サポートベクターマシンを用いた対訳表現の機械翻訳辞書登録適切性の自動判定

九津見 毅† 吉見 毅彦‡/†† 小谷 克則†† 佐田 いち子† 井佐原 均††

†シャープ(株) ‡龍谷大学 ††情報通信研究機構

1. はじめに

近年、対訳コーパスを利用して、対訳辞書や対訳表現集などの、翻訳に有用な情報を獲得するための研究が盛んになってきている[1, 2, 3, 4, 5, 6]。しかし、対訳コーパスから獲得できた対訳用語や対訳表現などを、機械翻訳システム用の辞書や翻訳規則に利用することを考えた場合、得られた表現が対訳として正しくても、それをシステムの辞書や規則に登録すべきかどうかを判断する必要がある。

たとえば、コーパスから

（ａ）Japan and U.S.

（ア）日本とアメリカ

といった、辞書未登録の英日対訳表現が獲得できたとしても、一般的な英日機械翻訳システムだと（ａ）という表現はたとえば「日本及び米国」のように訳され、それで概ね正しいことから、上記対訳を辞書に追加登録する必要性は乏しいと言える。

一方、

(b) Japanese Foundation for Research and Promotion of Endoscopy

（イ）内視鏡医学研究振興財団

のような対訳が得られたとすると、これは辞書に登録する必要があると言える。この対訳情報が未知の状態で (b) を機械翻訳したとすると例えば「内視鏡検査の研究及び促進の日本の基金」のようになり、（イ）のような的確な表現を構成的に得ることが難しいからである。

従来の研究では、対訳語句の正しい対応付けに焦点を当てたものや[1, 2, 3, 4]、正しい対応付けとその後のブラッシュアップ（たとえば、同義表現の獲得など[7]）までを含めた獲得方法を述べたものが存在する。

本研究では、対応付けが正しいと考えられる対訳語句を辞書に登録するか否かを判定する選別処理に着目する。これは、実際の機械翻訳システム用辞書の開発において、対訳として正しい組であっても、その表現が本当に必要か、ある程度の汎用性があるか、その表現を別の文中に埋め込んで自然に読めるか、などの観点から、翻訳用辞書としてふさわしい表現のものを選別したり表記の手直しを行ったりすることが、機械翻訳システムの品質に大きく影響を及ぼす重要な工程だからである。[1]

従来こういった判断は各機械翻訳システム開発者の経験的なノウハウに委ねられてきた。本研究では、機械学習を用いてこの判断を自動化することにより、機械翻訳システム開発者を支援することを目指す。

学習用データとしては、実際の商用機械翻訳システム開発の過程において辞書開発者によって登録すると判断された英日表現対と、登録しないと判断された表現対とを利用する。機械学習の手段としてはサポートベクターマシンを用いる。学習のための素性としては、英語表現を現状の機械翻訳システムで翻訳した結果(現状訳)と、この英語表現に対応する日本語表現(新規訳候補)との差分情報を用いる。実験結果から、選別が誤りとなった例を分析し、誤りの原因を考察する。

２．手法の説明

２．１　利用するデータ

利用するデータは、英日機械翻訳システムの辞書開発の過程で収集された、英語見出し候補と新規訳候補の組を集積したものである。各組は結果的に辞書に登録すると辞書開発者によって判断されたか否かで分類される。

[1] 専門性の高い分野の用語においては、表現が定式化しているので、正しい対応付けさえ行われれば表現の選別・手直しの必要性が薄い場合もあるが[1, 2 など]、時事用語などを対象とする場合はこの工程は重要である。

辞書に登録すると判断された事例を以下に示す。

(英語見出し候補) Special Committee on Medical Devices

(現状訳) 医療用具上の特別委員会

(新規訳候補) 医療用具特別部会

現状訳は、新規訳候補を辞書に登録する前のシステムで英語見出し候補を翻訳して得られる結果である。辞書開発者は、現状訳と新規訳候補を比較して、新規訳候補を辞書に登録することによって翻訳品質の向上が期待できると判断したものと推測される。

他方、辞書に登録することが望ましくないと我々が判断する事例には次のようなものがある。

(英語見出し候補) United Kingdom and Scandinavia

(現状訳) 英国、及び、スカンジナビア

(新規訳候補) 英国スカンジナビア経済同盟

２．２　利用する素性

機械学習においては、対象となる対訳候補の辞書登録が行われたか否かという状態を、特徴ベクターに付与される正／負のラベルとして用いる。素性には、現状訳と新規訳候補の差分情報を用いる。

素性の求め方の例を以下に示す。まず、現状訳と新規訳候補を形態素解析する。本研究では形態素解析手段に「茶筌」[2]を用いた。

医療/用具/上/の/特別/委員/会

医療/用具/特別/部会

その結果から、下記に示すように、形態素そのもの、品詞情報、意味情報の差分を求める。

(a) 形態素そのもの

形態素そのもので特徴ベクターを構成した例を次に示す。

```
+1
(1) same   医療/用具
(2) diff   上/の;NULL
(3) same   特別
(4) diff   委員/会;部会
```

ここで、最上行の「+1」は、正のラベルであり、この対訳候補が辞書登録されたものであることを表している。また、左欄の「same」は、現状訳と新規訳候補の各々の形態素に由来する情報が同一であることを示し、「diff」は、前記の各々の形態素に由来する情報が異なる (「;」の左側が現状訳に由来、右側が新規訳候補に由来) ことを示している。

(b) 品詞情報

品詞情報での特徴ベクターの例を示す。素性(3)は、形態素そのものは「特別/委員」と「特別/部会」で両者異なるが、品詞情報は「委員」と「部会」から同一の「名詞-一般」が得られたため、この範囲での品詞情報は同一となった。

```
+1
(1) same   名詞-一般/名詞-一般(医療/用具)
(2) diff   名詞-接尾-副詞可能/助詞-連体化(上/の);NULL
(3) same   名詞-形容動詞語幹/名詞-一般(特別/[委員|部会])
(4) diff   名詞-接尾-一般(会);NULL
```

ここで、品詞情報は「茶筌」での形態素解析により獲得されたものを用いた。また、本項及び後述の(c)における ( ) 内の形態素は、説明のために、特徴ベクターの素性を構成する情報（品詞や意味）が由来する形態素を示したもので、特徴ベクターの素性そのものではない。

(c) 意味情報

意味情報は、EDR 電子化辞書[3]の概念識別子で表わす。意味情報での特徴ベクターの例を示す。

```
+1
(1) same   0fe1dd/3cedca(医療/用具)
(2) diff   1eb357/undef(上/の);NULL
(3) same   2016ed(特別)
(4) diff   3bcaa4/3ceda8(委員/会);107777(部会)
```

３．実験と考察

３．１　実験方法

実験には、シャープ(株)が開発している英日機械翻訳システムの開発過程において収集されたデータを用いた。

辞書登録候補の語句のうち、前置詞句または並列表現を含む固有表現の名詞句を対象とした。これらのうち、辞書に登録された対訳は 11,545 組、登録されなかった対訳は 9,761 組である。

上記データを５分割交差検定法で SVM 学習・評価し、辞書登録候補の選別結果と正解との合致状況を調べた。

機械翻訳手段はシャープの英日機械翻訳システム[4]、学習ツールには TinySVM[5] をそれぞれ用いた。

[2] http://chasen.naist.jp/

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

[4] http://www.sharp.co.jp/ej/

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

３．２　実験結果

５分割交差検定を行った結果、平均の適合率・再現率・Ｆ値は表１の通りである。Ｆ値を求めた結果からは、形態素そのものを特徴ベクターの素性としたものが最も性能が良かった。

	適合率	再現率	Ｆ値
形態素	0.881	0.708	0.785
品詞	0.704	0.807	0.752
意味	0.823	0.618	0.706

表 1　実験結果

３．３　辞書登録必要・不要な対訳表現の特徴

形態素そのものを特徴ベクターの素性とした場合について、各対訳候補の正解ラベルと選別結果の別に、出現の多かった素性を、以下に示す。出現数は、５分割した検定結果から得られた合計数である。表 2 は、正解ラベルが「登録」である事例をシステムが正しく「登録」と判断した事例に現れる素性である。

(1)　正解：登録／システム：登録

順位	素性		出現数
1	diff	の;NULL	924
2	same	協会	555
3	same	日本	418
4	same	委員/会	369
5	same	法	278
6	same	米国	227
7	diff	NULL;国際	181
8	diff	の/ため/の;NULL	167
9	same	研究	163
10	same	国際	154

表 2

(2) 正解：非登録／システム：非登録

順位	素性		出現数
1	same	の	703
2	same	、	218
3	diff	NULL;、	104
4	diff	の;NULL	86
5	diff	、;NULL	76
6	diff	NULL;の	75
7	same	日本	51
8	same	科学	44
9	diff	NULL;、/及び/、	38
10	diff	・;NULL	36
10	same	、/及び/、	36

表 3

(3)　正解：登録／システム：非登録

順位	素性		出現数
1	same	日本	94
2	diff	の;NULL	87
3	same	協会	71
4	same	の	42
5	same	委員/会	37
6	same	法	24
6	same	科学	24
8	diff	、/及び/、;NULL	23
9	same	研究	22
10	same	会議	21

表 4

(4)　正解：非登録／システム：登録

順位	素性		出現数
1	diff	の;NULL	107
2	same	の	77
3	diff	、/及び/、;NULL	20
4	diff	、;NULL	19
5	diff	部門;課	14
5	same	アメリカ	14
7	diff	日本;NULL	12
8	same	科学	11
9	diff	NULL;の	8
9	same	会議	8
9	same	システム	8
9	same	教育	8
9	same	研究	8

表 5

上記の４通りのうち、表４と表５は、システムによる選別結果が誤っていた場合であるので、これらについて考察する。

実験データ全体の素性のうち、「システム：登録」となったものの選別が誤っていた(「正解：非登録」)率が 20.1%、「システム：非登録」となったものの選別が誤っていた(「正解：登録」)率も 20.1%である。よって、ある素性について、たとえば「システム：登録」となったもののうち「正解：非登録」だったものが 20%よりも際だって多ければ、それは特徴的な素性と言え、誤った原因の分析に役立つ可能性が考えられる。そのような素性のうち主要なものについて、それを含む現状訳と新規訳候補の差分情報を調べてみた。

「正解：登録／システム：非登録」及び「正解：非登録／システム：登録」に属する素性の、出現数トップ 20 のうち、その素性の選別が正しかった

場合（たとえば「正解：登録／システム：非登録」に対しては「正解：非登録／システム：非登録」）の出現数が、誤っていた出現数の（平均なら４倍程度あるところ）0.5 未満のものを以下に挙げた。

素性		出現数	
		正解:登録／システム:非登録	正解:非登録／システム:非登録
same	協会	71	30
same	委員/会	37	10
same	法	24	5
same	研究	22	10
same	会議	21	7
same	米国	20	2
same	国際	17	4
same	開発	17	7
diff	、/及び/、;・	16	6
diff	の/ため/の;NULL	14	5
same	会	13	0
same	安全	13	4

表 6　「正解：登録／システム：非登録」の素性のうち「正解：非登録／システム：非登録」の出現数が少ないものの例

素性		出現数	
		正解:非登録／システム:登録	正解:登録／システム:登録
diff	NULL;の	8	0
diff	NULL;国際	7	0
diff	、 及び 、;と	7	2

表 7　「正解：非登録／システム：登録」の素性のうち「正解：登録／システム：登録」の出現数が少ないものの例

これらは、特に「誤りの場合に特徴的な素性」である可能性が高い。よって、これらについて、その属する素性ベクトルの傾向を調べてみた。詳細は割愛するが、「正解：登録／システム：非登録」で「正解：非登録／システム：非登録」の出現数が少ない素性の場合（表６）は、リストアップされている素性自体が「協会」「委員/会」など団体名・機関名の一部に使われるものである傾向が強いが、特徴ベクター（対訳候補）もやはりそういうものが多く、このような場合は選別結果が「非登録」であっても実際は登録である傾向がみられる。一方、「正解：非登録／システム：登録」で「正解：登録／システム：登録」の出現数が少ない素性の場合（表７）は、特徴ベクター（対訳候補）を見ても傾向が見いだしにくく、たとえば英見出し候補の現状訳が「航空/保険/業者/の/国際/労働/組合」で新規訳候補が「国際/航空/保険/連合」のように、人が見ても登録した方が一見望ましそうにみられる対訳が多い。これらが登録されなかった原因は未解明であるが、それぞれ個別の事情があった例外的現象という可能性もある。

４．おわりに

本研究では、ある英日対訳句を辞書に登録すべきか否かの自動判定のために、辞書開発者によって登録すると判断された英日表現対と、登録しないと判断された表現対とを利用した機械学習を試みた。素性として、英語表現を現状の機械翻訳システムで翻訳した結果(現状訳)と、この英語表現に対応する日本語表現(新規訳候補)とで、形態素単位で求めた差分情報を用いた。形態素そのもの、品詞情報、意味情報でそれぞれ機械学習を行った結果、形態素そのものを用いた場合に最良のＦ値 0.785 が得られた。選別誤りの分析から、辞書に登録すべき対訳の傾向の一部も判明した。

参考文献

[1] 石本浩之，長尾眞．対訳文章を利用した専門用語対訳辞書の自動作成―訳語対応における両立不可能性を考慮した手法について―．研究報告 NL102-11, 情報処理学会, 1994.

[2] 熊野明，平川英樹．対訳文書からの機械翻訳専門用語辞書作成．情報処理学会論文誌，**35**(11), pp.2283-2290, 1994.

[3] 高尾哲康，富士秀，松井くにお．対訳テキストコーパスからの対訳語情報の自動抽出．研究報告 NL115-8, 情報処理学会, 1996.

[4] 梶博行，相薗敏子．共起語集合の類似度に基づく対訳コーパスからの対訳語抽出．情報処理学会論文誌, **42**(9), pp.2248-2258, 2001.

[5] 辻慶太．対訳コーパスからの低頻度訳語対の抽出：翻字・頻度情報の統合的利用．第 49 回日本図書館情報学会研究大会発表要綱，pp.59-62, 2001.

[6] Y. Al-Onaizan and K. Knight. Translating Named Entities Using Monolingual and Bilingual Resources. In *Procs. of the 40th Annual Meeting of the ACL*, pp. 400-408, 2002.

[7] 下畑光夫，渡辺太郎，隅田英一郎，松本裕治．パラレルコーパスからの機械翻訳向け同義表現抽出．情報処理学会論文誌，**44**(11)，pp.2854-2863, 2003.