日米対応特許データに基づく対訳自動抽出

○樋口重人*1，福井雅敏*1，藤井　敦*2,3 石川徹也*2

Extracting Bilingual Lexicons Based on Japan-US Patent Families

○Shigeto HIGUCHI*1, Masatoshi FUKUI*1
Atsushi FUJII*2,3 Tetsuya ISHIKAWA*2

Abstract

To facilitate retrieving patent information across languages, we are developing a multi-lingual patent retrieval system, where user queries are translated into the target language by way of a dictionary. In this paper, aiming to enhance the translation dictionary, we propose a method to automatically extract translations from Japan-US patent families consisting of Japanese/English comparable texts. Our method computes the association score for each combination of Japanese/English words in patent families, and selects those with greater scores as translations. We also show the effectiveness of our method through experiments.

1. はじめに

多言語情報をオンラインで容易に入手できるようになったことを主な背景として，言葉の違いを意識せずに欲しい情報を検索するための言語横断情報検索や多言語情報検索の研究が近年盛んに行われている．

特許検索の分野では，一国における出願が国際的な影響を及ぼすプロパテント政策（特許重視・保護強化政策）によって，日本語を用いた外国特許の検索と，外国語を用いた日本特許の検索を支援するシステムの実現が求められている．このニーズに応えるために（株）パトリス[注1)]は藤井ら[1)2)]が提案した手法を用いて，多言語特許検索システム「PRIME」(*Patent Retrieval In Multi-lingual Environment*)を開発している[3)4)]．PRIMEは，検索キーワードを検索対象の言語に翻訳することで多言語検索を実現する．しかし，日々増え続ける新しい発明に関する新語を適切に翻訳するためには，対訳辞書の定常的な更新が必要である．

自然言語処理の研究では，対訳関係にある多言語コーパス（例文集）から単語や句の対訳を自動抽出する手法が提案されている[5)6)]．しかし，一般的に対訳コーパスの入手や作成は高価である．そこで本研究は，優先権主張制度に基づいて出願された特許から単語対訳を自動抽出する手法を提案する．

以下，２章で日米対応特許から対訳コーパスを作成する手法について説明し，３章で対訳の自動抽出手法について説明する．４章で抽出結果を評価し，５章でまとめと今後の研究課題について述べる．

注 1）平成 13 年 4 月より（財）日本特許情報機構の民需部門は民営化され，（株）パトリスに譲渡移管された．

2. 日米対応特許を用いた対訳コーパスの作成

2.1 対応特許

特許制度には優先権主張を伴う出願（特許法第四三条　パリ条約による優先権主張の手続）がある．優先権主張を伴う出願とは，パリ条約に加盟している国（1998 年時点で 171 ヶ国）の在国人であれば，自国で出願した特許を元にして，同一内容の特許をパリ条約に加盟している他国にも出願できる制度である．さらに，パリ条約加盟国に出願した特許は，自国で出願した日まで出願日が遡及され，出願人に時間的に有利に働く．米国や一部の国を除くと，先願主義（先に出願した方が優先される特許制度）を採用している国が多いため，本制度は国際的に大きな効力を発している．

優先権主張に基づいて複数の国に出願された特許の集合は「パテントファミリー(対応特許)」と呼ばれる．パリ条約には，対応特許は完全に同一内容である必要性は明文化されていない．しかし，自国の出願を元に加盟国に出願するため，対応特許の内容は比較的類似していることが多い．そこで，まず優先権主張制度を利用して出願された日本特許を抽出した．次に，優先権主張番号に基づいて対応する米国特許を特定し，日英対訳コーパスとして利用した．なお，同一内容の特許を複数の国に出願する方法には外国直接出願もある．しかし，この方法で出願された場合，対応特許の特定は困難であるため，本研究では対象外とした．

2.2 対訳コーパスの作成

日本で出願された特許として，特許庁発行の公開特許公報データを用いた．図 1 に公開特許公報の抜粋を示す．本データには「(31)優先権主張番号」，「(33)優先権主張国」の項目がある．ここでは優先権主張国に「米国(US)」とあるので，本出願は米国で先に出願され，ほぼ同一内容で日本に出願されたことがわかる．次に，図 1 の特許と同じ優先権主張番号を持つ米国特許公報を図 2 に示す．図 2 において，優先権主張番号は「[21]Appl.No.」で示されている．

日本公開特許公報として 1995 年～1999 年の 5 年間に公開された約 175 万件を用いた．この内，米国での優先権主張を伴う出願は 32,590 件存在した．優先権主張番号を元に米国登録公報を抽出した結果，特許公報 32,896 件が得られた[注2)]．以上まとめると，日本と米国の対応特許に基づいて約 3 万件の日英対訳コーパスを作成した．

注 2) 日本公報 1 件に対して複数の米国登録公報が対応する場合もあるため，日本と米国で特許公報数に違いが生じた．

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平8－114278

(51)Int.Cl.[6]　識別記号　庁内整理番号　ＦＩ

Ｆ１６Ｋ　31/66
Ｇ０５Ｄ　7/03
// Ｈ０１Ｌ　21/306

(43)公開日　平成８年(1996)５月７日

技術表示箇所

Ｈ０１Ｌ　21/ 306　Ｌ

審査請求　未請求　請求項の数18　ＦＤ　（全 11 頁）

(21)出願番号　特願平7－239230

(22)出願日　平成７年(1995)８月24日

(31)優先権主張番号　２９５，１２７
(32)優先日　1994年８月24日
(33)優先権主張国　米国（ＵＳ）

(71)出願人　590000400
ヒューレット・パッカード・カンパニー
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000
(72)発明者　クリストファー・シー・ベティー
アメリカ合衆国ペンシルバニア州ランデンバーグ、ピルズ・ウエイ　23
(72)発明者　ジェイムズ・ダブリュー・ベイカー
アメリカ合衆国メリーランド州エルクトン、カーター・ロード　110
(74)代理人　弁理士　上野　英夫

(54)【発明の名称】　マイクロアクチュエータ

(57)【要約】
【課題】断熱構造を備えるマイクロアクチュエータ。
【解決手段】フローチャネルを介して運搬される流体流を制御する超小型バルブの形態をなすマイクロアクチュエータであり、サーマルアクチュエータによって選択的に駆動される熱駆動部材を有し、これが駆動されることによって熱エネルギを生成する第１基板と、対向する第１、第２主要面を有する第２基板よりなる。第２基板が第１主要面で第１基板に取付けられる。第２の主要面は第２基板が支持体に取付られると絶縁セルを画定し、これによってマイクロアクチュエータの熱容量を減少させ、第１基板を支持体から熱遮断する。

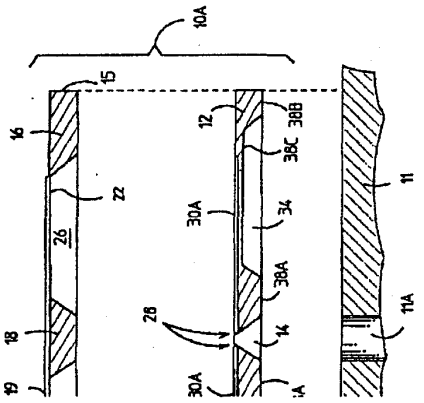


図１　日本公開特許公報(抜粋)

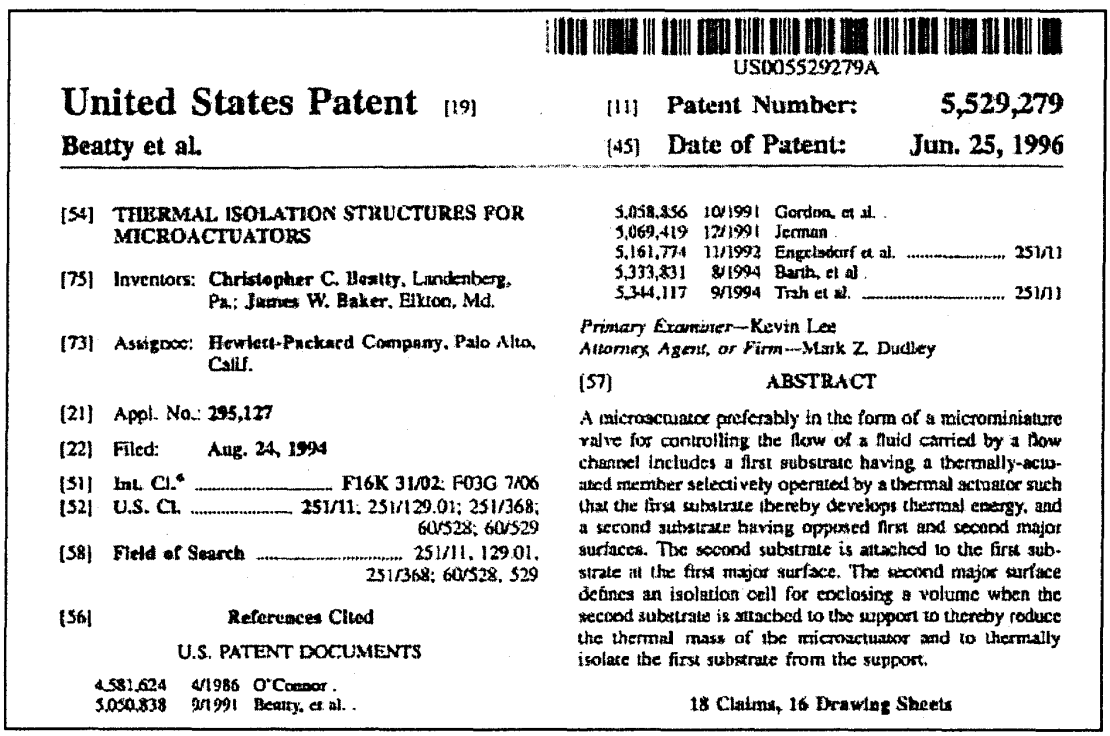
3.　対訳自動抽出法

3.1 概要

対訳コーパス（対応特許）において高い相関で共出現する日本語と英語は対訳である可能性が高い．対訳抽出の単位としては単語や句が考えられる．しかし，句の自動認定は依然として困難であるため，本研究は名詞性の単語を対象に対訳抽出を行った．そこで日米特許から名詞を抽出し，統計的な相関スコアに基づいて対訳を特定した．

3.2 対訳箇所の特定

特許公報は項目によって構造化されているので，日米で対応する項目を特定できれば対



US005529279A

United States Patent [19]

Beatty et al.

[11] **Patent Number: 5,529,279**

[45] **Date of Patent: Jun. 25, 1996**

[54] **THERMAL ISOLATION STRUCTURES FOR MICROACTUATORS**

[75] Inventors: **Christopher C. Beatty**, Landenberg, Pa.; **James W. Baker**, Elkton, Md.

[73] Assignee: **Hewlett-Packard Company**, Palo Alto, Calif.

[21] Appl. No.: **295,127**

[22] Filed: **Aug. 24, 1994**

[51] Int. Cl.[6] F16K 31/02; F03G 7/06

[52] U.S. Cl. 251/11; 251/129.01; 251/368; 60/528; 60/529

[58] Field of Search 251/11, 129.01, 251/368; 60/528, 529

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,581,624	4/1986	O'Connor .	
5,050,838	9/1991	Beatty, et al. .	
5,058,856	10/1991	Gordon, et al. .	
5,069,419	12/1991	Jerman .	
5,161,774	11/1992	Engelsdorf et al.	251/11
5,333,831	8/1994	Barth, et al. .	
5,344,117	9/1994	Trah et al.	251/11

Primary Examiner—Kevin Lee

Attorney, Agent, or Firm—Mark Z. Dudley

[57] **ABSTRACT**

A microactuator preferably in the form of a microminiature valve for controlling the flow of a fluid carried by a flow channel includes a first substrate having a thermally-actuated member selectively operated by a thermal actuator such that the first substrate thereby develops thermal energy, and a second substrate having opposed first and second major surfaces. The second substrate is attached to the first substrate at the first major surface. The second major surface defines an isolation cell for enclosing a volume when the second substrate is attached to the support to thereby reduce the thermal mass of the microactuator and to thermally isolate the first substrate from the support.

18 Claims, 16 Drawing Sheets

図２　米国登録公報(抜粋)

訳の探索範囲が限定され，対訳抽出の精度を高めることができる．

日本特許公報は【公開番号】，【出願日】，【出願人名】，【出願人住所】，【発明の名称】等が記載された書誌的事項と呼ばれる部分と，【要約】，【請求の範囲】，【発明の詳細な説明】，【実施例】，【図面】などから構成されている．項目の分布や項目名の表記は特許ごとにばらつきがある．これは米国登録公報でも同じである．しかも，対応特許は完全に同一内容ではないため，日米対応特許の間で項目にずれが生じる．すなわち，対応する日米特許中の対訳箇所を特定することは容易ではない．そこで予備調査を行った結果，２章で作成したコーパスにおいて発明の名称と要約は全件対応したので，当該項目を対象に自動抽出を行った．図１，２において，それぞれの項目は【発明の名称】と[54]，【要約】とABSTRACTで示されている．

3.3 対訳自動抽出

まず，抽出対象項目について，日本特許公報に対して「茶筌」[注3)]を用いて形態素解析(単語分割と品詞付与)を行い，名詞および未知語を抽出した．ただし，数名詞，代名詞，非自立名詞，接尾名詞の形容語幹は除いた，抽出した単語数は延べ1,103,024語，異なり26,918語であった．

さらに，米国登録公報に対して「Brill Tagger」[注4)]を用いて品詞付与を行い，普通名詞，固有名詞を抽出した．抽出した単語数は延べ689,788語，異なり35,232語であった．

注3)　http://chasen.aist-nara.ac.jp/index.html.ja

注4)　http://www.cs.jhu.edu/~brill/home.html

最後に，北村ら[5]が提案した「重み付き Dice 係数」を用いて日米単語対の相関スコア(score)を計算し，スコアが高い単語対を対訳として出力した．重み付け Dice 係数の計算方法を式1に示す．

$$score\ (w_J, w_E) = \log f_{JE} \times \frac{2 \times f_{JE}}{f_J + f_E} \quad (式1)$$

ここで w_J と w_E はそれぞれ日本語と英語の単語であり，f_J と f_E は対訳コーパスにおけるそれぞれの出現頻度である．また f_{JE} は抽出対象項目における w_J と w_E の共出現頻度である．右辺の log 成分は，通常の Dice 係数に対して，共出現頻度の重要度を強める効果がある．

4. 評価

3.2 節で抽出した対訳について人手で正解判定を行った結果を表 1 に示す．スコアの閾値を高く設定すると，抽出される対訳数は減少するものの正解率が向上し，より精密な対訳を抽出できることが分かった．

本研究の目的は，既存の辞書に定義されていない対訳を自動的に抽出する点にある．そこで 3,035 語の正解対訳を「専門語辞書（約 100 万語収録）」[注5]と照合した．その結果，202 語は専門語辞書に定義されていない対訳であった．これらの対訳例を表 2 に示す．

優先権主張制度に基づく出願が続く限り，対訳コーパスは今後も増加する．すなわち，対訳辞書の定常的な更新を実現できる見通しを得ることができた．

表1 スコアと正解率の関係

スコア	対訳数	正解対訳数	正解率
0以上	26,569	3,035	11.4%
0.5以上	2,613	1,543	59.1%
1.0以上	982	760	77.4%
2.0以上	195	182	93.3%
3.0以上	33	32	97.0%

表2 自動抽出した辞書未登録対訳の例

日本語	英語
アシルオキシシラン	acyloxysilane
イオノグラフィック	ionographic
キノリジニウム	quinolizinium
メチルトリクロロシラン	methyltrichlorosilane
懸架	suspension
多方向	multi-direction
誘電	dielectric

注5）（株）ノヴァ http://www.nova.co.jp/

5. おわりに

本研究では日本と米国の対応特許公報を用いて日英対訳コーパスを作成し，コーパス中の特定の特許項目から単語対訳を自動抽出する手法を提案した．抽出対象項目の拡張や句（フレーズ）の対訳抽出は，今後の研究課題である．

参考文献

1) 藤井敦，石川徹也．技術文書を対象とした言語横断情報検索のための複合語翻訳．情報処理学会論文誌，Vol.41, No.4, pp.1038-1045, 2000.
2) 藤井 敦，石川 徹也．質問翻訳と文書翻訳を統合した日英言語横断情報検索．電子情報通信学会論文誌，Vol.J84-D-II, No.2, pp.362-369, 2001.
3) Masatoshi Fukui, Shigeto Higuchi, Youichi Nakatani, Masao Tanaka, Atsushi Fujii and Tetsuya Ishikawa. Applying a Hybrid Query Translation Method to Japanese/English Cross-Language Patent Retrieval. ACM SIGIR Workshop on Patent Retrieval, 2000.
4) 樋口重人，福井雅敏，藤井敦，石川徹也．特許情報を対象とした言語横断検索システムの開発．言語処理学会第7回年次大会発表論文集，pp.445-447, 2001.
5) 北村美穂子,松本裕治．対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol.38, No.4, pp.727-736, 1997.
6) Frank Smadja, Kathleen R. McKeown and Vasileios Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, Vol.22, No.1, pp.1-38, 1996.

---

*1 （株）パトリス（〒135-0043 江東区塩浜二丁目4番29号）
*2 図書館情報大学（〒305-8550 つくば市春日1-2）
*3 科学技術振興事業団 CREST

Shigeto Higuchi (s_higuchi@patolis.co.jp) PATOLIS Co.
Masatoshi Fukui (m_fukui@patolis.co.jp) PATOLIS Co.
Atsushi Fujii (fujii@ulis.ac.jp) University of Library and Information Science; CREST, Japan Science and Technology Corporation
Tetsuya Ishikawa (ishikawa@ulis.ac.jp) University of Library and Information Science