対訳特許文を用いた同義対訳専門用語の推移的収集*

梁 冰† 豊田 樹生‡ 宇津呂 武仁§ 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 理工学群工学システム学類‡ 筑波大学 システム情報系§

1 はじめに

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り，これまでに，対訳特許文書を情報源として，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．森下らは，NTCIR-7の特許翻訳タスクで配布された日英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 [3]．この研究では，句に基づく統計的機械翻訳モデル [1] 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，要素合成法，Support Vector Machines (SVMs) [5] による機械学習を用いることによって，専門用語対訳対獲得を行った．しかし，森下らの手法では，ある日本語専門用語に対する英訳語を推定する際に，その日本語専門用語が出現する一つの対訳文に出現する英訳語のみを推定対象としているため，他の対訳文に出現している同義の専門用語対訳対を同定することができていない，という問題点があった．

そこで，先行研究 [2] では，ある日本語専門用語が出現する複数の対訳文を入力として，同義の専門用語対訳対を同定する手法を提案する．提案手法では，対訳特許文および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専門用語対訳対を収集し，それに対して，SVMを適用することにより，専門用語対訳対の同義・異義関係の判定を行う．この手法は，評価実験において，およそ98%の適合率と40%以上のF値を実現した．

しかし，高い適合率に対して再現率が低いという問題点も見られた．そこで，本論文では，再現率の改善方法として，同義対訳専門用語の推移的同定の枠組みを提案する．この枠組みでは，SVMによって高適合率で同義と判定された専門用語対訳対を新たな中心的対訳対として選定し，それらの同義集合の和集合を元の中心的対訳対の同義集合として出力するという手順を再帰的に行う．この手法に対して行った評価実験の結果，95%の適合率と32%の再現率を達成し，推移的同定の枠組みを適用しない場合と比べ，再現率が4%向上した．さらに，推移的同定の枠組みにおいて，人手の介在を併用する場合では，95%以上の適合率と50%以上の再現率を達成し，再現率をさらに20%改善することができた．

2 機械学習を用いた同義対訳専門用語の同定

2.1 適用手順

本論文では，先行研究 [2] の場合と同様に，まず，表1に示すように，134個の専門用語対訳対同義候補集合を生成した．そして，134個の専門用語対訳対同義候補集合 $CBP(s_J)$ を全事例集合 CBP とし，互いに素な事例部分集合 $CBP_i(i = 1, \ldots, 10)$ に10分割する[1]．本論文では，機械学習のツールキットであるTinySVM[2]を利用して，評価実験を行った．カーネル関数として，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 (すなわち，中心的対訳対と同義) 判定に下限閾値を設定した．訓練の手順について，$CBP_1, \ldots, CBP_{10}$ の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て2種類のパラメータの調整を行い，最後の1個を評価用事例集合とした．以上の手順を10通り繰り返し，その平均値を算出し同義判定の性能評価

*Transitive Collection of Bilingual Synonymous Technical Terms from Parallel Patent Sentences

†Bing Liang, Graduate School of Systems and Information Engineering, University of Tsukuba

‡Itsuki Toyota, College of Engineering Systems, University of Tsukuba

§Takehito Utsuro, Mikio Yamamoto, Faculty of Engineering, Information and Systems, University of Tsukuba

[1]ただし，ここでは，134個の中心的対訳対の集合を10個に分割した．その際，各 $CBP_i(i = 1, \ldots, 10)$ における正例 (中心的対訳対と同義)・負例 (中心的対訳対と異義) の数が，各 $CBP_i(i = 1, \ldots, 10)$ の間で均等になるように，中心的対訳対の集合を分割した．

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

表 1: 作成された専門用語対訳対の同義候補集合中の対訳対数

	総要素数	134 個の集合の間の平均対数
同義候補集合 $\bigcup_{s_J} CBP(s_J)$	22,473	167.7
人手で同定した同義集合 $\bigcup_{s_{JE}} SBP(s_{JE})$	1,680	12.5

表 2: 同義判定の性能評価 (%)

手法		適合率	再現率	F 値
ベースライン		67.0	54.3	60.8
SVM	適合率最大	**97.5**	28.7	43.9
	F 値最大	73.5	68.1	**70.5**

を行った．なお，本論文で調整の対象としたパラメータは，SVM のソフトマージンを制約するパラメータ，および，分離平面から評価用事例までの距離の下限閾値である．

2.2 同義・異義判定のための素性

同義専門用語対訳対の同定に用いた素性は大きく，対訳対 $\langle t_J, t_E \rangle$ の特性を規定するものおよび，対訳対 $\langle t_J, t_E \rangle$ と中心的対訳対 $\langle s_J, s_E \rangle$ の間の関係を規定するものの 2 種類に分けられる．

2.3 評価結果

表 2 に，同義判定における性能の評価結果を示す．ベースラインとしては，「t_J と s_J が同一，または，t_E と s_E が同一」という条件を用いた．距離下限閾値およびソフトマージンのパラメータに対して，同義判定の適合率を最大化する調整を行った場合は，97.5%の適合率と 43.9%の F 値を達成した．一方，距離下限閾値およびソフトマージンのパラメータに対して，同義判定の F 値を最大化する調整を行った場合は，適合率 73.5%，適合率 68.1%，F 値 70.5%を達成した．

3 同義対訳専門用語の推移的同定

SVM(2 節) による専門用語対訳対同義・異義自動同定の評価実験結果によって，適合率が高いものの，再現率が低い問題が存在していることが分かった．この問題を解決するため，SVM の同定結果に基づく推移的同定の枠組みを提案する．SVM により高適合率で同定した同義集合は，中心的対訳対との同義同定が相対的に容易な事例の集合と考えられる．そこで，このような高適合率での同義同定を漸進的に行うことにより，中心的対訳対との同義同定を直接行うことが困難な事例を同定することができ，同義同定の再現率の改善につながるというのが，この枠組みの基本的な考え方である．

3.1 推移的同定の手順

以下では，同義専門用語対訳対の推移的同定の手順を述べる．

ステップ 1 専門用語対訳対の同義候補集合 $CBP(s_{JE})$ の要素に対して，あらゆる $u_{JE} = \langle u_J, u_E \rangle$ と $v_{JE} = \langle v_J, v_E \rangle$ の組 (ただし，$u_{JE} \neq v_{JE}$) を作成し，それらの組に SVM(2 節) を適用し，同義・異義関係を判定する．

ステップ 2 それぞれの $u_{JE} = \langle u_J, u_E \rangle (\in CBP(s_{JE}))$ に対し，$u_{JE} = \langle u_J, u_E \rangle$ と同義の $v_{JE} = \langle v_J, v_E \rangle (\in CBP(s_{JE}))(\neq u_{JE})$ を集合 $X(u_{JE})$ の要素とする (図 1 (a)，図 2 (a))[3].

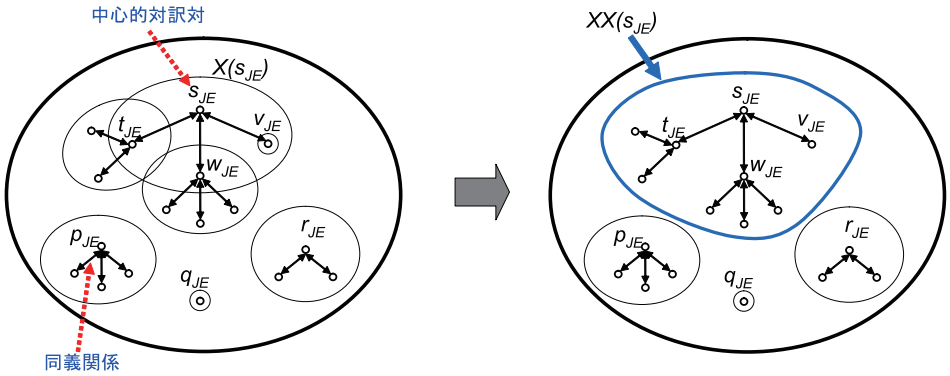
$$X(u_{JE}) = \Big\{ v_{JE} = \langle v_J, v_E \rangle \ (\in CBP(s_J)) \ \Big| \ v_{JE} = u_{JE}, \text{ または，SVM(2 節) により } v_{JE} \text{ と } u_{JE} \text{ を同義であると判定.} \Big\}$$

ステップ 3 このステップは，人手の介在を併用するか否かにより，以下の 2 つの方式に分けられる．

人手の介在を併用しない推移的同定 人手の介在を併用しない推移的同定は，複数の中心的対訳対間の同義・異義関係を判定する際，SVM による自動同定結果を利用する方式である．SVM(2 節) により中心的対訳対 s_{JE} と同義であると判定された専門用語対訳対 u_{JE} の集合を $SBP'(u_{JE})$ と定義する．

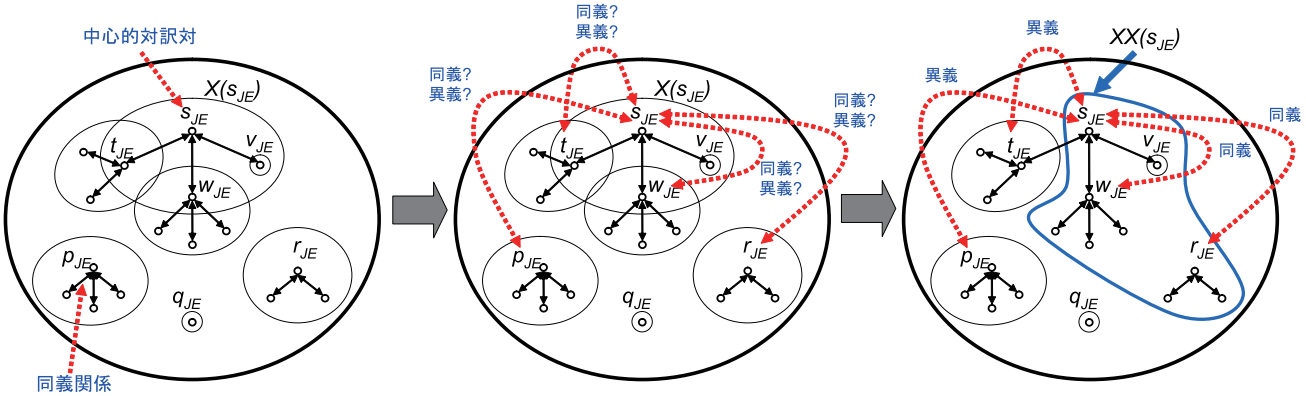
$$SBP'(s_{JE}) = \Big\{ u_{JE} = \langle u_J, u_E \rangle \ (\in CBP(s_J)) \ \Big| \ \text{SVM (2 節) により } u_{JE} \text{ と } s_{JE} \text{ を同義であると判定.} \Big\}$$

[3] ここで，v^1_{JE} および v^2_{JE} のいずれも，SVM により，u_{JE} と同義であると判定され，その一方で，v^1_{JE} と v^2_{JE} は異義であると判定される場合は，本論文では，v^1_{JE} と v^2_{JE} の両方を $X(v_{JE})$ の要素とする．



(a) SVMにより同義対訳対を同定　　(b) $X(u_{JE})$ の和集合を$XX(s_{JE})$として出力

図 1: 同義対訳専門用語の推移的同定手順 (人手の介在を併用しない)



(a) SVMにより同義対訳対を同定　　(b) $|X(u_{JE})|$ >1の場合のみ, u_{JE} が中心的対訳対と同義か否かを人手で判定　　(c) $X(u_{JE})$ の和集合を$XX(s_{JE})$として出力

図 2: 同義対訳専門用語の推移的同定手順 (人手の介在を併用)

また，SVM により，中心的対訳対 s_{JE} と同義であると判定した専門用語対訳対 u_{JE} に対して，ステップ 2 で定義した $X(u_{JE})$ の和集合を $XX(s_{JE})$ と定義する (図 1 (b)).

$$XX(s_{JE}) = \bigcup_{u_{JE} \in SBP'(s_{JE})} X(u_{JE})$$

この方式では，$XX(s_{JE})$ を中心的対訳対 s_{JE} の同義対訳専門用語集合として出力する．

人手の介在を併用した推移的同定 人手の介在を併用した推移的同定は，複数の中心的対訳対間の同義・異義関係を判定するとき，人手による判定を利用する方式である．

それぞれの専門用語対訳対 $u_{JE} = \langle u_J, u_E \rangle (\in CBP(s_{JE}))$ に対し，$|X(u_{JE})| > 1$ の場合のみ[4]，u_{JE} は中心的対訳対 s_{JE} と同義であるか否か (すなわち，$u_{JE} \in SBP(s_{JE})$) を人手で判定する (図 2 (b)).

また，人手により，中心的対訳対 s_{JE} と同義であると判定した専門用語対訳対 u_{JE} に対して，ステップ 2 で定義した $X(u_{JE})$ の和集合を $XX(s_{JE})$ と定義する (図 2 (c)).

$$XX(s_{JE}) = \bigcup_{\substack{u_{JE} \in SBP(s_{JE}) \\ |X(u_{JE})| > 1}} X(u_{JE})$$

この方式では，$XX(s_{JE})$ を中心的対訳対 s_{JE} の同義対訳専門用語集合として出力する．

[4]この条件は，SVM が少なくとも一つの専門用語対訳対 v_{JE} が u_{JE} と同義であると判定する場合に相当する．この条件が成り立たない場合は，u_{JE} が中心的対訳対 s_{JE} と同義であるか否かの人手による判定を行わない．

表 3: 同義対訳専門用語の推移的同定の評価結果 (%)

調整用事例における適合率の条件	適合率 / 再現率 / F 値		
	推移的同定なし	推移的同定あり	
		人手の介在を併用しない	人手の介在を併用
$> 80\%$	79.3 / 53.9 / 63.6	78.4 / **59.1** / **66.6**	81.3 / **89.9** / **85.1**
$> 85\%$	85.1 / 46.4 / 59.7	84.2 / **49.6** / **61.6**	86.9 / **80.9** / **83.4**
$> 90\%$	89.0 / 38.6 / 53.3	89.7 / **42.7** / **57.5**	91.3 / **69.1** / **78.2**
$> 95\%$	94.1 / 27.6 / 42.4	95.2 / **32.1** / **47.9**	95.2 / **53.1** / **67.9**

3.2 評価結果

本論文では，複数の中心的対訳対間の同義・異義関係を判定し，複数の同義対訳専門用語集合を統合することにより，同義同定の再現率を改善するという推移的同定の枠組みのもとで，評価実験を行った．具体的には，3.1 節で述べた方式を評価した．

2.1 節で述べたように，調整用事例集合を用いて，距離下限閾値を調整することにより，判定結果の適合率を変化させることができる．評価実験において，調整用事例における判定結果の適合率が 80%以上，85%以上，90%以上，95%以上のときのそれぞれの距離下限値を利用した場合の評価用事例における評価結果を表 3 に示す[5]．

全体として，人手の介在を併用なしの推移的同定の評価結果においては、推移的同定なしのときの評価結果と比べ，再現率を平均 4%以上改善した．さらに，人手の介在を併用した推移的同定の評価結果においては，人手の介在を併用しない推移的同定の評価結果と比べ，適合率は平均 2%以上向上し，再現率はさらに平均 30%改善された．しかし，人手の介在を併用しない推移的同定方式においては，再現率の増加は一サイクル目の推移的同定において最大となり (表 3 の評価結果に示した結果)，それ以降のサイクルにおいて，再現率をさらに改善することができなかった．一方，人手の介在を併用した推移的同定方式は，高い適合率を保ちながら，再現率を大幅に改善した．言い換えると，この再現率は推移的同定という枠組みの現段階における再現率の上限値であるといえる．

[5]参考として，各参照用専門用語対訳対の同義集合 $SBP(s_{JE})$ の要素 u_{JE} のうち，$\left|X(u_{JE})\right| = 1$ の平均要素数を測定した．この数は，実際，どのくらいの要素 u_{JE} に対して，中心的対訳対 s_{JE} と同義であるか否かの判定を行う必要がないかを表す．一つの中心的対訳対あたりの参照用同義対訳専門用語の数は 12.5 個であるが，表 3 に示す結果においては，この数は，それぞれ，“$> 80\%$” の場合は 0.9，“$> 85\%$” の場合は 1.4，“$> 90\%$” の場合は 2.2，“$> 95\%$” の場合は 4.0 となった．

4 関連研究

文献 [4] は，対訳専門用語の同義判定に機械学習を用いており，手法の点においても，また，機械学習で用いている素性の点においても，本論文の手法と密接に関連している．しかし，文献 [4] では，同義判定の対象とする対訳専門用語の収集を手動で行っており，手法の適用範囲が非常に限定されている．一方，本論文の手法は，毎年に公開される対訳特許テキストから，同義判定の対象とする対訳専門用語の収集を自動で行っており，文献 [4] と比較して，手法の適用範囲が限定されないという点で，優れていると言える．

5 おわりに

本論文では，同義対訳専門用語の自動同定において再現率が低いという問題点を改善するため，推移的同定の枠組みを構築した．評価実験において，95%以上の適合率と 32%の再現率を達成し，再現率を 4%改善した．さらに，新たな中心的対訳対を選定する際に，人手の介在を併用した場合では，95%以上の適合率と 50%以上の再現率を達成し，再現率をさらに 20%改善した．今後の課題としては，中心的対訳対の同義候補集合の生成 (文献 [2] を参照) の過程を再帰的に行う方式を開発することが重要であると考えられる．

参考文献

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[2] 梁冰, 宇津呂武仁, 山本幹雄. 対訳特許文を用いた同義対訳専門用語の同定と収集. 言語処理学会第 17 回年次大会論文集, pp. 963–966, March 2011.

[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] T. Tsunakawa and J. Tsujii. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pp. 457–464, 2008.

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