LE-003

サポートベクターマシンを用いた対訳辞書登録候補の自動選別

Selection of Dictionary Entries from Bilingual Pairs by using Support Vector Machines

九津見毅* 吉見毅彦† 小谷克則‡ 佐田いち子* 井佐原均‡
T. Kutsumi T. Yoshimi K. Kotani I. Sata H. Isahara

1 はじめに

英日機械翻訳システムなどの対訳辞書を拡張するための手段の一つとして，対訳コーパスなどから語彙知識を自動的に獲得する方法が有望である．適切な語彙知識を獲得するためには，(1) 対訳コーパスにおいて英語表現と日本語表現を正しく対応付ける処理と，(2) 対応付けられた英日表現対を辞書に登録するか否かを判定する処理の二つが必要である．後者の処理が必要な理由は，対応付けられた英日表現対には，辞書に登録することによって翻訳品質が向上することがほぼ確実なものとそうでないものがあるため，これらを選別する必要があるからである．例えば，対訳コーパスから次のような英日表現対の対応付けが得られたとする．

Customs and Tariff Bureau	関税局
Minshuto and New Komeito	民主党や公明党
Miyagi and Yamagata	宮城，山形両県

これらのうち第一の英日表現対は辞書に登録すべきであるが，第二，第三の英日表現対はそうではない．なぜならば，“Minshuto and New Komeito” を我々の機械翻訳システム*1で処理すると「民主党，及び，公明党」という翻訳が得られるが，この翻訳と「民主党や公明党」とでは翻訳品質に大きな差はないと判断できるからである．また，第三の英日表現対は，“Miyagi” と “Yamagata” が県名を表わしていない文脈では不適切となり，文脈依存性が高いからである．このように，翻訳品質が変化しなかったり，低下することが予想されたりする英日表現対はふるい落とさなければならない．

我々が英日表現対の対応付けと選別を分けて考えるもう一つの理由は，前者はシステム依存性が低いのに対して，後者は依存性が高いという違いがあるからである．対応付けが正しいか否かは個々の機械翻訳システムにほとんど依存しない．このため，正しい対応付けを得るための判定基準を設定する際には特定のシステムを想定する必要がない．これに対して，対応付けられた英日表現対 (辞書登録候補) を登録するべきか否かは個々の機械翻訳システムに依存するため，選別は，特定の機械翻訳システムを想定した判定基準に基づいて行なわれなければならない．例えば，我々の機械翻訳システムには “the Bank for *ABC*” を「*ABC* 銀行」のように訳す (前置詞 “for” を訳出しない) 規則が存在しない．このため，“the Bank for International Settlements” が「国際決済のための銀行」と訳されてしまう．従って，我々のシステムの場合はこの英語名詞句と「国際決済銀行」の対を辞書に登録すると判定するのが妥当である．しかし，もし前置詞 “for” を訳出しないという規則を持つシステムが存在すれば，そのシステムにとっては登録する必要がないと判定するのが妥当であろう．従って，対応付けと選別とでは異なる正解判定基準を導入する必要がある．

従来の研究では，異なる言語の表現同士を正しく対応付けることに焦点が当てられていることが多く [1, 2, 3, 4, 5, 6, 7, 8]，(正しく) 対応付けられた表現対を辞書に登録するか否かを判定する処理について，選別のシステム依存性を意識した上で明確に議論した研究はほとんど見当たらない．

本稿では，対訳辞書に登録する目的で収集された英日表現対のうち，前置詞句と等位構造の両方または一方を持つ英語固有名詞句 (以下では単に英語名詞句と呼ぶ) とそれに対応する日本語名詞句を辞書登録候補とし，この辞書登録候補を自動的に選別して適切な語彙知識を獲得する方法を提案する．提案手法ではサポートベクトルマシンによる機械学習を利用する．

2 着目した素性

学習に用いる素性を決定するために，まず，人間の辞書開発者が候補の選別をどのように行なっているかについて考える．ある英語名詞句とそれに対応する日本語表現 (以下では新規訳と呼ぶ) から成る辞書登録候補を辞書に登録するか否かを判定する際に辞書開発者は開発に携わっているシステムの特性 (辞書や規則など) を考慮に入れつつ様々な観点から検討を加え，最終的な判断を下している．そのうち最も重要な判断基準の一つは，辞書登録候

* シャープ (株)，SHARP
† 龍谷大学 / 情報通信研究機構，Ryukoku University / NICT
‡ 情報通信研究機構，NICT
*1 シャープ (株) の英日翻訳支援ソフトウェア「翻訳これ一本」の開発段階のバージョン．

補を登録した場合それに値するだけの改善が翻訳品質に見られるかどうかであろう．もし十分な品質向上が達成できると辞書開発者が考えればその辞書登録候補を登録すると判定し，そうでなければ登録しないと判定する．辞書登録候補を登録することによって達成される改善の度合いは，その辞書登録候補が登録されていない状態での辞書を用いて英語名詞句を翻訳した結果 (以下では既存訳と呼ぶ) と新規訳を比較することによって見極めることができる．

本研究では，辞書開発者のこのような作業を機械的に模倣し，既存訳と新規訳を比較して得られる差異に着目して辞書登録候補を選別することを試みる．具体的には，既存訳と新規訳で異なる部分 (差分部分) と両者に共通する部分 (共通部分) が辞書登録候補の選別に影響しうる要因であると考え，それらを素性とする．

既存訳と新規訳の差分部分と共通部分を表現する手段としては，表記情報 (形態素)，品詞情報，意味情報などが挙げられる．まず，形態素による表現について述べる．例えば “Special Committee on Medical Devices” という英語名詞句が辞書に登録されておらず，この英語名詞句の既存訳が「医療用具上の特別委員会」であり，新規訳が「医療用具特別部会」であるとする．このとき，「医療用具上の特別委員会」と「医療用具特別部会」に対して茶筌[*2]によって形態素解析を行ない，両者の差分部分と共通部分を，文字を単位として差分検出ツール mdiff [*3]によって求めて出力形式を若干変更すると，図1のような結果が得られる．従って，「医療用具上の特別委員会」と「医療用具特別部会」は，図1に示す四つの素性 same(医療/用具)，diff(上/の, NIL)，same(特別)，diff(委員/会, 部会) の値を 1，その他の素性の値を 0 とする素性ベクトルに写像される．なお，diff(A, B) は A と B が差分部分であることを表わし，same(C) は C が共通部分であることを表わす．また，NIL は対応する部分が存在しないことを意味し，記号 ‘/’ は形態素の区切りを表わす．

same(医療/用具)
diff(上/の, NIL)
same(特別)
diff(委員/会, 部会)

図1　形態素単位での差分・共通部分

既存訳と新規訳の差分部分や共通部分を表わす素性として，品詞情報や意味情報などの，表記よりも抽象化された (粒度が粗い) 情報を利用することも考えられる．既存訳「医療用具上の特別委員会」と新規訳「医療用具特別部会」に対して茶筌の品詞単位で差分部分と共通部分を求めた結果を図2に示す．図1と図2を比べると，形態素による表現では「特別」が共通部分であり「委員/会」と「部会」が差分部分であると解釈されていたのに対して，品詞による表現では「特別/委員」と「特別/部会」の品詞が共通部分であり，既存訳の「会」の品詞に対応する品詞が新規訳には存在しないと解釈されている．

same(名詞-一般/名詞-一般)
diff(名詞-接尾-副詞可能/助詞-連体化, NIL)
same(名詞-形容動詞語幹/名詞-一般)
diff(名詞-接尾-一般, NIL)

図2　品詞単位での差分・共通部分

意味情報による表現では，EDR 日本語単語辞書[*4] に記述されている概念識別子に表記 (形態素) を写像する．このとき，概念識別子の曖昧性は考慮しない．すなわち，ある形態素に対して概念識別子が二つ以上存在する場合それらの中から無作為に概念識別子を一つ選ぶ．このように曖昧性の解消を放棄せざるを得ない理由は，処理対象の表現が文脈から切り離されているため，曖昧性解消に必要な情報が十分には得られないことにある．EDR 辞書から概念識別子が得られなかった場合は，未定義 (undef) とする．なお，既存訳や新規訳の構成要素のうち茶筌品詞が名詞と未知語であるもののみを概念識別子への写像の対象とし，それ以外の構成要素は削除する．既存訳「医療用具上の特別委員会」と新規訳「医療用具特別部会」の差分部分と共通部分を，概念識別子で表現した場合の素性を図3に示す．

same(0fe1dd/3cedca)
diff(1eb357, NIL)
same(2016ed)
diff(3bcaa4/3ceda8, 107777)

図3　概念識別子単位での差分・共通部分

これまでに述べた差分部分と共通部分の表現では，差分部分や共通部分の出現順序が考慮されていない．そこで，差分部分や共通部分の出現順序を考慮するためにこれらの N グラム ($N = 2, 3$) を素性として利用する．

また，差分部分と共通部分を表わす素性として，1 グラムと 2 グラムを合成したものや 1 グラムと 2 グラムと 3 グラムを合成したものを利用することも考えられる．

まとめると，本稿では，形態素，品詞，概念識別子という選択肢と N グラム ($N = 1$，$N = 2$，$N = 3$，$N \leq 2$，$N \leq 3$) の選択肢の組み合わせについて，それぞれどれくらいの精度で辞書登録候補の選別が行なえるのかを検証する．

[*2] http://chasen.aist-nara.ac.jp/chasen/

[*3] http://www2.nict.go.jp/jt/a132/member/murata/software/mdiff/mdiff.html

[*4] http://www2.nict.go.jp/kk/e416/EDR/J_index.html

3 訓練事例集

訓練事例集の作成では，それに必要な労力を省くために，我々が保有している既存の言語資源を利用した．

3.1 正例の作成

我々の機械翻訳システムの対訳辞書に登録されている英日表現対は，当然，これまでの辞書開発過程において辞書開発者によって登録すると判定されたものである．[*5] 従って，このような英日表現対における日本語表現を新規訳とみなすことができる．また，この新規訳に対する既存訳は，この英日表現対をシステムの対訳辞書から削除した状態で英語名詞句を翻訳すれば得ることができる．

今回の実験で利用する事例は，英語名詞句 NP と既存訳 CT と新規訳 NT の三つ組 $< NP, CT, NT >$ のうち既存訳と新規訳の部分だけに着目して，既存訳と新規訳の対を取り出したものである．このため，<Dept. of Transport, トランスポートの部門, 運輸省 > と <Department of Transport, トランスポートの部門, 運輸省 > のように英語名詞句は異なるが既存訳と新規訳の部分は同じである三つ組から既存訳と新規訳の対を取り出すと，二つの事例が重複する．このような場合，重複は許さず，事例を一つだけ事例集に含めることにする．

英語名詞句と新規訳の英日表現対をシステムの対訳辞書から削除しても英語名詞句の翻訳として新規訳が得られることがある[*6]．このような場合には，既存訳と新規訳の間に差分が全く生じない．既存訳と新規訳の間に差分がないような英日表現対は翻訳品質の向上に貢献しておらず，このような対を正例とみなすのは適切ではない．このため，事例集には含めないことにする．

3.2 負例の作成

我々は，これまでの辞書開発過程で候補には挙がったが辞書に登録されなかった英日表現対の一覧表を保有している．このような一覧表に掲載されている日本語表現は負例における新規訳とみなすことができる．また，この新規訳に対する既存訳は現状の機械翻訳システムで英語名詞句を翻訳して得られる日本語表現である．

正例作成の場合と同じく，英語名詞句と既存訳と新規訳の三つ組としては異なるが既存訳と新規訳の部分は同じであるものは，一つだけを事例集に含める．

正例の場合，差分が全くない既存訳と新規訳の対は事例集に含めないが，負例の場合には，差分がないことが辞書に登録しないことの理由であると考えられるため，事例集に含める．

負例として事例集に含めようとしている既存訳と新規訳の対が，既に正例として存在している場合，この対は事例集に含めないことにする．これは，既に辞書に登録済みであるため，重複登録を避ける目的で登録しないと判断された可能性があるからである．

3.3 訓練事例集のシステム依存性

本研究で作成した訓練事例集は，我々の機械翻訳システムの特性 (辞書や規則など) を反映したものであるため，他のシステムの対訳辞書の拡張に直接利用することは望ましくない．また，本研究で扱っている選別問題において一般的に利用可能な訓練事例集を見つけることは容易ではないであろう．しかしながら，この点は問題にならないと考える．なぜならば，我々が利用した言語資源と同様の資源は，機械翻訳システムの研究開発に携わる他の組織にも存在する可能性が高いため，その組織で開発されている機械翻訳システムの特性に合わせて訓練事例集を作成できるからである．このように，対象システムの辞書開発時の経験に基づいてそのシステムに適した訓練事例集を作成するという方針は，辞書登録候補の選別は個々のシステムに依存するものであり一般的に論じることは必ずしも適切ではないという，1節で述べた考えに基づくものである．

4 実験と考察

サポートベクトルマシンによる機械学習には TinySVM [*7] を利用した．カーネル関数は一次の多項式とした．いずれの実験でも五分割の交差検定を行なった．評価には F 値を用いた．ただし，再現率 (辞書登録すべき英日表現対のうち，正しくそのように判定されたものの割合) よりも適合率 (辞書登録すべきでない英日表現対のうち，正しくそのように判定されたものの割合) を重要視することにし，$\beta = 0.5$ とした．実験に用いた訓練事例集における正例は 10154 件，負例は 8878 件である．

4.1 差分・共通部分の表現方法と選別性能

既存訳と新規訳の間の差分部分と共通部分を形態素，品詞，概念識別子と N グラム ($N = 1$, $N = 2$, $N = 3$, $N \leq 2$, $N \leq 3$) の各組み合わせで表現したときの選別性能を表 1 に示す．数値は五分割の交差検定の平均値である．表 1 から，全体で最も高い選別性能を示す表現方法は形態素 1 グラムであることが分かる．

4.1.1 N グラムごとの比較

差分部分や共通部分の出現順序を考慮すると，選別性能にどのような影響が出るのかを検証する．1 グラム ($N = 1$)，2 グラム ($N = 2$)，3 グラム ($N = 3$) を用いた場合の各平均値を比較すると，1 グラムの場合 (0.752) よりも，2 グラムの場合 (0.594)，3 グラムの場合 (0.631) のほうが性能が大きく低下している．1 グラムの場合と 2 グラムの場合をより詳しく比較すると，品詞による表現を除き，2 グラムより 1 グラムのほうが性能が高い．また，1 グラムの場合と 3 グラムの場合をより詳しく見ると，品詞

[*5] この際の辞書開発者による判断は，ある英日表現対を登録した際の悪影響が生じる恐れを多くの観点から検討された結果 (たとえば，登録しようとする表現を含むより大きな表現の翻訳結果が不適切にならないか，等) であると考えてよい．

[*6] このような場合，対訳辞書に登録されている英語名詞句と新規訳の対は，翻訳品質の観点からは冗長な登録である．

[*7] http://chasen.org/ taku/software/TinySVM/

表 1　各表現での選別性能 (F 値)

	$N=1$	$N=2$	$N=3$	$N\leq 2$	$N\leq 3$	平均
形態素	0.803	0.515	0.592	0.798	0.796	0.701
品詞	0.705	0.792	0.708	0.743	0.739	0.737
概念識別子	0.749	0.475	0.594	0.744	0.744	0.661
平均	0.752	0.594	0.631	0.762	0.760	

による表現のときに 3 グラムのほうが僅かに高いだけで，それ以外の表現のときには 1 グラムのほうが高いことが分かる．これらのことより，N グラム ($N=2,3$) を単独で用いることが選別性能の向上につながるとは限らないと言える．

1 グラムだけを用いた場合 ($N=1$) の平均値，1 グラムと 2 グラムを合成した場合 ($N\leq 2$) の平均値，1 グラムと 2 グラムと 3 グラムを合成した場合 ($N\leq 3$) の平均値は，それぞれ 0.752，0.762，0.760 であり，その差はあまり大きくない．1 グラムだけを用いた場合と，1 グラムと 2 グラムを合成した場合をより詳しく比べると，品詞による表現による表現のときには後者のほうが性能が高いが，形態素，概念識別子による表現のときには前者のほうが高い．また，1 グラムだけを用いた場合と，1 グラムと 2 グラムと 3 グラムを合成した場合を比較しても同様である．これらのことより，N グラムを合成することが選別性能の大幅な向上にはつながっていないと言える．

4.1.2 形態素，品詞，概念識別子での比較

差分部分や共通部分を形態素，品詞，概念識別子のそれぞれで表現した場合の選別性能を比較する．形態素，品詞，概念識別子による表現での各平均値を比べると，品詞で表現した場合に最も高い選別性能が得られていることが分かる．品詞による表現での差分情報の粒度は，それ以外の表現での粒度よりも粗い．このため品詞による表現での選別性能が最も悪くなるだろうと当初予想していたが，平均値で比較する限りはこの予想に反する結果となった．ただし，1 グラムの場合で比較すると，差分情報の粒度が細かい表現方法 (形態素) から粗い表現方法への順で性能が低下している．

形態素，概念識別子による表現では，2 グラムと 3 グラムで性能の大幅な低下が見られるが，品詞による表現では，そのような低下は見られず，比較的安定した性能を示していることが分かる．

5 おわりに

機械翻訳システムなどで必要とされる語彙知識を獲得するためには，対訳コーパスにおいて二言語の表現を正しく対応付ける処理と，対応付けられた表現対を辞書に登録するか否かを判定する選別処理の二つが必要であるが，対応付けと選別は特定のシステムへの依存性に関して性質の異なる問題である．本稿では，このような点を指摘し，従来あまり扱われてこなかった辞書登録候補の選別問題を採り上げ，この問題を機械学習によって解く方法を示した．学習に用いる素性として，既存訳と新規訳で異なる部分と両者に共通する部分に着目し，差分部分や共通部分を表現する手段として，表記 (形態素)，品詞，概念識別子を用いた．さらに，差分部分や共通部分の出現順序を考慮するためにこれらの N グラムを導入した．評価実験の結果，最も高い選別性能を示す表現方法は形態素 1 グラムであることが明らかになった．

今後の課題の一つとして，このような対訳表現対の機械的選別を，実際の辞書開発作業に導入することによる効率や品質への影響を調査することを検討している．

参考文献

[1] N. Ayan, B. Dorr, and N. Habash. Multi-Align: Combining Linguistic and Statistical Techniques to Improve Alignments for Adaptable MT. In *Proc. of AMTA*, pp. 17–26, 2004.

[2] 出羽達也. 対訳文書から自動抽出した用語対訳による機械翻訳の訳語精度向上. 電子情報通信学会論文誌, Vol. J87-D-II, No. 6, pp. 1244–1251, 2004.

[3] F. Sadat, M. Yoshikawa, and S. Uemura. Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval. In *Proc. of ACL*, pp. 141–144, 2003.

[4] M. Sahlgren. Automatic Bilingual Lexicon Acquisition Using Random Indexing of Aligned Bilingual Data. In *Proc. of LREC*, pp. 1289–1292, 2004.

[5] 佐藤健吾, 斎藤博昭. サポートベクタマシンを用いた対訳表現の抽出. 自然言語処理, Vol. 10, No. 4, pp. 109–124, 2003.

[6] D. Tufis. A Cheap and Fast Way to Build Useful Translation Lexicons. In *Proc. of COLING*, pp. 1030–1036, 2002.

[7] T. Utsuro, T. Horiuchi, and T. Chiba, Y.and Hamamoto. Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites. In *Proc. of AMTA*, pp. 165–176. 2002.

[8] K. Yamamoto, T. Kudo, Y. Tsuboi, and Y. Matsumoto. Learning Sequence-to-Sequence Correspondences from Parallel Corpora via Sequential Pattern Mining. In *Proc. of HLT-NAACL Workshop*, pp. 73–80, 2003.