情報学基礎 64－4
自然言語処理 145－4
（2001. 9. 10）

日米対応特許コーパスを用いた対訳抽出手法

福井 雅敏† 樋口 重人† 藤井 敦††,††† 石川 徹也††

† （株）パトリス
〒135-0043 東京都江東区塩浜 2-4-29
†† 図書館情報大学
〒305-8550 つくば市春日 1-2
††† 科学技術振興事業団 CREST

E-mail: fujii@ulis.ac.jp

あらまし インターネット上の外国語情報を母国語によって活用するためには，新語に対応して迅速に翻訳辞書を更新する必要がある．本研究は，同一内容に関して日本語と米国に出願された対応特許をコーパスとして利用し，複合語単位の日英対訳を抽出する手法を提案する．本手法は，発明名称や要約などの特許項目から品詞パターンに基づいて日本語と英語の複合語を抽出し，さらに統計的な関連度スコアを計算して適切な対訳を選択する．1995～1999 年に出願された約 32,000 件の対応特許を対象に評価実験を行った結果，本手法によって年間数千件の新語対訳を半自動的に取得できる見通しを得た．

Bilingual Lexicon Extraction Using Japan-US Patent Family Corpora

Masatoshi Fukui†, Shigeto Higuchi†, Atsushi Fujii††,†††, Tetsuya Ishikawa††

†PATOLIS Corporation
2-4-29 Shiohama Koto-ku, 135-0043, Japan
††University of Library and Information Science
1-2 Kasuga Tsukuba, 305-8550, Japan
†††CREST, Japan Science and Technology Corporation

E-mail: fujii@ulis.ac.jp

Abstract To facilitate utilizing foreign information over the Internet, it is crucial to update dictionaries by way of new translations. We propose a method to extract Japanese/English phrasal translations from patent families consisting of Japanese-US patents associated with the same invention. Our method first uses part-of-speech patterns to extract phrases from patent fields, such as titles and abstract, and then computes a statistical association score between each combination of Japanese and English phrases to select appropriate translations. For the purpose of experiments, we used approximately 32,000 patent families filed in 1995-1999, and showed that several thousand new translations per year could be obtained semi-automatically through our method.

1 はじめに

インターネットの普及によって様々な外国語情報を容易に入手できる時代になり，それらを母国語で読みたい，また逆に自分の情報を他言語ユーザにも読んでほしいという要求が増加している．

しかし，インターネット上の情報（Webページなど）は，既存の翻訳辞書に登録されていない新語や専門用語などを数多く含んでいるため，機械翻訳や多言語検索において，辞書未登録語への迅速な対応が今後ますます重要になる．

上記問題への対処として，対訳関係にある多言語コーパスから単語や句の単位で対訳を自動抽出する統計的手法がある [2, 6, 8]．しかし，これらの手法が必要とする大規模な対訳コーパスは高価である．また，新語の発生に追随して最新のコーパスを迅速に入手する見通しがなければ，実用化は容易ではない．

最近では，Webから対訳情報を抽出する手法 [4, 5] が提案されている．Webから対訳関係にあるページや対訳情報を掲載したページを抽出できれば，定期的に安価な対訳コーパスを取得することができる．しかし，Webには低品質な情報が混在していることが多い．また，ヨーロッパ言語間に比べると，日本語と英語の均一な対訳ページは比較的少ない．

本研究では，優先権主張制度に基づいて出願された日米の「対応特許」を対訳コーパスとして利用し，複合語単位の対訳を抽出する手法を提案する．著者らは既に，多言語特許検索システム「PRIME」の辞書更新機能に当該手法を応用している [3, 7]．

対応特許とは，同一（あるいは類似の）内容に関して複数国に出願された特許である．通常これらは異なる言語で書かれているため，特許制度が存続する限り，新語・専門用語を含む高品質かつ大規模なコーパスを比較的容易に更新することができる．

以下，2章で日米対応特許公報から対訳コーパスを作成する手法について説明し，3章で対訳抽出手法を説明する．4章で本手法を評価し，実用化に向けての今後の研究課題について議論する．

2 対応特許を用いたコーパスの作成

2.1 優先権主張に基づく対応特許

特許制度にはパリ条約による優先権主張を伴う出願制度がある．パリ条約に加盟している国（2000年1月時点で157ヶ国）の在国人であれば，第1国で出願した特許に基づいて，同一内容の特許をパリ条約に加盟している第2国にも出願することができる．第2国に出願した特許は，第1国で出願した日まで出願日が遡及される．米国など一部の国を除くと，先願主義（先に出願した者が優先的に特許権利を得る制度）を採用している国が多いため，優先権主張制度は国際的に大きな効力を発している．

このように，同一の発明に関して複数国に出願された特許の集合をパテントファミリーと呼び，パテントファミリーを構成する特許を対応特許と呼ぶ．

対応特許は第1国と第2国でそれぞれ出願した特許間の構成部分が明らかにされていれば完全に同一内容である必要はない．しかし，一般的に内容が大きく逸脱することはないため，対応特許の内容は非常に類似している．

同一の発明を複数国に出願する方法には，優先権主張制度を利用する以外にも，各国への個別出願や国際出願がある．しかし，これらの方法で出願された特許に関しては，対応特許を特定することが容易ではない．それに対して，優先権主張制度に基づいて出願された場合は，特許に固有の優先権主張番号によって対応特許を機械的に特定することができる．

2.2 対訳コーパスの作成

本研究の趣旨は対応特許コーパスから対訳情報を抽出する点にあり，原理的には言語の種類は問わない．しかし，今回は日本と米国に出願された対応特許から日英対訳コーパスを作成した．

日本は公開制度を採用しているため，特許が出願されると，まず特許公開公報が発行され，特許が登録されると登録公報が発行される．すなわち，同一特許に対して2種類の公報が存在する．

公開公報と登録公報を比較すると，前者は件数が圧倒的に多いのに対して，後者は件数が少ないものの発明内容の質が高い．しかし，言語的な質に顕著な違いはないので，本研究では公開公報を用いた．

他方において，米国には公開制度がないため，登録公報のみ発行される．そこで，米国の特許については登録公報を用いた．

パテントファミリーを構成する日本公開公報と米国登録公報の例（抜粋）を図1と2に示す．この例では，図1の「(31) 優先権主張番号」と「(33)【優先権主張国】米国（ＵＳ）」によって，米国に先に出願した内容に基づいて日本に優先権主張され，米国と日本がそれぞれ第1，2国であることが分かる．

さらに，図1の「(31) 優先権主張番号」と図2の「【21】Appl.No.」を対比させることで，両者が対応特許であることが分かる．しかし，第1国に出願された特許には優先権主張を示す項目はない．この例とは逆に日本が第1国である場合は，日本公開公報には優先権主張に関する項目は含まれない．

そこで，原理的には日本公開公報と米国登録公報の両方を全件探索して，第2国での出願をまず特定し，次に優先権主張番号に基づいて第1国での出願を特定し，両者を対応付ける処理が必要である．

しかし，本研究では特許オンライン検索システム「PATOLIS」[1]を用いて，1995年～1999年の5年間に公開された日本公開公報（約175万件）だけを探索することで，米国での優先権主張を伴う32,590件を特定した．さらに該当する米国登録公報だけを選

[1] http://www.patolis.co.jp/

択的にUSサイトから取得し，約32,000件の日英対訳コーパスを作成した[2].

(11)【公開番号】特開平８－１１４２７８
(43)【公開日】平成８年（１９９６）５月７日
(54)【発明の名称】マイクロアクチュエータ
(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）
(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: 日本公開特許公報の例（抜粋）

[11] Patent Number 5,529,279
[45] Date of Patent June 25, 1996
[54] Thermal isolation structures for microactuators
[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.
[21] Appl. No.: 295127
[22] Filed: August 24, 1994

図 2: 米国登録特許公報の例（抜粋）

3 対訳抽出手法

3.1 概要

対訳コーパス（対応特許）において共出現する日本語と英語は対訳である可能性が高い．しかし，コーパスにおける出現位置を考慮しないと探索空間が不必要に広がるため，対応箇所を制限する必要がある．従来の手法では，文単位の対応を利用したり [6, 8]，語の統計頻度によって自動的に対応を付ける [2].

しかし，日米対応特許を文単位に対応付けることは困難である．自動対応付けは計算コストが高く，また必ずしも高精度が期待できない．そこで本研究では，特許が複数の項目（フィールド）によって構造化されている点に着目し，構造情報に基づいて日英の対応箇所を特定する．

さらに，対応する日本語と英語の項目から複合語を抽出し，統計的な関連度スコアに基づいて適切な対訳を特定する．図 3 に対訳抽出の概要を示す．それぞれの処理について以下の節で具体的に説明する．

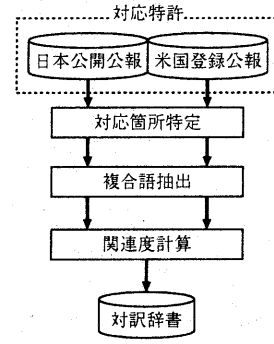

図 3: 対訳抽出手法の概要

3.2 対応箇所の特定

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することで，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】などが記載された書誌的事項と,【要約】,【請求の範囲】,【発明の詳細な説明】,【実施例】,【図面】などの項目から構成されている（図 1 参照）．しかし，項目の分布や項目名の表記は特許公報ごとにばらつきがある．

これは米国登録公報においても同様である．しかも対応特許は完全に同一内容ではないため，日米間で項目のずれも生じる．そこで，対応する日米特許中の対訳箇所を完全に特定することは容易ではない．

一部の対応特許について各項目間の対応を人手で分析した結果「発明の名称」と「要約」は全件で対応したので，当該項目を対象に対訳自動抽出を行った．なお，米国登録公報において「発明の名称」と「要約」は，それぞれ [54] と [ABSTRACT] で示されている（図 2 参照）．

他方において，請求項や実施例にも新語が多く含まれるので，これらの項目の利用についても今後検討を行う必要がある．

3.3 日本語複合語の抽出

複合語の特定は依然として難しい問題である．そこで本研究では，機械翻訳用のノヴァ専門用語辞書[3]の項目（多くは複合語）を事前に分析し，品詞パターンに基づく複合語抽出規則を作成した．本辞書は 19

[2] 第 1 国での特許 1 件に対して第 2 国で複数の特許が出願される場合もあるため，日本と米国では対応特許数に若干の差異が生じる．しかし，本研究では第 2 国での出願が複数ある場合には，無作為に 1 つを選択して第 1 国特許と対応付けた．

[3] （株）ノヴァ http://www.nova.co.jp/

分野で構成され（図4），日英・英日共に約100万語の対訳が登録されている．

航空・宇宙，バイオテクノロジー，ビジネス，化学，コンピュータ，土木・建築，防衛，地球環境，電気・電子，原子力・エネルギー，金融，法律，数学・物理，機械工学，医療・医学，金属，海洋・船舶，プラント，貿易

図 4: ノヴァ専門用語辞書の構成分野

専門用語辞書内の日本語を「茶筌」[4]を用いて形態素解析し，品詞パターンを分析したところ，異なりで33,378パターンが存在し，辞書項目は平均2.5単語で構成されていた．表1に頻出品詞パターンを示す．

最頻出の品詞パターンは「名詞+名詞」で，専門用語辞書内の品詞パターン中約35%存在した．表1より，専門用語辞書中の語のほとんどが名詞，未知語，接頭詞の組合せで構成されていることが分かる．これらを踏まえて以下の規則を作成した．

- 文章中の名詞，未知語，接頭詞，自立動詞（体言接続特殊），自立形容詞の連続を複合語として抽出する．ただし，非自立名詞，代名詞，数詞は含まない．

  例）粉砕した焙煎コーヒーの混合物

  （解析結果）

粉砕（名詞-サ変接続） し（動詞-自立） た（助動詞） 焙（動詞-自立） 煎（動詞-自立） コーヒー（名詞-一般） の（助詞-連体化） 混合（名詞-サ変接続） 物（名詞-接尾-一般）

  （抽出結果）粉砕，焙煎コーヒー，混合物

- 複合語を抽出する際，接頭詞・自立形容詞を末尾の単語としない．以下の例では「超（接頭詞）」が末尾にくるような複合語は抽出しない．

  例）電極含有超音波トランスデューサー

  （解析結果）

電極（名詞-一般） 含有（名詞-サ変接続） 超（接頭詞-名詞接続） 音波（名詞-一般） トランスデューサー（名詞-一般）

  （抽出結果）電極含有，超音波トランスデューサー

- 抽出した複合語が3単語以上で構成されている場合は，2語以上の単語組合せも複合語として別途抽出する．ただし，接尾名詞は直前の単語に連結し，単語として計上しない．

  例）ヒートポンプ加熱装置

  （解析結果）

ヒートポンプ（未知語） 加熱（名詞-サ変接続） 装置（名詞-サ変接続）

  （抽出結果）ヒートポンプ加熱装置，ヒートポンプ加熱，加熱装置

- カタカナ語は実際には複合語であっても茶筌で未分割の場合が多いため，単独で複合語として抽出する．

  例）熱可塑性シンチレーター材料

  （解析結果）

熱（名詞-一般） 可塑（名詞-一般） 性（名詞-接尾-一般） シンチレーター（未知語） 材料（名詞-一般）

  （抽出結果）熱可塑性シンチレーター材料，シンチレーター，熱可塑性シンチレーター，可塑性シンチレーター材料，熱可塑性，シンチレーター材料，可塑性シンチレーター

- 複合語を抽出した後，特許に特有の表現（該システム，本発明，各プロセス，前記処理中の，該，本，各，前記）を削除する．

表 1: 日本語の頻出品詞パターン

日本語品詞パターン	件数	割合%
名詞+名詞	411,211	35.31
名詞+名詞+名詞	208,839	17.93
名詞	165,677	14.22
名詞+名詞+名詞+名詞	67,955	5.83
接頭詞+名詞+名詞	19,641	1.69
名詞+名詞+名詞+名詞+名詞	18,315	1.57
接頭詞+名詞	14,812	1.27
名詞+動詞	12,549	1.08
カタカナ未知語	10,887	0.93
カタカナ以外の未知語	10,392	0.89

3.4 英語複合語の抽出

日本語と同様に，ノヴァ専門用語辞書内の英語項目を分析し，品詞パターンに基づく英語複合語の抽出規則を作成した．英語項目に対して「Brill tagger」[5]を用いて品詞付与したところ，異なりで12,016パターンが存在し，辞書項目は平均2.5単語で構成されていた．表2に頻出品詞パターンを示す．

最頻出の品詞パターンは「名詞+名詞」で，専門用語時書内の品詞パターン中約23%存在した．表2より，専門用語辞書中の語のほとんどが名詞と形容詞の組合せで構成されていることが分かる．これらを踏まえて，以下の規則を作成した．

- 形容詞と名詞の組合せを複合語として抽出する．英語複合語には接続詞（for, in, of など）を含むものがあるため，複合語の境界判定が難しい．しかし，接続詞を含む複合語は専門用語辞書中で2.5%ときわめて少なかったため，接続詞は全て複合語の境界として利用した．

[4] http://chasen.aist-nara.ac.jp/index.html.ja

[5] http://www.cs.jhu.edu/~brill/home.html

例）method and apparatus for channel equalization

（解析結果）

method（名詞） and（接続詞） apparatus（名詞） for（接続詞） channel（名詞） equalization（名詞）

（抽出結果）method, apparatus, channel equalization

- 抽出した複合語が3単語以上で構成されている場合は，2語以上の単語組合せも複合語として別途抽出する．
- 形容詞が末尾になるような複合語は抽出しない．以下の例では「band folding」のような単語連続は抽出しない．

例）multiple band folding antenna

（解析結果）

multiple（名詞） band（名詞） folding（形容詞） antenna（名詞）

（抽出結果）multiple band folding antenna, band folding antenna, multiple band, folding antenna

- 抽出した語に対して，WordNet [1] の活用情報を用いて接辞処理を行う．

表 2: 英語の頻出品詞パターン

英語品詞パターン	件数	割合%
名詞+名詞	253,805	23.26
名詞	237,400	21.76
形容詞+名詞	190,944	17.50
名詞+名詞+名詞	57,144	5.24
形容詞	48,043	4.40
形容詞+名詞+名詞	46,094	4.22
動詞+名詞	45,516	4.17
名詞+接続詞+名詞	18,497	1.70
形容詞+形容詞+名詞	12,124	1.11
名詞+動詞	10,448	0.96

3.5 関連度の計算

日本語と英語の関連度スコアとして北村ら [8] が提案した「重み付き Dice 係数」を用い，スコアが高い日英対を対訳として出力する．スコア（$Score$）の計算方法を式 (1) に示す．

$$Score(J,E) = \log F(J,E) \cdot \frac{2F(J,E)}{F(J)+F(E)} \quad (1)$$

ここで，J と E はそれぞれ日本語と英語の複合語であり，$F(J)$ 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 $F(J,E)$ は抽出対象箇所（発明の名称，要約など）における J と E の共出現頻度である．右辺の log 成分は，高頻度の共起に対してより大きな値を与える作用がある．

4 評価実験

4.1 実験方法

2章で作成した約 32,000 件の対訳コーパスを用いて対訳抽出手法の評価実験を行った．特許公報は文体が特殊であるため，単語数や文数などによってコーパスの規模を示すことが難しい．ファイル容量は，日本語と英語がそれぞれ 1.3GB と 1.6GB であった．

本実験の目的は大きく分けて2つある．まず，特許という特殊なコーパスからどの程度の精度で対訳を自動抽出できるかを評価した．さらに，既存の辞書に未登録の新語をどの程度抽出できるかを評価した．

具体的には，特許項目のうち「発明の名称」「要約」から個別に自動抽出した対訳に対して人手で正解判定を行い，さらにノヴァ専門用語辞書に登録されていない正解対訳数を調査した．

4.2 実験結果

3.3，3.4 節の手法によって発明の名称と要約から抽出された日本語と英語の複合語数を表 3 に示す．

表 3: 抽出された複合語数

	異なり	のべ
日本語	345,291	1,100,416
英語	439,020	1,233,070

発明の名称，要約に対する対訳抽出の実験結果を表 4 と 5 にそれぞれ示す．ここでは，スコアの閾値を段階的に上げていき，閾値以上の対訳だけを抽出した場合の正解率の変化を調査した．

表 4: スコアと正解率の関係（発明の名称）

スコア	対訳数	正解数	正解率%	新語数
0 以上	38,790	4,455	11.48	2,594
0.1 以上	27,756	4,206	15.15	2,488
0.2 以上	22,955	3,754	16.35	2,285
0.5 以上	12,816	2,193	17.11	1,467
1.0 以上	1,624	406	25.00	253
1.5 以上	275	83	30.18	48
2.0 以上	54	22	40.74	12

表 5: スコアと正解率の関係（要約）

スコア	対訳数	正解数	正解率%	新語数
0.1 以上	259,774	16,935	6.52	10,714
0.2 以上	230,113	13,399	5.82	8,551
0.5 以上	129,898	5,695	4.38	3,552
1.0 以上	10,902	1,007	9.24	476
1.5 以上	3,196	202	6.32	99
2.0 以上	530	39	7.36	16

表 4 より，発明の名称だけを用いた場合は，スコア閾値を上げるにつれて正解率も向上することが分かった．また，いずれの閾値に対しても辞書未登録の正解対訳数（表中「新語数」）は，正解総数のほぼ半数であった．図 5 に辞書未登録の対訳例を示す．

アルケニル含有ポリジオルガノシロキサン	alkenyl-containing polydiorganosiloxane
イオントラップ質量スペクトロメータ	ion trap mass spectrometer
インドリルアルキルピペラジニルピリジン	indolylalkylpiperazinyl pyridine
エアバッグキャニスタ	air bag canister
ジオルガノポリシロキサンポリマー	diorganopolysiloxane polymer
シリコーン感圧接着剤組成物	silicone pressure sensitive adhesive compositions
セルローストリアセテート写真	cellulose triacetate photographic
セルローストリアセテート写真フィルムベース	cellulose triacetate photographic film base
加硫性エラストマーコンパンド	vulcanizable elastomeric compound
感熱色素転写システム	thermal dye transfer system
小型走査共焦点顕微鏡	miniature scan confocal microscope
電子マネーシステム	electronic-monetary system
蠕動ポンプ	peristaltic pump

図 5: 辞書未登録対訳の例

他方において，表 5 より，要約を用いた場合は正解率がスコア閾値によらず比較的低かった．しかし，いずれの閾値に対しても辞書未登録語の正解対訳数は，正解総数のほぼ半数であった．なお，表 5 では人手のコストを抑えるために，スコア閾値 0.1 以上の対訳に対してのみ正解判定を行った．

以上まとめると，正解率については，特許以外のコーパスを用いた関連研究に比べて高くはなかったものの，最大で万単位の新語を抽出することができた．

さらに，新語の抽出数を年単位で調査した結果を表 6 に示す．ここでは，前年までに抽出された対訳は既知語と見なしている点に注意を要する．表 6 より，年平均 6,500 件の日米対応特許が出願されたことが分かる．また，専門用語辞書中の 100 万語を既知語と見なしたにも拘らず，年平均で 2,700 件の新語を抽出することができた．

表 6: 辞書未登録語の年間件数

年	1995	1996	1997	1998	1999	合計
特許件数	6,795	8,126	5,703	8,074	3,892	32,590
新語数	5,951	3,043	1,709	1,707	898	13,308

4.3 実用化に関する議論

現時点では，本手法を用いて対訳抽出を全自動化することは精度の面から難しい．しかし，対訳抽出精度が今後向上したとしても，100%の精度が保証されない限り，人手によるチェックは必須である．

本実験において，表 4 中の 38,790 対訳の正解判定に要した工数は約 4 人日である．言い替えれば，本抽出手法をフィルターとして用いることで，人手のコストを抑えた半自動の対訳抽出が可能である．

また，今回は発明の名称と要約のみを用いて実験を行い，それ以外の特許項目に対しては正解判定を行っていないものの，請求項には要約の約 10 倍の対訳が潜在的に含まれていることが分かっている．そこで，請求項を利用することで，表 6 に示した件数をはるかに越える新語を取得できる可能がある．

さらに，特許公報に付与された特許分類に基づいて分野別に対訳を抽出すれば，より実用的な辞書更新が可能である．

5 おわりに

日々増え続ける多言語情報を活用するためには，新語に対応した翻訳辞書の更新が必要である．本研究では，日本と米国に同一内容に関して出願された対応特許をコーパスとして利用し，複合語単位の新語対訳を自動抽出する手法を提案した．

特許中の「発明の名称」と「要約」を対象に対訳抽出実験を行った結果，年間千単位の新語対訳を半自動的に取得できる見通しを得た．今後は，請求項などの項目を対象にさらなる研究を行う予定である．

謝辞

専門用語辞書は（株）ノヴァの許諾を得て使用させて頂きました．この場を借りて深謝致します．

参考文献

[1] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[2] Pascale Fung. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 236–243, 1995.

[3] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. Prime: A system for multi-lingual patent retrieval. In *Proceedings of MT Summit VIII*, 2001. (To appear).

[4] Masaaki Nagata, Teruka Saito, and Kenji Suzuki. Using the Web as a bilingual dictionary. In *Proceedings of the ACL-EACL Workshop on Data-Driven Machine Translation*, 2001.

[5] Philip Resnik. Mining the Web for bilingual texts. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 527–534, 1999.

[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.

[7] 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.

[8] 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.