社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－173 (6)
2006／5／19

単語長と出現回数を考慮した対訳文書からの訳語検出

北村 美穂子　　下畑 さより

沖電気工業株式会社 研究開発本部
〒541-0053 大阪市中央区本町 2-5-7
Email: {kitamura655, shimohata245}@oki.com

特許などの技術文書を翻訳する場合，専門用語を正しく認定し，適切な訳を付与することが重要である．我々は，既存の対訳文書から，精度の高い専門用語の対訳辞書を自動的に作成する手法を提案する．本手法は，(1) 単言語の文書から隣接単語の分散に基づく手法によって専門用語を抽出し，(2) 対訳文書から単語長と出現回数を考慮した手法によってその専門用語の訳語を検出する．という二段階から成る．米国特許公報 6 万件から (1) の手法を用いて抽出した専門用語約 12,000 語に対して，(2) の手法を検証した結果，一位正解率 80.0%，一位以外も含めての正解獲得率 96%で専門用語訳が検出できることを確認した．

Translation Extraction Method using Word Length and Word Frequency

Mihoko Kitamura　　Sayori Shimohata

Corporate Research and Development Center, Oki Electric Industry Co., Ltd.
2-5-7 Honmachi, Chuo-ku, Osaka-shi, Osaka 541-0053, Japan
Email: {kitamura655, shimohata245}@oki.com

When translating technical documentations such as patent publications, it is important to adequately recognize technical terms and translate them accurately. We propose a method to automatically develop bilingual dictionaries of technical terms with a high degree of accuracy using existing bilingual corpora. This method is comprised of two main steps: (1) extracting technical terms from monolingual corpora using a statistical method by co-occurrences and word order constraints, (2) detecting the translations of the extracted terms from bilingual corpora using a statistical method on the basis of word lengths and word frequency. We examined the detection accuracy of step (2) in about 12,000 technical terms extracted from 60,000 US-Patent publications using step (1), and confirmed that the method achieved accuracy of 80.0% when the top-candidate translation detected was correct, and 96% when either of the candidates were correct.

1 はじめに

我々は，米国特許公報を対象にした言語横断検索 RIPWAY[1]のための機械翻訳システムを開発している．RIPWAY は，ユーザが入力したキーワードを翻訳して検索するキーワード翻訳方式ではなく，検索対象の文書を予め翻訳しておき，ユーザの入力を直接検索する文書翻訳方式を用いている．

文書翻訳方式において，機械翻訳を用いる場合，最も重要なことは，専門用語訳の正しさである．特に，RIPWAY のように検索対象文書が特許公報である場合では，検索精度の決め手になるのは，ユーザがキーワードとして入力する専門用語である．その専門用語が翻訳済文書において正しく翻訳されていないと，検索精度は著しく下がる結果となる．さらに，特許公報がカバーする分野は多岐に渡る．専門用語訳に関しても，それぞれの分野に応じて適切な訳語を付与することが求められる．

上記の課題に対処するため，我々は自動又は半自動的な専門用語の辞書獲得に力を入れている [3, 7]．

対訳辞書の自動獲得方法として従来より用いられている方法は，文対応済みの対訳文書から，原言語と対象言語の語の対応関係を統計的に見つけ出すという方法である [7, 8]．

この手法は，(1) 二言語のうちのどちらかの言語の用語リストを用意し，その用語リストの訳語を，対訳文書から検出する手法，(2) 用語リストを必要とせず，原言語と対象言語の対訳文書から対応関係の高い対訳表現を網羅的に抽出する方法，の二つに大別される．

後者の手法は，我々は既に提案している [7]．前者の手法に関しては，我々が過去に提案した隣接文字の分散値に基づく定型表現の自動抽出手法 [2] をさらに発展させることにより，より効果的な対訳辞書獲

[1] http://www.ripway.net/, RIPWAY はリコーテクノシステムズ株式会社の登録商標です．

得方法が実現できると考える．

そこで，本稿では前者の手法として，次の二段階の手法を提案する．

1. 我々が過去に提案した，隣接文字の分散値を利用した定型表現の自動抽出手法 [2] を利用して，まず，どちらか一方の言語の文書から専門用語を抽出する．

2. 文対応済み対訳文書を利用して，その専門用語の訳語を検出する．

2. の訳語検出では，検出対象とする専門用語と対訳文書から統計的に抽出されたその訳語候補において，「訳語候補の単語長と出現回数を考慮した対応度計算式」を定義し，その対応度の計算結果から，最も高い対応度を持つ訳語候補をその専門用語の訳語とする．

以下，2節では，1. の専門用語リストを作成するための「隣接単語の分散値を利用した専門用語抽出手法」について説明する．次に，3節では，単語長と出現回数を考慮した訳語検出の手法について説明する．4節では，本手法による実験及びその結果を報告し，結果について考察する．5節では関連研究について紹介する．最後に，6節で結論を述べる．

2 隣接単語の分散値を用いた専門用語抽出

本稿では，「ある特定の文書集合に特徴的に出現する意味のある表現のまとまり」(以下，表現ユニットと呼ぶ) は，分野特有の定型表現や専門用語であるという考えに基づき，単語列の前後に出現する単語の分散の度合を基準に表現ユニットを抽出する．さらに，構成単語の品詞情報と分野情報を使って，表現ユニットから専門用語を抽出する．

分散の度合は，ある単語列に隣接する単語の種類と各単語の生起する確率で表す．隣接単語の種類が多いほど，また，隣接単語の生起確率が均等であるほど，分散の度合は大きくなる．逆に，隣接単語の種類が少ないほど，また，隣接単語の生起確率が偏っているほど，分散の度合は小さくなる．つまり，分散の度合の大きい単語列はその位置で分割される表現ユニットである可能性が高く，小さい文字列はその単語列は隣接単語を含むより長い単語列の一部，すなわち，断片的単語列である可能性が高いと判断する．分散の度合の計算には，エントロピー基準を用いる．エントロピーは事象の不確定さ，乱雑さを表す量で，等確率で選択肢が多いほど増大し，偏った確率で選択肢が少ないほど減少する．

上記の考え方に基づいて，以下のステップで，単言語の文書から専門用語抽出を行う．なお，以下では，4節の実験に準じるため，抽出対象を英語専門用語として説明する．

1. 単言語文書からの候補単語列の抽出 英語文書から表現ユニットの候補となる英語単語列とその出現を求める．抽出する単語列は，基本的に文書中に出現する全ての単語N-gramであるが，特定の記号や句読点など，専門用語の構成単語として不適切な単語を含むものはここで除外する．

2. 各候補単語列の隣接単語情報の取得 候補単語列の前後に出現する単語の種類とその生起確率を求める．後接単語の場合を例に以下説明する．単語数 l の単語列 $S = c_1, c_2, \cdots c_l$ は単語列 $l-1$ の単語列 $S\prime = c_1, c_2, \cdots c_{l-1}$ の後ろに c_l が出現する状態を表す．この時，$S\prime$ の後ろに c_l が出現する回数 $f(S\prime, c_l)$ は S の出現回数 $f(S)$ と等しい．このことから 1. で得られた全ての候補単語列 S を $S\prime$ と最後の単語に分割し，$S\prime$ 毎に最後の単語列の出現回数を修正することにより，後接単語の種類と出現回数を求めることができる．また，$S\prime$ の出現回数を $f(S\prime)$ とすると，$S\prime$ の後ろに c_l が生起する確率 $P(S\prime, c_l)$ は次式で求められる．

$$P(S\prime, c_l) = \frac{f(S\prime, c_l)}{f(S\prime)}$$

3. 候補単語列のエントロピー値の計算 隣接単語情報から，各候補文字列の前方エントロピー及び後方エントロピーを計算する．単語列 S の隣接文字集合を $W(S) = \{w_i | w_1, \ldots w_n\}$，隣接単語 w_i の生起確率を $P(A, w_i)$ とする時，エントロピー $H(S)$ は以下の式で求められる．

$$H(S) = -\sum_{i=1}^{n} p(S, w_i) \cdot \log P(S, w_i)$$

4. エントロピー基準による表現ユニットの抽出 候補単語列を 3. で求めたエントロピー値に基づいて，表現ユニットとそうでない単語列に分類する．4節で用いた実験では，エントロピー値が 1.0 以上，かつ，総出現回数が 10 回以上の候補単語列を表現ユニットとして抽出した．

5. 品詞情報による専門用語の抽出 4. で求めた表現ユニットから，品詞情報と分野情報を用いて，専門用語を絞り込む．品詞情報による絞込みでは，専門用語は名詞句が中心となることから，既知の英語名詞句の品詞情報をもとに，抽出対象を名詞句となりうる品詞列の表現ユニットに限定する．さらに，分野情報による絞込みでは，表現ユニットの分野固有性を求め，特定の分野に偏って出現する表現ユニットのみを抽出する．4節の実験では，分野固有性の基準として，tf・idf[1] を用いた[2]．

表1に，コンピュータ分野の米国特許公報 7,343 件 (約 43,000 文) から上記手法を用いて抽出した専門用語抽出結果の例 (上位 10 位及び下位 10 位) を示す．第一列はエントロピー値，二列目は tf・idf 値，三列目は出現回数，四列目は抽出結果を示す．

[2] 4 節の実験では，一単語構成語の抽出語数が非常に多かったため，一単語構成語の場合は未知語のみに抽出対象を絞った．

$H(S)$	tf・idf	$f(S)$	抽出結果
4.86	5952.04	3105	image data
4.73	1206.95	426	host computer
4.55	791.45	326	original image
4.40	501.84	362	recording medium
4.26	668.64	236	peripheral device
4.18	686.60	220	central processing unit
4.17	1151.93	273	cache memory
4.16	730.97	258	image processing method
4.14	1259.88	623	computer system
4.09	624.89	309	storage device
:	:	:	:
1.01	118.15	28	input image
1.01	105.49	25	process apparatus
1.00	88.61	21	processor unit
1.00	71.73	17	store programme
1.00	67.51	16	pulse series
1.00	63.47	18	grey level value
1.00	63.47	18	generate system
1.00	59.07	14	small region
1.00	198.32	47	logic unit
1.00	185.66	44	graphic call

表 1: 専門用語抽出結果

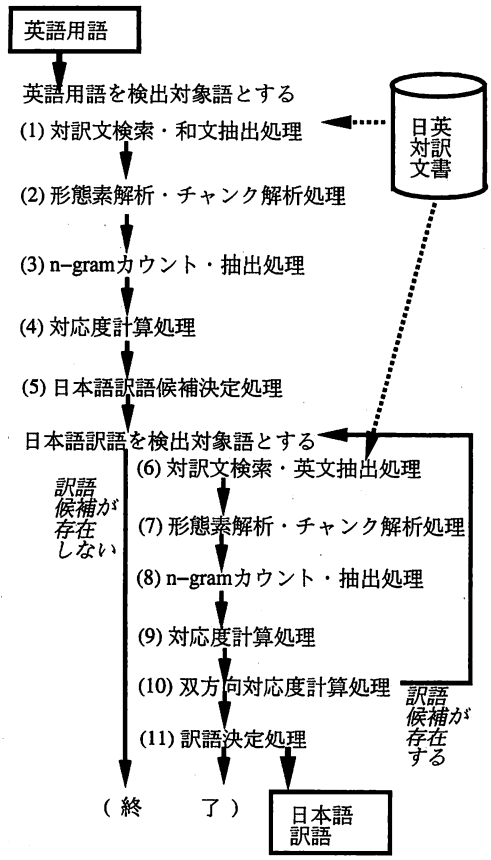

図 1: 訳語検出の流れ

日本語 n-gram 単語列	出現回数	構成単語数
アスペルギルス オリザエ	8	2
アスペルギルス	11	1
オリザエ	9	1

表 2: "aspergillus aryzae" の対応度計算例

3 単語長と出現回数を考慮した訳語検出手法

2 節で抽出した英語の専門用語に対して，日英の文対応済の対訳文書を利用して日本語の訳語を検出する．本処理の流れを図 1 に示し，各処理の詳細について以下に説明する．

(1) 対訳文検索・和文抽出処理 予め準備した文対応済の日英対訳文書から 2 節で抽出された英語の専門用語を検索し，その英語専門用語を含む英文に対応付けられた和文を全て抽出する．

(2) 形態素解析・チャンク解析処理 (1) の和文に対して，形態素解析及びチャンク解析 (文節区切りの認定処理) を行なう．

(3) n-gram カウント・抽出処理 (2) の解析結果を利用して，文節を超えない範囲で生成される全ての日本語 n-gram 単語列を作成する．

(4) 対応度計算処理 (3) で作成された日本語 n-gram 単語列に対して，英語の専門用語との対応度を計算する．英語専門用語 "org" と日本語 n-gram 単語列 "tar" との対応度 $sim(\text{“}org\text{”} \Rightarrow \text{“}tar\text{”})$ は単語長と出現回数を考慮した以下の式で定義する．

$$sim(\text{“}org\text{”} \Rightarrow \text{“}tar\text{”}) = \sum_{\text{“}sub\text{”} \in \text{“}tar\text{”}\text{の部分単語列}} \frac{conNum(\text{“}sub\text{”})}{conNum(\text{“}tar\text{”})} \cdot freq(\text{“}sub\text{”})$$

$conNum(\text{“}exp\text{”})$: “exp” の構成単語数
$freq(\text{“}exp\text{”})$: “exp” の出現回数

具体例を挙げて説明する．英語の専門用語 "aspergillus oryzae" と日本語 n-gram 単語列「アスペルギルス オリザエ」の対応度の計算は，次のように計算する．なお，各 n-gram 単語列の出現回数は表 2 の通りとする．「アスペルギルス オリザエ」の部分単語列は，「アスペルギルス オリザエ」，「アスペルギルス」，「オリザエ」の 3 単語列であり，その 3 つに対して

$$\frac{conNum(\text{“}sub\text{”})}{conNum(\text{“}tar\text{”})} \cdot freq(\text{“}sub\text{”})$$

を求め，それらを合計することにより対応度を求める．表 2 の例では，$\frac{2}{2} \cdot 8 + \frac{1}{2} \cdot 11 + \frac{1}{2} \cdot 9 = 18$ となり，その結果，$sim(\text{“}aspergillusoryzse\text{”} \Rightarrow \text{“}$アスペルギルス オリザエ$\text{”}) = 18$ となる．

この式は $freq(\text{“}sub\text{”})$ によって部分単語列の出現回数が多いほど対応度が高くなるように設定されている．さらに，$\frac{conNum(\text{“}sub\text{”})}{conNum(\text{“}tar\text{”})}$ によって，その部分単語列の構成数の大小に応じた重みを与えている．また，その総和を計算することにより，構成単語数が多く (つまり単語長が長く) なるほど，部分単語列の総数が多くなり，その総和も大きくなる．これによって，単語長が長いほど，対応度が高くなるという性質を対応度に与えることができる．

(5) 日本語訳語決定処理 予め設定された閾値を超える対応度を持つ日本語 n-gram 単語列をその専門用語の訳語候補とする．

閾値は予備実験の結果，出現回数 10 回以上の英語専門用語と 10 回未満の専門用語に分けて設定する．前者に関しては，訳語候補の出現回数が英語専門用語の出現回数の 1/10 回以上出現し，かつ，英語専門用語の出現回数の 1/10 以上の対応度をもつ日本語単語列を訳語候補とする．後者に関しては，1 以上の対応度をもつ日本語単語列を訳語候補とする．後者の閾値を下げる理由は出現回数が低い場合，閾値の設定により逆に検出漏れが起こるためである．

(6) 対訳文・英文抽出処理 以降，(5) により決定された訳語候補に対して，逆方向の訳語検出を行なう．つまり，(5) により決定された日本語訳語候補に対して (1) (5) と同様の処理を行なう．

まず，(5) により決定された日本語訳語候補を日英対訳文書から検索し，その訳語候補を含む和文に対応付けられた英文を全て抽出する．

但し，正解となる日本語訳語候補の出現回数は，元の英語専門用語の出現回数とほぼ等しいことが予測されるので，無駄な計算を避けるため，英語専門用語の出現回数の 100 倍以上の出現回数を持つ日本語訳語候補はここで除外する．

(7) 形態素解析・チャンク解析処理 (6) の英文に対して，形態素解析・チャンク解析処理を行なう．

(8) n-gram カウント・抽出処理 (7) の解析結果を利用して，文節を超えない範囲で生成される全ての英語 n-gram 単語列を作成する．

(9) 対応度計算処理 (8) で作成された英語 n-gram 単語列に対して，日本語の訳語候補との対応度 $sim(\text{“}tar\text{”} \Rightarrow \text{“}org\text{”})$ を (4) と同様の計算式で計算する．

(10) 双方向対応度計算処理 (9) の英語 n-gram 単語列の中に，訳語検出対象の英語専門用語が存在し，かつ，日本語訳語候補とその専門用語の対応度が，予め設定された閾値を超えるならば，$sim(\text{“}org\text{”} \Rightarrow \text{“}tar\text{”})$ と $sim(\text{“}tar\text{”} \Rightarrow \text{“}org\text{”})$ との和を求めることにより，双方向対応度 $sim(\text{“}org\text{”} \Leftrightarrow \text{“}tar\text{”})$ を計算する．

$$sim(\text{“}org\text{”} \Leftrightarrow \text{“}tar\text{”}) = sim(\text{“}org\text{”} \Rightarrow \text{“}tar\text{”}) + sim(\text{“}tar\text{”} \Rightarrow \text{“}org\text{”})$$

ここでの閾値は予備実験の結果，出現回数 100 回以上の英語専門用語と 100 回未満の専門用語に分けて設定する．

前者に関しては，訳語候補の出現回数が英語専門用語の出現回数の 1/100 回以上出現し，かつ，英語専門用語の出現回数の 1/100 以上の対応度をもつ日本語単語列を訳語候補とする．後者に関しては，1 以上の対応度をもつ日本語単語列を訳語候補とする．

閾値の設定が (5) よりも緩い理由は，双方向による検出により，不適切な候補語は除外されるため，閾値による制約の必要性が (5) に比べて低いためである．

全ての日本語候補語に対して，(6) (10) の処理を繰り返し適用する．

(11) 訳語決定処理 全ての日本語訳語候補を双方向対応度の高いものから順に並べ変え，出力し，処理を終了する．

4 実験

4.1 実験手法

上記手法の有効性を確かめるために，提案手法における専門用語抽出及び訳語検出の実験を行った．

専門用語抽出は，2 節の手法を用いて，米国特許公報のタイトル及びアブストラクトの 6 万件 (約 26 万文) から専門用語とみなされる英語単語列を抽出した．なお，1 単語からなる語の場合は，未知語のみを抽出対象とした．一方，複合語は，前置詞を含まない全ての複合語を対象とした[3]．

最終的に訳語検出に用いた未知語の数は 734 語，複合語の数は 11,501 語であった[4]．

一方，訳語検出に用いた日英の文対応済の対訳文書は，特許公報において，その文対応が確実な「発明の名称」の部分 (以降，タイトルと呼ぶ) を用いた．日本の特許公報とその英文抄録 (PAJ) のタイトル，約 350 万文と，米国特許公報とその日本語訳 (JAU) のタイトル，約 233 万文の計 583 万文を用いた．

3 節で述べた手法は，閾値を超えた日本語訳語候補語が対応度順に全て出力されるので，結果は 1 つと限らず，複数の結果が出力される．また，閾値を満たす訳語が存在しない場合は検出できない．したがって，全体のうち検出できた割合をカバー率として求め，検出できたもののうち一位で正解した割合

[3] 3 節で述べた方法は，(3) 及び (7) で文節を超えない範囲に候補語を限定しているため，現状では前置詞を含む専門用語の訳語検出をすることができない．このため前置詞を含む複合語は検出の対象外とした

[4] 今回はこの結果の評価はしておらず，良好な結果が得られていることを確認しただけである．

	未知語	複合語
訳語検出対象語数	734 語	11,501 語
検出語数	575 語	8,543 語
一位正解語数	462 語	6,959 語
一位正解率	80.3%	81.5%
含正解語数	556 語	8,241 語
含正解率	96.7%	96.5%
カバー率	78.3%	74.3%
一用語当たりの訳語候補数	4.2 語	4.4 語

表 3: 訳語検出結果

(以下，一位正解率と呼ぶ) と出力中に正解を含む割合 (以下，含正解率と呼ぶ) の 2 つを求めた．

カバー率 (%) = (検出語数／訳語検出対象語数) × 100
一位正解率 (%) = (一位正解数／訳語検出数) × 100
含正解率 (%) = (正解を含む結果数／検出数) × 100

4.2 実験結果及び考察

本実験結果を表 3 に示す．未知語，複合語共，第一位の候補語は，約 80%の正解率であった．また，一用語当たり抽出される訳語候補語数は，4.2　4.4 語であったが，その中に正解が含まれる割合は，未知語，複合語共，約 96%とかなり高かった．カバー率に関しては，未知語は，78.3%，複合語はそれよりやや低く，74.3%となった．

検出できなかった複合語の英語専門用語について，検出できなかった原因を調査した結果，主に 3 つの原因が存在した．

第一の原因は，出現回数が少なく，かつ，検出に用いた対訳文書中の訳語が一意でなかったため，閾値を満たす訳語候補語が存在しなかったためである．大半の未検出の用語がこれに相当した．第二の原因は，品詞認定や文節区切りの認定間違いによるものである．例えば，"feeding means" は means を動詞と認定されたため文節の範囲を超えると判断され，抽出することができなかった．第三は，英語用語の出現回数と，正解となる訳語候補語の出現回数が極端に異なるため，正解となる訳語候補語が，3 節の (9) の段階で除外されてしまった原因によるものである．

今後，カバー率を向上させるためには，第一の原因を解消することが有効だと思われる．例えば，検出できなかった英語専門用語に対しては，閾値を低くし再検出する，または，双方向対応度の計算によって訳語候補語を検出するのではなく，英語用語から日本語候補語の対応度の計算のみで検出する，等の方法が考えられる．

一方，間違った結果が得られた用語を分析すると，未知語，複合語共に，一用語当たりの訳語候補語数が 1.8 語と，平均に比べてかなり少なかった．しかし，候補語が 1 語の場合の正解率は 82.7%と全体に比べて少し高いという結果となった．

間違った結果は，次の 3 タイプに分類される．

1. 用語の訳が訳語候補語の一部に含まれる場合 (表 4 における (1) の例)
2. 用語の訳の一部のみが訳語候補になっている場合 (表 4 における (2) の例)
3. 対応関係がない (表 4 における (3) の例)

これらの間違いの原因の 1 つとして，我々が定義した対応度の特徴である単語長の長い候補語を優先することによる弊害が挙げられる．表 4 の (1) のように "arylene" の訳語として「アリーレンスルフィドポリマー」が抽出されたのはこの弊害のためである．

もう 1 つの間違いの原因は，使用した対訳文書において，正解となるべき日本語訳語が定まっておらず，様々な訳語で表現されていたため，正解となるべき日本語訳語の対応度が低くなり，別の間違った候補が選択されてしまうことによる問題である．例えば，表 4 の英語用語 " liquid vehicle " を含む対訳文における "liquid vehicle" の訳語は，「液体」，「液状の媒体」，「溶剤」，「液体賦形剤」，「液体ビヒクル」と全て異なっていた．これらは，今後，解決すべき課題である．

表 4 に，本手法で抽出された訳語候補の例を示す．評価の欄の◎印が一位正解のもの，○印が含正解のもの，×印が不正解のものである．このように，提案手法は，一般的な対訳辞書には見られない数多くの専門用語の訳語を検出することができる．

提案手法は，精度が 100%でないため，検出された訳語候補は，人手で正しいか否かを確認した後，登録しなければならないという課題が残る．しかし，96%の含正解率ということは大半の場合，正解が含まれているということである．本研究で扱った特許文書では，対象とする英語用語は技術用語が大半であるので，日本語訳語候補になる語は，表 4 のように，英語のカタカナ表記 (例：xerography の訳語はゼログラフィ) や英語表現と同一 (例: ALU の日本語訳語候補は ALU) である場合が多い．したがって訳語が正しいかどうかを辞書で確認する場合は少なく，その分野の専門知識がない人でも容易に正解を決定することができる．

5 関連研究

今回我々が提案した，二言語のうちのどちらかの言語の用語リストを用意し，その用語リストの訳語を文対応済み対訳文書から発見する手法は，既にいくつか提案されている [4, 8]．また，我々と同じく特許公報を利用して対訳抽出を行なう研究も数多く行なわれている [4, 5, 6]

熊野ら [8] は，統計情報と言語情報 (既存の対訳辞書) を併用することにより対訳文書から専門用語の対訳を抽出している．さらに，出羽 [6] はこの手法を用いて，特許公報から機械翻訳用の専門用語辞書を構築し，その辞書を用いて特許公報を翻訳することにより，翻訳品質の改善が見られたことを報告している．

我々の手法との違いは，上記の研究は言語情報を利用している点にある．言語情報を用いることを前提にした手法の場合，未知語の訳語検出精度は低いものとなる．また，特許公報に出現する専門用語は，既知語であっても特別な語の定義を用いる場合が多

英語 専門用語	日本語訳語	対応度	評価	
xerography	ゼログラフィ	11.6	◎	
varifocal	可変焦点レンズ	20.0		
	可変焦点	17.8	○	
	バリフォーカルレンズ	4.0		
arylene	アリーレンスルフィドポリマー	37.0	×	(1)
backpatching	-			
magnetic thin film	磁性薄膜	1132.5	◎	
actuator assembly	アクチュエタ	132.8		
	アクチュエタアセンブリ	69.5	○	
	作動装置	66.3		
amorphous semiconductor layer	非晶質半導体	12.5	×	(2)
liqiud vehicle	スルホポリエステル	7.2		
	ヒドロアルコール	6.0	×	(3)
	エアゾールヘアースプレー	6.0		
back focal distance	-			

表 4: 日本語訳語検出結果の例

く，既存の対訳辞書における対応関係がない場合も多い．

我々の手法は言語情報を利用しなくても，約 80%の精度で訳語検出をすることができた．我々の対応度計算式を言語情報も利用する計算式へと改良することにより精度が向上するかどうか，今後，是非検証したい．

その他，特許公報を対象にする手法としては，特許公報の数情報を利用して対応関係を推定する手法 [5]，統計情報と特許文書特有の構成単語の品詞情報を利用する方法 [4] が提案されている．

前者は数情報を利用しているが，今回我々が利用した発明の名称 (タイトル) には，数情報はほとんど存在しない．

一方，後者は専門用語辞書に登録されている用語の品詞パターンを予め分析し，その分析結果に基づいて，対訳文書から両言語の専門用語候補を作成し，両言語の専門用語候補の対応度の高いものを対訳として抽出する手法である．我々の手法でも日本語の専門用語を抽出する際，品詞パターンを利用しており，この点では上記の手法と類似する．しかし，訳語候補に関しては統計情報のみを利用し，品詞パターンの間違いによる悪影響を抑えている．さらに，我々は対応度計算式に対して独自の工夫を施すことにより，高い精度を実現している．

6 結論

本稿では，(1) 単言語文書から隣接単語の分散に基づく手法によって専門用語を抽出する，(2) 対訳文書から単語長と出現回数を考慮した手法によって (1) の専門用語訳語を検出する，という二段階の手法を用いて，専門用語の対訳を自動的に抽出する手法を提案した．

専門用語が多数存在する特許公報を抽出対象とし，米国特許公報約 6 万件から英語の専門用語を抽出し，日本語特許公報と米国特許公報のタイトルの対訳文計 583 万文を利用して，その英語専門用語の訳語検出を行なったところ，未知語においては精度 80%，複合語においても精度 81%で正しい訳語を検出することができた．

本稿で対象とした特許公報は，多数の分野に分類されており，分野によって訳語が異なる場合もある．例えば，“administration” の訳語は，情報化学分野では「管理」であるが，医学薬学分野では「投薬」であるのが望ましい．

本提案では，カバー率を向上させるため，分野による訳語の違いは考慮に入れず，検出対象とする特許公報全件を 1 つの検出対象として訳語検出を行なった．今後，カバー率を下げることなく，分野による訳語の違いも検出できるような手法に改良したいと考える．

謝辞

本研究は，情報通信研究機構平成 14 年度民間基盤技術研究促進制度に係る研究開発課題「多言語標準文書処理システムの研究開発」の一環として行われたものである．

参考文献

[1] Salton, G. and McGill, M. J.: *Introduction to modern information retrieval.*, McGraw-Hill (1987).

[2] Shimohata, S., Sugio, T. and Nagata, J.: Retrieving Collocations by Co-occurrences and Word Order Constraints, *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics*, pp. 476–481 (1997).

[3] 下畑さより, 山崎貴弘, 坂本仁, 北村美穂子, 村田稔樹: 特許翻訳における専門用語辞書構築, 言語処理学会第 11 回年次大会論文集, pp. 356–359 (2005).

[4] 福井雅敏, 樋口重人, 藤井敦, 石川徹也: 日米対応特許コーパスを用いた対訳抽出手法, *NL-145-4*, pp. 23–28 (2001).

[5] 高橋博之, 川崎立八, 牧田光晴, 樋口重人, 藤井敦, 石川徹也: 日米特許公報を用いた対訳辞書および翻訳メモリの構築, *NL-155-7*, pp. 39–46 (2003).

[6] 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, *NL-144-1*, pp. 1–7 (2001).

[7] 北村美穂子, 松本裕治: 言語資源を活用した実用的な対訳表現抽出, 自然言語処理, Vol. 13, pp. 3–25 (2006).

[8] 熊野明, 平川秀樹: 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, Vol. 35, pp. 2283–2290 (1994).