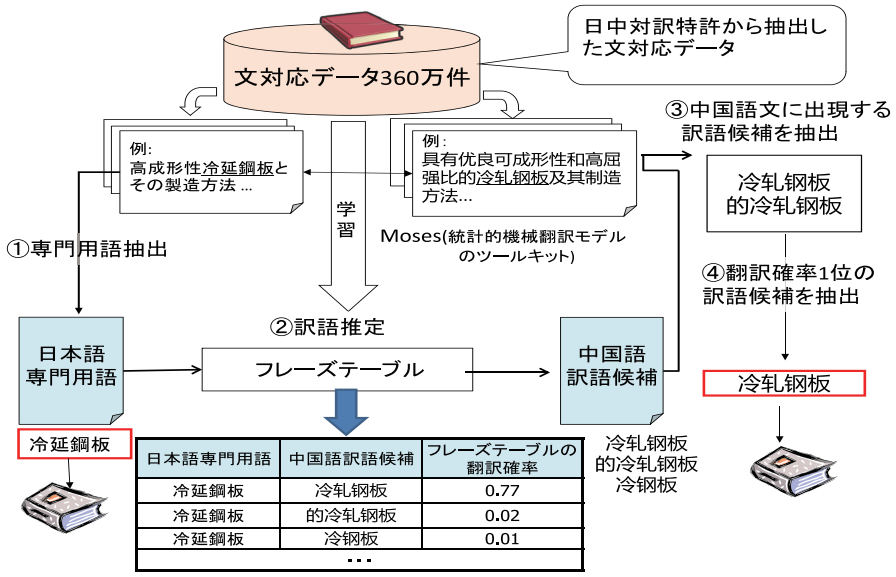

図 1: 対訳文およびフレーズテーブルを用いた対訳専門用語獲得の流れ

表 1: 日本語名詞句の分類および例

日本語名詞句の分類	例
専門用語	有機 化合 物 膜
	希土類 焼 結 磁石
一般名詞句	製造 方法
	化合 物
区切り位置誤り	二 次 加
	式 リフト クレーン

表 2: 訳語候補集合における正・負例数の内訳

正例	負例	総数
1,531	1,002	2,533

テーブルにおける翻訳確率 1 位の訳語候補を生成し，訳語推定結果とした．

5 複数の日中対訳文からの情報を素性とする SVM の適用

5.1 訳語推定対象の選定

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，訳語推定手法の評価を行う対象とする日本語名詞句を選定する．具体的に，まず，全対訳文データ 360 万件の日本語文を形態素解析し，1,244,480 万件の日本語名詞句を抽出する[2]．次に，全日本語名詞句に対して，13 の日本語名詞句の頻度レンジを設定し，各頻度レンジにごとに，無作為に 90 例 (合計 1,148 例)[3]の日本語名詞句を抽出する．これらの日本語名詞句 1,148 例を大別すると，表 1 に示すように，専門用語，および，それ以外の一般名詞句，区切り位置誤り，あるいはそれらの混在したものに分類することができる．最後に，日本語名詞句 1,148 例のうち，578 例の日本語専門用語のみを評価対象の日本語名詞句として選定した．

5.2 参照用対訳対集合の作成

前節で述べた手順によって得られた日本語専門用語 578 例に対して，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，2,533 例の日中対訳専門用語を生成した．次に，生成した 2,533 例の日中対訳専門用語に対して，専門用語の対訳対として適切か否かの判定を行い，表 2 に示すように，全 2,533 例中，正例を 1,531 例，負例を 1,002 例とした．

5.3 SVM の適用

前節で生成した 2,533 例の日中対訳専門用語を全事例集合として，互いに素な部分集合に 10 分割した．ただし，同一の日本語専門用語を共有する複数の日中対訳専門用語は，同一の部分集合に含めた．また，本論文では，TinySVM[4]を利用して，評価実験を行った．カーネル関数としては，一次多項式カーネルおよび二次多項式カーネルを評価し，相対的によい二次多項式カーネルを用いた．また，SVM の分離平面から評価事例までの距離を信頼度とし，正例判定における信頼度の下限を設定した．具体的には，10 個の部分集合のうち，8 個を訓練用事例集合として SVM の訓練を行い，残りのうちの 1 個を調整用事例集合とし，最後の 1 個を評価用事例集合とした．調整用事例集合を用いたパラメータの調整においては，分離平面から評価用事例までの距離の下限のパラメータの調整を行った[5]．

[2]日本語文の形態素解析結果に対して，名詞・接頭辞・接尾辞・未知語のいずれかの品詞の最長の形態素列に加えて，数字・アルファベットの列が連接することを許容したものを日本語名詞句として抽出する．

[3]頻度レンジ 10,000 以上の場合のみ，全体で 68 例のみである．

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

[5]本研究では，日中対訳専門用語の適合率および F 値を最大化する調整を行った．ただし，適合率を最大化する場合は，再現率が

表 3: 日中対訳専門用語同定のための素性

分類	素性名	定義
単言語素性	f_1: 日本語専門用語の頻度	日本語専門用語が属する頻度レンジの番号 (1~13)
	f_2: 中国語専門用語の頻度	中国語専門用語が属する頻度レンジの番号 (1~13)
二言語素性	f_3: 翻訳確率	フレーズテーブルにおける翻訳確率
	f_4: 訳語候補の順位 (翻訳確率の降順)	同一日本語専門用語に対する訳語候補の順位 (翻訳確率の降順)
	f_5: 日中対訳専門用語の頻度	日中対訳専門用語が属する頻度レンジの番号 (1~13)
	f_6: 日本語専門用語と対訳共起頻度の頻度差	日本語専門用語の頻度 − 日中対訳共起頻度が上限値 (本論文では 105) 以下の場合 1，上限値を超える場合 0
	f_7: 訳語数	同一の日本語専門用語に対する中国語訳語候補数
	f_8: 文単位の句対応制約の違反のない対訳文の割合	$f_8 = \frac{\text{文単位の句対応制約の違反のない対訳文対の数}}{\text{当該日中対訳専門用語の共起頻度}}$
	f_9: 要素合成法のスコア	要素合成法により出力された訳語候補のスコア

表 4: 評価結果 (%)

		適合率	再現率	F 値
ベースライン		60.4	100	75.3
SVM	適合率最大	91.8	59.7	72.3
	F 値最大	76.2	88.6	81.9

以上の訓練，調整，評価の手順を 10 通り繰り返し，その評価結果のマイクロ平均を算出し，日中対訳専門用語判定の性能評価を行った．

5.4 素性

本論文の手法において用いた素性は，表 3 に示すように，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，全 360 万対訳文における日本語専門用語 t_J の頻度 (f_1)，および，中国語専門用語 t_C の頻度 (f_2) を用いた．

二言語素性としては，フレーズテーブルによって各訳語候補に付与された翻訳確率の素性 (f_3)，および，同一日本語専門用語に対する訳語候補を翻訳確率の降順に順位付けした場合の順位の素性 (f_4) を用いた．その他，日中対訳専門用語 $\langle t_J, t_C \rangle$ の共起頻度の素性 (f_5) を用いた．また，日本語専門用語の頻度と日中対訳専門用語の共起頻度の差の素性 (f_6)，同一日本語専門用語に対する中国語訳語候補の数の素性 (f_7)，文単位の句対応制約 [4] の違反のない対訳文の割合の素性 (f_8) を用いた．さらに，日本語専門用語 t_J と中国語専門用語 t_C との間で，要素合成法に基づき，フレーズテーブルを用いて構成要素の間の翻訳を行った際の翻訳確率の積を求めた値を要素合成法のスコアの素性 (f_9) として用いた．ただし，フレーズテーブルを用いて構成要素を翻訳する際の翻訳確率に対して下限値 (本論文では 0.005) を設けるとともに，日本語専門用語 t_J および中国語専門用語 t_C を構成要素に分割する際に二通り以上の分割の仕方が可能な場合には，それぞれの分割の仕方における要素合成法スコアの和を用いた．

60%以上となるという条件のもとで，パラメータの調整を行った．

5.5 評価結果

表 4 に評価結果を示す．ベースラインとして，全事例が正しいと判定した場合，適合率は 60.4%，再現率は 100%，F 値は 75.3%となった．適合率を最大化する調整を行った場合の適合率は 91.8%，F 値を最大化する調整を行った場合の F 値は 81.9%となった．

表 5 に SVM によって選定された訳語候補の正解例および誤り例を示す．

表 5(a)「SVM による正解例」のうち，日本語専門用語「水性/樹脂/組成/物」および中国語訳語“水性/树脂/组合物”の組においては，フレーズテーブルにおける翻訳確率 (素性 f_3) は 0.95 であり，フレーズテーブルにおける中国語訳語の順位 (素性 f_4) は 1 位である．また，「水性/樹脂/組成/物」の訳語数 (素性 f_7) は 1 であり，要素合成法により出力された訳語候補のスコア (素性 f_9) は 0.73 である．これらの素性の効果によって，SVM により正解の対訳専門用語であると判定できた．一方，日本語専門用語「気/液/分離/器」および中国語訳語“气液/反应器”の組においては，フレーズテーブルにおける翻訳確率および要素合成法スコアの素性 (f_3，f_9) の値が十分に小さいことの効果によって，誤りの対訳専門用語であると判定できた．

表 5(b)「SVM による不正解例」のうち，日本語専門用語「生物/処理/反応/槽」および中国語訳語“生物/处理/反应/槽中”の組においては，素性 (f_3,f_4,f_7,f_9) が原因となって，誤って正解訳語と判定されてしまった．逆に，日本語専門用語「非/晶/質/シリコン」および中国語訳語“非晶质/硅”の組においては，要素合成法スコアの素性 (f_9) が 0 となったことが原因で，誤りの対訳専門用語であると判定されてしまった．これら二組の例においては，いずれも，中国語側の形態素解析の誤りが原因となって，SVM による判定結果が不正解となっている．一つ目の例においては，日本語専門用語中の構成要素「槽」に対応するべき中国語側の形態素は“槽”であるが，中国語文の形

表 5: SVM による正解例および不正解例

(a) SVM による正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_2	素性 f_3	素性 f_4	素性 f_7	素性 f_9	人手による判断	SVMによる判断
水性/樹脂/組成/物	水性/树脂/组合物	$11 \leqq jf \leqq 15$	$11 \leqq jcf \leqq 15$	0.95	1	1	0.73	正解	正解
気/液/分離/器	气液/反应器	$1{,}001 \leqq jf \leqq 10{,}000$	$jcf=1$	0.0008	9	13	0	誤り	誤り

(b) SVM による不正解例

日本語専門用語	中国語専門用語	素性 f_1	素性 f_2	素性 f_3	素性 f_4	素性 f_7	素性 f_9	人手による判断	SVMによる判断
生物/処理/反応/槽	生物/处理/反应/槽中	$21 \leqq jf \leqq 30$	$2 \leqq jcf \leqq 5$	0.06	3	5	0.05	誤り	正解
非/晶/質/シリコン	非晶质/硅	$501 \leqq jf \leqq 1{,}000$	$jcf=1$	0.005	7	13	0	正解	誤り

態素解析結果において，“中”と“槽”が分割されなかったため，中国語訳語が“生物/处理/反应/槽中” となってしまった．一方，二つ目の例では，日本語側が「非/晶/質」と一文字ごとに一形態素へと分割されたのに対して，中国語側が“非晶质” と三文字が一形態素に連結された形態素解析結果となったため，要素合成法において構成要素に分割して対応付けることができず，要素合成法のスコアが 0 となってしまった．これらの二例における誤りを回避するためには，中国語側において一文字を一単語として学習した文字単位フレーズテーブルを併用し，この文字単位フレーズテーブルを扱うための素性を新たに導入する必要があると考えられる．

6 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 [4, 1] においては，パテントファミリーから抽出した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，専門用語の訳語推定を行う．訳語推定手法においては，一つの対訳文を対象として，その対訳文に出現する用語の訳語対を推定する．一方，本論文においては，複数の対訳文から得られる素性を用いた SVM によって，高信頼度な日中間対訳専門用語を同定する手法を提案しており，文献 [4, 1] と比べると，これらの点が本論文の新規性となる．また，文献 [3] においては，中米パテントファミリーから対訳特許文を抽出して，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，および SVM を用いることによって，中英対訳対を獲得する．本論文と文献 [3] の間の最も大きな相違点として，本論文の手法においては，フレーズテーブルから得られるスコアの情報だけではなく，頻度の情報 (f_1，f_2，f_5，f_6)，訳語候補の順位 (f_4)，訳語数 (f_7)，文単位の句対応制約の情報 (f_8) を素性として用いて SVM を適用する点が挙げられる．

7 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，対訳専門用語を収集する手法を提案した．提案手法では，評価対象日本語名詞句に対して，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，日中対訳専門用語の候補を生成し，生成した日中対訳専門用語候補に対して，SVM を適用した．提案手法により，再現率 60%以上という条件のもとで，90%程度の適合率を達成した．今後は，中国語側の語の区切り単位として，形態素および文字の二種類の単位を併用した上で SVM を適用することによって，日中対訳専門用語同定の性能を改善する方式に取り組む．

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] 董麗娟, 龍梓, 豊田樹生, 宇津呂武仁, 三橋朋晴, 山本幹雄. 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. 言語処理学会第 20 回年次大会発表論文集, pp. 368–371, 2014.

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

[3] B. Lu and B. K. Tsou. Towards bilingual term extraction in comparable patents. In *Proc. 23rd PACLIC*, pp. 755–762, 2009.

[4] 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.

[5] 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.

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

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