

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4431759号
(P4431759)

(45) 発行日 平成22年3月17日(2010.3.17)

(24) 登録日 平成22年1月8日(2010.1.8)

(51) Int.Cl.

F I

G 0 6 F 17/28 (2006.01)

G 0 6 F 17/28

C

請求項の数 8 (全 18 頁)

(21) 出願番号	特願2005-10712 (P2005-10712)	(73) 特許権者	393031586
(22) 出願日	平成17年1月18日(2005.1.18)		株式会社国際電気通信基礎技術研究所
(65) 公開番号	特開2006-201873 (P2006-201873A)		京都府相楽郡精華町光台二丁目2番地2
(43) 公開日	平成18年8月3日(2006.8.3)	(74) 代理人	100099933
審査請求日	平成18年12月21日(2006.12.21)		弁理士 清水 敏
(出願人による申告)平成16年度独立行政法人情報通信研究機構、研究テーマ「大規模コーパス音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第30条の適用を受ける特許出願		(72) 発明者	坂本 仁
特許権者において、実施許諾の用意がある。			京都府相楽郡精華町光台二丁目2番地2
		(72) 発明者	大竹 清敬
			株式会社国際電気通信基礎技術研究所内
			京都府相楽郡精華町光台二丁目2番地2
			株式会社国際電気通信基礎技術研究所内
		審査官	成瀬 博之
		最終頁に続く	

(54) 【発明の名称】 未登録語自動抽出装置及びプログラム、並びに未登録語自動登録装置及びプログラム

(57) 【特許請求の範囲】

【請求項1】

第1の言語の文と第2の言語の文との間の書換パターンと、前記第1の言語と前記第2の言語との間の複数の対訳用例文とに基づいて、前記第1の言語の機械可読な辞書に未登録な語を自動的に抽出する未登録語自動抽出装置であって、

前記書換パタンの前記第1の言語の文と前記第2の言語の文とは、それぞれ固定的な文字列と、互に対応する、可変な部分を示す特定の文字列とを含み、

前記未登録語自動抽出装置は、

前記書換パターンと前記対訳用例文とを照合し、前記対訳用例文から書換パターンにマッチする対訳を抽出して、当該対訳のうち前記第1の言語の文中で、当該対訳にマッチした書換パターン内の前記第1の言語の文中の前記特定の文字列に対応する位置に存在する文字列を特定する用例文・書換パターン照合手段と、

前記用例文・書換パターン照合手段により特定された文字列を前記辞書内で検索し、登録されているか否かを判定する辞書検索手段と、

前記辞書検索手段により前記辞書に登録されていないと判定された前記第1の言語の文字列について、当該第1の言語の文字列を含む対訳と、当該対訳がマッチした書換パターンとに基づいて、当該対訳内の前記第2の言語の文中で、当該マッチした書換パターン内の前記第2の言語の文内の前記特定の文字列に対応する位置に存在する前記第2の言語の文字列を、前記第1の言語の文字列とともに前記辞書に登録すべき情報として特定する登録情報特定手段とを含む、未登録語自動抽出装置。

【請求項 2】

前記辞書検索手段により前記辞書に登録されていると判定された前記第 1 の言語の文字列について、前記辞書に当該第 1 の言語の文字列に関連して登録されている品詞分類又は意味分類とともに前記書換ボタンと関連付けて可変部情報として記憶する可変部情報記憶手段をさらに含み、

前記登録情報特定手段は、

前記辞書検索手段により前記辞書に登録されていないと判定された前記第 1 の言語の文字列について、当該第 1 の言語の文字列を含む対訳と、当該対訳がマッチした書換ボタンとに基づいて、当該対訳のうち、前記第 2 の言語の文中で、当該マッチした書換ボタン内の前記第 2 の言語の文内の前記特定の文字列に対応する位置に存在する文字列を特定することにより、当該第 1 の言語の文字列に対応する訳語として前記辞書に登録されるべき前記第 2 の言語の文字列を特定する訳語特定手段と、

10

前記第 1 の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて前記可変部情報記憶手段に記憶されている可変部情報の一部又は全部を所定の基準にしたがって選択することにより、前記第 1 の言語の文字列に対応する前記品詞分類又は意味分類を前記辞書に登録すべき付加情報として特定する付加情報特定手段とを含む、請求項 1 に記載の未登録語自動抽出装置。

【請求項 3】

前記付加情報特定手段は、前記第 1 の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて前記可変部情報記憶手段に記憶されている可変部情報の和集合を、前記第 1 の言語の文字列に関する前記付加情報として特定する手段を含む、請求項 2 に記載の未登録語自動抽出装置。

20

【請求項 4】

前記用例文・書換ボタン照合手段による照合の結果に前記第 1 の言語の文字列が出現するたびに、当該文字列の出現回数を計数する出現回数計数手段をさらに含み、

前記付加情報特定手段は、前記第 1 の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて前記可変部情報記憶手段に記憶されている可変部情報のうち、前記出現回数計数手段により計数された出現回数が所定回数以上のものの和集合を、前記第 1 の言語の文字列に関する前記付加情報として特定する手段を含む、請求項 2 に記載の未登録語自動抽出装置。

30

【請求項 5】

前記用例文・書換ボタン照合手段による照合の結果に前記第 1 の言語の文字列が出現するたびに、当該文字列の出現回数を計数する出現回数計数手段をさらに含み、

前記付加情報特定手段は、前記第 1 の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて前記可変部情報記憶手段に記憶されている可変部情報の数が所定のしきい値より大きいときに、当該可変部情報のうちで、前記出現回数計数手段により計数された出現回数が所定回数以上のものの和集合を、前記第 1 の言語の文字列に関する前記付加情報として特定する手段を含む、請求項 2 に記載の未登録語自動抽出装置。

【請求項 6】

コンピュータを、請求項 1～請求項 5 のいずれかに記載の未登録語自動抽出装置の全ての手段として機能させるための、未登録語自動抽出プログラム。

40

【請求項 7】

請求項 1～請求項 5 のいずれかに記載の未登録語自動抽出装置と、

前記未登録語自動抽出装置により前記辞書に登録すべきとして特定された情報を、当該辞書に登録する辞書登録手段とをさらに含み、未登録語自動登録装置。

【請求項 8】

コンピュータを、請求項 7 に記載の未登録語自動登録装置の全ての手段として機能させる、未登録語自動登録プログラム。

【発明の詳細な説明】

50

【技術分野】

【0001】

この発明は機械翻訳等の自然言語処理のための辞書作成技術に関し、特に、翻訳等のための辞書に登録されていない未登録な語を自動的に抽出又は登録可能な未登録語自動抽出・登録装置に関する。

【背景技術】

【0002】

機械翻訳のためには、機械可読な翻訳辞書が不可欠である。翻訳辞書に登録されている単語数が多いほど、翻訳の精度が高くなることが期待できる。しかし、辞書を保守するためには、文献を収集し、その中で辞書に未登録である語を探し、その単語に対する訳語を探して辞書に登録する、という手間がかかる。

10

【0003】

上記した作業には3段階の作業がある。第1段階は未登録語の発見という作業である。第2段階はその訳語の特定という作業である。第3段階は、未登録語と訳語、品詞分類、及び意味分類等の付加情報を辞書に登録する作業である。例えば、自動翻訳システムで辞書の不備により翻訳が失敗した場合、リライトの済んだ訳文と原文とを突合せながら、リライトされた部分の原文を突き止めて、辞書に登録可能な語句の形にし、訳語、品詞、意味分類等の付加情報を付与してはじめて辞書登録を行なうことができる。このような作業をきちんと続けていかなければ、例えば自動翻訳システムはまた同様の部分で翻訳に失敗し、再度リライトが必要ということになる。従来は、翻訳作業が自動化されても、リライトや辞書登録作業を含めた作業全体では、それほど大きな効率改善効果が見られない場合が大半であった。

20

【0004】

そのため、効率を改善するための一つの方策として、翻訳辞書への未登録語の登録を自動的に行なう技術が必要と考えられる。こうした機能を以下「自動登録」と呼ぶ。

【0005】

自動登録の技術として、下記特許文献1に記載されたものがある。特許文献1に記載された技術は、日本語と中国語との間で、固有名詞に限定して適用可能なものである。この技術では、日本語形態素辞書と日中翻訳辞書とを準備しておき、日本語形態素辞書には日本語側の語が登録されているが日中翻訳辞書には未登録である場合に、当該語を日中翻訳辞書に登録する。すなわち、原語と訳語の対として翻訳辞書に未登録であるもののうち、一部のみについてその登録を自動化する。

30

【0006】

特許文献2には、日英翻訳辞書において、カタカナ語と英語との対に限定して辞書への自動登録を行なう技術が開示されている。特許文献2に開示された技術では、英語辞書と日英翻訳辞書とを準備し、英語辞書には英語側の語が登録されているが日英翻訳辞書には未登録の場合に、当該英語の単語と、対応するカタカナ語とを新たな見出しとして日英翻訳辞書に登録する。

【特許文献1】特開平4-256171号公報

【特許文献2】特許第2995783号

40

【特許文献3】特開2003-6193号

【発明の開示】

【発明が解決しようとする課題】

【0007】

ところで、技術の進歩等により新造語が必要となり使用されるようになるのはもちろん、必要ではなくても「新鮮味」を打ち出そう等として次々に新しい言葉が造られていく。この傾向は、テレビジョン放送及び携帯電話等の情報伝達メディアの一般化及び日常化とともに強まる一方である。そのような「新鮮味」を出そうとする新造語の場合、普通の辞書に登録されているような語を意図的に避けて案出し、合成し、省略し、又は転用して造られる。いわゆる和製英語又は多言語からの合成等による新造語が濫用されていると見る

50

むきはあるものの、これら新造語を避けて現実社会の会話又は文章を成立させることは非常に困難である。

【 0 0 0 8 】

一方で、技術進歩は、こうした言語を処理して人間の活動を支援しようとする、いわゆる自然言語処理技術にも大きく影響している。例えば、処理に必要な辞書を装置として構成するための記憶装置は、日本で機械翻訳装置が商品化され始めた 1 9 8 0 年代半ばと比較して容量比で 1 0 万分の 1 以下の価格となっている。つまり、1 9 8 0 年当時の記憶装置の 1 0 0 万円相当の容量を用いて構成されていた翻訳辞書であれば、現在では記憶装置の 1 0 円相当の容量を用いて構成可能である。

【 0 0 0 9 】

こうしたことは、辞書を構成する記憶装置の容量を抑えるために辞書に登録する語を限定する等という、当時の「常識」を完全に陳腐化させている。すなわち、日本語形態素辞書には登録されているが日中翻訳辞書には未登録であるとか、英語辞書には登録されているが日英翻訳辞書には未登録であるとかいう辞書の構成の仕方は、それによって多少の容量が節約できるとしても、それによるコスト低減効果が、利用者の受ける不利益に対してあまりにも小さくなっている。

【 0 0 1 0 】

このため、現在のような状況下では、従来のように、一部の辞書にだけ単語が未登録である、というような辞書登録の不揃いを解消するような自動登録では、機械翻訳等の自然言語処理の有用性に対してはごく限定的な効果しか持ち得ないという問題がある。

【 0 0 1 1 】

それゆえに本発明の目的は、辞書に全く登録されていない未登録語を自動的に抽出することができる未登録語自動抽出装置、及びそうした未登録語を自動的に辞書に登録可能な未登録語自動登録装置を提供することである。

【課題を解決するための手段】

【 0 0 1 2 】

本発明の第 1 の実施の形態に係る未登録語自動抽出装置は、第 1 の言語の文と第 2 の言語の文との間の書換ボタンと、第 1 の言語と第 2 の言語との間の複数の対訳用例文とに基づいて、第 1 の言語の機械可読な辞書に未登録な語を自動的に抽出する未登録語自動抽出装置であって、書換ボタンと対訳用例文とを照合し、対訳用例文から書換ボタンにマッチする対訳を抽出して、当該対訳のうち第 1 の言語の文中で可変部を構成する文字列を特定するための用例文・書換ボタン照合手段と、用例文・書換ボタン照合手段により特定された文字列を辞書内で検索し、登録されているか否かを判定するための辞書検索手段と、辞書検索手段により辞書に登録されていないと判定された第 1 の言語の文字列について、当該第 1 の言語の文字列を含む対訳と、当該対訳がマッチした書換ボタンとに基づいて、当該第 1 の言語の文字列に対応する第 2 の言語の文字列を、第 1 の言語の文字列とともに辞書に登録すべき情報として特定するための登録情報特定手段とを含む。

【 0 0 1 3 】

用例文・書換ボタン照合手段が、書換ボタンと対訳用例文とを照合し、対訳のうち第 1 の言語の文中で可変部を構成する文字列を抽出する。辞書検索手段がこの文字列を辞書内で検索し、登録されているか否かを判定する。登録されていない場合、この文字列に対応する第 2 の言語の文字列を登録情報特定手段が特定する。第 1 の言語の文字列と、当該文字列に対応するとして特定された第 2 の言語の文字列とが、辞書に登録すべき情報として自動的に抽出される。この処理では、抽出されるべき文字列に関する何らかの情報が辞書に登録されている必要はない。人手によりこのような文字列の対を抽出する必要がなく、辞書に登録すべき見出しを特定するための労力が大きく改善される。その結果、辞書に全く登録されていない未登録語を自動的に抽出することができる未登録語自動抽出装置を提供できる。

【 0 0 1 4 】

好ましくは、未登録語自動抽出装置は、辞書検索手段により辞書に登録されていると判

10

20

30

40

50

定された第1の言語の文字列について、辞書に登録されている情報とともに書換ボタンと関連付けて可変部情報として格納するための可変部情報格納手段をさらに含む。登録情報特定手段は辞書検索手段により辞書に登録されていないと判定された第1の言語の文字列について、当該第1の言語の文字列を含む対訳と、当該対訳がマッチした書換ボタンとに基づいて、当該第1の言語の文字列に対応する訳語として辞書に登録されるべき第2の言語の文字列を特定するための訳語特定手段と、第1の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて可変部情報格納手段に格納されている可変部情報に基づいて、第1の言語の文字列に関して辞書に登録すべき付加情報を特定するための付加情報特定手段とを含む。

【0015】

10

同じ書換ボタンにマッチする対訳用例文は、互いに共通する種類の情報に関するものであると考えられる。すなわち、ある書換ボタンにマッチすることにより特定された文字列に対し付加すべき情報は、その書換ボタンにマッチした他の文字列であってかつ辞書から抽出された情報と類似していると考えられる。したがって、上記したように付加情報特定手段が特定する情報は、最初に特定された第1の言語の文字列に関しても妥当する情報と考えられ、これら当該第1の言語の文字列に付加して辞書に登録すべき情報を作成することで、辞書に登録すべき情報をさらに充実させることができる。

【0016】

さらに好ましくは、付加情報特定手段は、第1の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて可変部情報格納手段に格納されている可変部情報の和集合を、第1の言語の文字列に関する付加情報として特定するための手段を含む。

20

【0017】

ある書換ボタンにマッチすることにより特定された文字列に対し付加すべき情報の和集合をとることにより、最初に特定された第1の言語の文字列に関しても妥当する、漏れない情報を作成できる。

【0018】

未登録語自動抽出装置はさらに、用例文・書換ボタン照合手段による照合の結果に第1の言語の文字列が出現するたびに、当該文字列の出現回数を計数するための出現回数計数手段をさらに含んでもよく、付加情報特定手段は、第1の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて可変部情報格納手段に格納されている可変部情報のうち、出現回数計数手段により計数された出現回数が所定回数以上のものの和集合を、第1の言語の文字列に関する付加情報として特定するための手段を含んでもよい。

30

【0019】

所定回数出現する文字列に付加されている情報の和集合をとることにより、特定の書換ボタンにマッチする文字列として頻繁に出現する文字列に関して辞書から抽出された情報が得られる。こうした文字列に付随する情報は、書換ボタンにマッチする文字列の持つ性質をよく反映していると考えられる。したがって、特定された第1の言語の文字列にこれら和集合の情報を付すことにより、第1の言語の文字列に関する付加情報であって、かつ正確な情報を特定することができる。

【0020】

40

好ましくは、未登録語自動抽出装置は、用例文・書換ボタン照合手段による照合の結果に第1の言語の文字列が出現するたびに、当該文字列の出現回数を計数するための出現回数計数手段をさらに含む。付加情報特定手段は、第1の言語の文字列を含む対訳がマッチした書換ボタンと関連付けて可変部情報格納手段に格納されている可変部情報が所定のしきい値より大きいときに、出現回数計数手段により計数された出現回数が所定回数以上のものの和集合を、第1の言語の文字列に関する付加情報として特定するための手段を含む。

【0021】

第1の言語の文字列を含む対訳がマッチした書換ボタンと関連付けられている可変部情報が多数の時には、それらのうちで所定回数以上出現したものの和集合をとる。したがっ

50

て、第１の言語の文字列に関する情報として正確な情報が得られることが期待できる。一方、当該書換ボタンと関連付けられている可変部情報が少数のときには、それらの和集合をとることで、できるだけ漏れの少ない、包括的な情報を得ることができる。

【００２２】

本発明の第２の局面に係る未登録語自動抽出プログラムは、コンピュータにより実行されると、上記したいずれかの未登録語自動抽出装置として当該コンピュータを動作させるものである。したがって、これら未登録語自動抽出装置と同様の効果を得ることができる。

【００２３】

本発明の第３の局面に係る未登録語自動登録装置は、上記したいずれかの未登録語自動抽出装置と、未登録語自動抽出装置により辞書に登録すべきとして特定された情報を、当該辞書に登録するための辞書登録手段とを含む。

10

【００２４】

この未登録語自動登録装置により、辞書に登録すべき文字列と、その文字列に対する訳語等の情報とが自動的に抽出でき、かつ自動的に辞書に登録される。登録すべき文字列のピックアップ、訳語の特定、品詞分類又は意味分類等、文字列に付加すべき情報の作成、辞書への登録等の処理が自動的に行なわれる。その結果辞書に全く登録されていない未登録語を自動的に抽出し辞書に登録することができる未登録語自動登録装置を提供できる。

【００２５】

本発明の第４の局面に係る未登録語自動登録プログラムは、コンピュータにより実行されると、上記した未登録語自動登録装置として当該コンピュータを動作させる。したがって、これら未登録語自動登録装置と同様の効果を得ることができる。

20

【発明を実施するための最良の形態】

【００２６】

〔概略〕

以下に説明する本発明の一実施の形態では、完全な未登録語を、すでに翻訳された対訳文から自動的に語句として切出し、訳語及び意味的情報等を付加して翻訳辞書に登録することができる機構を設ける。この機構により、未登録語を探したり、訳語を決定したり、意味分類を付与したり、という一連の辞書登録作業が不要となり、自動翻訳システム等の自然言語処理を行なうシステムにおける作業効率を改善することができる。

30

【００２７】

〔機能的構成〕

以下、本実施の形態に係る辞書自動登録装置の構成について説明する。なお、以下の実施の形態では、説明のために日英翻訳を行なう自動翻訳システムを例にとる。原言語は日本語、目的言語は英語である。

【００２８】

図１に本実施の形態に係る未登録語自動登録装置３０の構成をブロック図形式で示す。図１を参照して、未登録語自動登録装置３０は、日英翻訳を行なう翻訳システムが備える、自動翻訳で使用する書換ボタンを記憶する記憶部３２と、予め準備された日本語と英語との多数の対訳用例文群３４と、上記した翻訳システムが備える翻訳用の辞書３６とともに用いられ、対訳用例文群３４中の用例文から辞書３６に未登録の単語（文字列）を切出し、所定の方法にしたがって訳語、意味情報等の付加情報を加えて辞書３６に自動的に登録する機能を持つ。

40

【００２９】

未登録語自動登録装置３０は、記憶部３２から書換ボタンを抽出するための書換ボタン抽出部５０と、書換ボタン抽出部５０により抽出された書換ボタンを記憶するための書換ボタン記憶部５２とを含む。

【００３０】

図２に、書換ボタン記憶部５２に記憶される書換ボタンの一例として、書換ボタン８０を示す。図２を参照して、書換ボタン８０は少なくとも、日本語の原文ボタン９０と、英

50

語の訳文ボタン 9 2 とを含む。各ボタンは少なくとも、文字列の情報を持つ固定部と、固定部に対する位置情報及び原文ボタンの可変部と訳文ボタンの可変部との対応情報を持つ可変部とを含む。

【 0 0 3 1 】

図 2 に示す例では、原文ボタン 9 0 は「私は...に ~ を送った」であり、その中で「私は」「に」「を送った」の部分が固定部である。訳文ボタン 9 2 は「I sent ~ to ...」であり、その中で「I sent」と「to」とが固定部である。両ボタンにおいて「...」で示される文字列 1 0 0 及び 1 0 6、並びに「~」で示される文字列 1 0 2 及び 1 0 4 がそれぞれ可変部である。原文と訳文とで文字列が共通している可変部は、互いに対応している。すなわち、文字列 1 0 0 及び 1 0 6 の部分が互いに対応しており、文字列 1 0 2 及び 1 0 4 の部分がそれぞれ対応している。さらにこの可変部は、原文ボタン及び訳文ボタンの双方において固定部の文字列中の特定位置に挿入されていることで、固定部との位置情報を持っている。

10

【 0 0 3 2 】

以上のとおり、書換ボタンとは、原文ボタンと訳文ボタンとの対を含む。原文ボタンと訳文ボタンとの各々は、固定文字列（固定部）と、可変部とを含む。書換ボタンは、各可変部と固定部との間の位置関係が分かるように構成されている。さらに書換ボタンは、原文ボタンと訳文ボタンとの間で、それぞれの可変部の対応関係が分かるように構成されている。

【 0 0 3 3 】

20

再び図 1 を参照して、未登録語自動登録装置 3 0 はさらに、書換ボタン記憶部 5 2 に接続され、対訳用例文群 3 4 に記憶された対訳用例文を書換ボタン記憶部 5 2 に記憶された書換ボタンと照合し、書換ボタンのいずれかに原文と訳文との双方がマッチする用例文を抽出し、さらに抽出された対訳文の日英両方の文において、書換ボタンにより指定された可変部に対応する文字列（以下「可変部文字列」と呼ぶ。）を特定する処理を行なうための用例文・書換ボタン照合部 5 4 と、用例文・書換ボタン照合部 5 4 が特定した可変部文字列と、対応する書換ボタンとを受けてこれらを解析し、当該書換ボタンにマッチした用例文の可変部文字列に関する情報（単語、品詞、意味情報等。以下「可変部情報」と呼ぶ。）を作成又は更新するための処理を行ない、当該可変部情報を書換ボタンと関連付けて書換ボタン記憶部 5 2 に記憶させる処理を行なうための可変部解析部 5 6 とを含む。可変部解析部 5 6 はこのために、当該可変部文字列が対応する書換ボタンに関する可変部情報として書換ボタン記憶部 5 2 に登録されているか否かを判定する機能を持つ。

30

【 0 0 3 4 】

未登録語自動登録装置 3 0 はさらに、可変部解析部 5 6 において、入力された可変部文字列が書換ボタン記憶部 5 2 中の該当書換ボタンの可変部情報に登録されていないと判定されたことに応答して、当該可変部文字列を未登録語として、未登録語情報（未登録語に対する訳語、並びに品詞及び意味情報等の付加情報）を書換ボタン記憶部 5 2 の当該書換ボタンの可変部情報に基づいて作成するための未登録語登録情報生成部 5 8 と、未登録語登録情報生成部 5 8 により作成された未登録語情報とともに、未登録語を翻訳システムの辞書 3 6 に自動的に登録するための辞書登録部 6 0 と、可変部解析部 5 6 によって書換ボタン記憶部 5 2 に記憶された書換ボタンに対する可変部情報として登録されていないと判定された可変部文字列に関し、この未登録語自動登録装置 3 0 が組込まれた自動翻訳システムの全ての辞書 3 6 を検索して登録されている情報を読出すための登録語辞書検索部 6 2 とを含む。

40

【 0 0 3 5 】

〔 コンピュータプログラムの構成 〕

後述するように、この未登録語自動登録装置 3 0 は、コンピュータと、当該コンピュータ上で実行されるコンピュータプログラムとにより実現できる。用例文・書換ボタン照合部 5 4、可変部解析部 5 6、未登録語登録情報生成部 5 8、辞書登録部 6 0 及び登録語辞書検索部 6 2 からなる未登録語自動登録処理を実現するためのコンピュータプログラムに

50

について、図3～図6を参照してその制御構造につき説明する。書換ボタン抽出部50の処理については、他の処理とは独立に実行でき、かつその処理内容については簡単であるのでここではその詳細については省略する。

【0036】

図3に、この未登録語自動登録処理プログラムの全体の処理フローを示す。図3を参照して、このプログラムの実行が開始されると、まず図1に示す対訳用例文群34を書換ボタン記憶部52に記憶された書換ボタンと照合し、書換ボタンのいずれかと原文ボタン及び訳文ボタンの双方において一致する用例文を特定し、その中の可変部文字列を抽出する処理と、抽出された可変部文字列の各々について可変部情報を作成する処理とからなる、可変部抽出・可変部情報作成処理を行なう(ステップ120)。可変部抽出・可変部情報作成処理が終了すると、ステップ120で抽出された可変部文字列のうち、翻訳システムの辞書36に登録されていない文字列について、作成された可変部情報から辞書とともに登録すべき未登録語情報とともに辞書36に登録する処理を行なう(ステップ122)。以上で処理終了である。

【0037】

次に、図4を参照して、図3のステップ120で実行される可変部抽出・可変部情報作成処理の詳細について説明する。まずステップ140で初期設定を行なう。この初期設定では、ワークエリアのクリア等を行なう。ステップ142で、書換ボタン記憶部52に記憶されている書換ボタンを讀出す。ステップ144で、全ての書換ボタンの讀出が終了したか否かについて判定する。もしも全ての書換ボタンの讀出が終了していればこの処理を終了する。全ての讀出が終了していなければ、ステップ146に進む。

【0038】

ステップ146では、対訳用例文群34中の用例文を讀出す。ステップ148で用例文を全て讀出したか否かを判定する。もしも全て讀出していたら制御はステップ142に戻り、次の書換ボタンについての処理を行なう。全て讀出していない場合にはステップ150に進む。

【0039】

ステップ150では、讀出された用例文が、現在の書換ボタンにマッチしているか否かについて判定する。この判定は、書換ボタンのうち原文ボタンに用例文の原文がマッチしているか否かと、訳文ボタンに用例文の訳文がマッチしているか否かとの、双方について行なわれる。もしもマッチしていなければ制御はステップ146に戻り、次の用例文に対する処理を行なう。もしもマッチしていれば、制御はステップ152に進む。

【0040】

ステップ152では、用例文のうち原文及び訳文の双方において、書換ボタンの可変部にマッチした文字列(可変部文字列)を特定する。

【0041】

図5を参照して、さらにステップ182で、書換ボタン記憶部52の当該書換ボタンの可変部情報に、この可変部文字列が登録されているか否かを判定する。もしも登録されていれば、ステップ184で当該文字列の出現数に1を加算し、図4のステップ146に戻る。登録されていない場合は、ステップ186に進む。

【0042】

ステップ186では、この文字列を翻訳システムの辞書36内で検索する。ステップ188で、翻訳システムの辞書36にこの文字列が登録されているか否かについて判定する。もしも登録されていればステップ190でこの可変部文字列と、翻訳システムの辞書36から当該文字列に対して検索された結果とを、書換ボタン記憶部52内の当該書換ボタンの可変部情報に追加する。この後制御はステップ146に戻る。一方、ステップ188でこの文字列が辞書36に登録されていないと判定されれば、ステップ192でこの可変部文字列に、未登録を示す情報を付加して、書換ボタン記憶部52内の当該書換ボタンに付随する可変部情報として書換ボタン記憶部52に追加する。このとき、当該可変部情報に関する出現回数を「1」に初期化しておく。この後制御はステップ146に戻り、次の

用例文について同じ処理を繰返す。

【 0 0 4 3 】

こうして、全ての書換ボタンと全ての用例文との組合せについて図 4 及び図 5 に示す処理が終了すると、制御は図 3 のステップ 1 2 2 に示す辞書更新処理に移る。その詳細について図 6 に示す。図 6 に示す処理のうち、ステップ 2 0 2 ~ ステップ 2 2 8 までは図 1 に示す未登録語登録情報生成部 5 8 に相当し、ステップ 2 3 0 が辞書登録部 6 0 に相当する。

【 0 0 4 4 】

図 6 を参照して、辞書更新処理では、最初にステップ 2 0 0 で初期設定を行なう。この初期設定では、ワークエリアのクリア、辞書 3 6 に登録すべき語のリスト（未登録語リスト）の初期化等を行なう。ステップ 2 0 2 で、書換ボタン記憶部 5 2 に記憶された書換ボタンを読出す。ステップ 2 1 0 では、全ての書換ボタンについて読出が終了したか否かを判定する。もしも終了していれば制御はステップ 2 3 0 に進む。さもなければ制御はステップ 2 1 2 に進む。

【 0 0 4 5 】

ステップ 2 1 2 では、この書換ボタンに付随する可変部情報中に、未登録を示す情報が付された可変部文字列があるか否かを判定する。もしもなければ制御はステップ 2 0 2 に戻り、次の書換ボタンについて処理を繰返す。もしも未登録を示す情報が付された可変部文字列が存在すれば、制御はステップ 2 1 4 に進む。

【 0 0 4 6 】

ステップ 2 1 4 では、未登録を示す情報が付された可変部文字列を読出す。ステップ 2 1 6 では、この書換ボタンに関する可変部情報中から、未登録を示す情報が付された可変部文字列を全て読出したか否かについての判定が行なわれる。もしも全て読出していれば制御はステップ 2 0 2 に戻り、次の書換ボタンについて処理が実行される。もしもまだ全て読出していなければ、ステップ 2 1 8 でその可変部文字列に対する訳語を特定する処理が実行される。

【 0 0 4 7 】

ステップ 2 1 8 で行なわれる処理の詳細について説明する。図 2 に示すように、書換ボタンは原文ボタンと訳文ボタンとを含み、その可変部は、可変部文字列の対応関係を示す情報を含む。この対応関係を用い、訳文の可変部文字列のうち、原文の可変部文字列に対応する文字列がどれかを判定することができる。この文字列が、原文の可変部文字列に対する訳語であると考えることができる。ステップ 2 1 8 ではこのようにして書換ボタンを使用して原文の可変部文字列に対する訳語を特定する。

【 0 0 4 8 】

続いてステップ 2 2 0 では、この書換ボタンに付随する可変部情報の数が所定のしきい値より大きいかな否かを判定する。可変部情報の数がしきい値より大きければステップ 2 2 6 に、さもなければステップ 2 2 4 に、それぞれ進む。

【 0 0 4 9 】

ステップ 2 2 6 では、この書換ボタンに付随する可変部情報のうち、出現回数がある一定値以上のものの可変部情報に含まれる品詞分類、意味分類等の情報の和集合をとり、それを当該可変部文字列に関する登録情報とする。一方、ステップ 2 2 4 では、全ての可変部情報に含まれる品詞分類、意味分類等の情報の和集合をとり、それを当該可変部文字列に関する登録情報とする。

【 0 0 5 0 】

このように本実施の形態では、第 1 の言語の文字列を含む対訳がマッチした書換ボタンと関連付けられている可変部情報が多数の時には、それらのうちで所定回数以上出現したものの和集合をとる。ある書換ボタンに頻繁にマッチする文字列に対し付加すべき情報の和集合をとることにより得られる情報は、最初に特定された第 1 の言語の文字列に関する情報としても妥当する可能性が高い。したがって、第 1 の言語の文字列に関する情報として正確な情報が得られることが期待できる。一方、当該書換ボタンと関連付けられている可変部情報

10

20

30

40

50

が少数のときには、それらの和集合をとることで、できるだけ漏れの少ない、包括的な情報を得ることができる。

【 0 0 5 1 】

もちろん、このような処理を行わず、どのようなときにも全ての可変部情報の和集合をとるようにしてもよい。この場合でも、ある書換ボタンにマッチすることにより特定された文字列に対し付加すべき情報の和集合をとることにより、最初に特定された第 1 の言語の文字列に関しても妥当する、漏れの無い情報を作成できると考えられる。

【 0 0 5 2 】

この後制御はいずれもステップ 2 2 8 に進み、当該可変部文字列と、その文字列に付随する未登録語情報とを未登録語リストに追加する。その後制御はステップ 2 1 4 に戻り、当該書換ボタンに付随する可変部情報のうち、次のものに対する処理を実行する。

【 0 0 5 3 】

こうして、全ての書換ボタンの全ての可変部情報に対し、ステップ 2 0 2 ~ ステップ 2 2 8 の処理を実行すると、制御はステップ 2 1 0 からステップ 2 3 0 に移る。

【 0 0 5 4 】

ステップ 2 3 0 では、ステップ 2 2 8 において未登録語情報が追加された未登録語リストの先頭から、翻訳システムの辞書 3 6 に追加していく処理が実行される。この処理が終了すると、辞書更新処理は終了である。

【 0 0 5 5 】

〔コンピュータハードウェア構成〕

上記したコンピュータプログラムを実行するコンピュータシステムの外観の一例を図 7 に、そのブロック図の例を図 8 に、それぞれ示す。

【 0 0 5 6 】

図 7 を参照して、このコンピュータシステム 3 3 0 は、F D (フレキシブルディスク) ドライブ 3 5 2 及び C D - R O M (コンパクトディスク読出専用メモリ) ドライブ 3 5 0 を有するコンピュータ 3 4 0 と、キーボード 3 4 6 と、マウス 3 4 8 と、モニタ 3 4 2 とを含む。

【 0 0 5 7 】

図 8 を参照して、コンピュータ 3 4 0 は、F D ドライブ 3 5 2 及び C D - R O M ドライブ 3 5 0 に加えて、C P U (中央処理装置) 3 5 6 と、C P U 3 5 6、F D ドライブ 3 5 2 及び C D - R O M ドライブ 3 5 0 に接続されたバス 3 6 6 と、ブートアッププログラム等を記憶する読出専用メモリ (R O M) 3 5 8 と、バス 3 6 6 に接続され、プログラム命令、システムプログラム、及び作業データ等を記憶するランダムアクセスメモリ (R A M) 3 6 0 とを含む。コンピュータシステム 3 3 0 はさらに、プリンタ 3 4 4 を含んでいる。

【 0 0 5 8 】

ここでは示さないが、コンピュータ 3 4 0 はさらにローカルエリアネットワーク (L A N) への接続を提供するネットワークアダプタボードを含んでもよい。

【 0 0 5 9 】

コンピュータシステム 3 3 0 に図 1 に示す未登録語自動登録装置 3 0 としての動作を行なわせるためのコンピュータプログラムは、C D - R O M ドライブ 3 5 0 又は F D ドライブ 3 5 2 に挿入される C D - R O M 3 6 2 又は F D 3 6 4 に記憶され、さらにハードディスク 3 5 4 に転送される。又は、プログラムは図示しないネットワークを通じてコンピュータ 3 4 0 に送信されハードディスク 3 5 4 に記憶されてもよい。プログラムは実行の際に R A M 3 6 0 にロードされる。C D - R O M 3 6 2 から、F D 3 6 4 から、又はネットワークを介して、直接に R A M 3 6 0 にプログラムをロードしてもよい。

【 0 0 6 0 】

このプログラムは、コンピュータ 3 4 0 にこの実施の形態の未登録語自動登録装置 3 0 としての動作を行なわせる複数の命令を含む。この動作を行なわせるのに必要な基本的機能のいくつかはコンピュータ 3 4 0 上で動作するオペレーティングシステム (O S) 又は

10

20

30

40

50

サードパーティのプログラム、もしくはコンピュータ 340 にインストールされる各種ツールキットのモジュールにより提供される。したがって、このプログラムはこの実施の形態の未登録語自動登録装置 30 を実現するのに必要な機能全てを必ずしも含まなくてよい。このプログラムは、命令のうち、所望の結果が得られるように制御されたやり方で適切な機能又は「ツール」を呼出すことにより、上記した未登録語自動登録装置 30 として動作を実行する命令のみを含んでいればよい。コンピュータシステム 330 の動作は周知であるので、ここでは繰返さない。

【0061】

〔動作〕

上に構成を説明した本実施の形態に係る未登録語自動登録装置 30 は以下のように動作する。図 1 を参照して、予め翻訳システムの記憶部 32 及び翻訳システムの辞書 36 がこの未登録語自動登録装置 30 に接続されているものとする。また、対訳用例文群 34 が予め準備されており、かつ各対訳用例文は互いに正しい訳となっているものとする。

【0062】

最初に、書換パタン抽出部 50 が翻訳システムの記憶部 32 から書換パタンを抽出し、図 2 に示すような形式で書換パタン記憶部 52 に格納する。全ての書換パタンを翻訳システムの記憶部 32 から抽出し書換パタン記憶部 52 に格納させると、書換パタン抽出部 50 は動作を終了する。

【0063】

続いて用例文・書換パタン照合部 54 が動作を開始する。具体的には、用例文・書換パタン照合部 54 は、まず書換パタン記憶部 52 から先頭の書換パタンを読出す。さらに対訳用例文群 34 から先頭の対訳用例文を読出す。用例文・書換パタン照合部 54 は、読出した書換パタンと対訳用例文とを比較し、対訳用例文の原文が書換パタンの原文パタンと一致し、かつ対訳用例文の訳文が書換パタンの訳文パタンと一致しているか否かを判定する。もしもそうした一致がなければこの対訳用例文は処理中の書換パタンとマッチしない。したがって用例文・書換パタン照合部 54 は次の対訳用例文を対訳用例文群 34 から読出し、同じ処理を実行する。

【0064】

もしも対訳用例文の原文が書換パタンの原文パタンと一致し、かつ対訳用例文の訳文が書換パタンの訳文パタンと一致している場合、用例文・書換パタン照合部 54 は、書換パタンと対訳用例文とを用いて、対訳用例文から原文の可変部文字列を抽出する。用例文・書換パタン照合部 54 はさらに、この原文の可変部文字列が、処理中の書換パタンに付随する可変部情報として登録されているか否かを判定する。もしも登録されていれば、当該可変部情報の出現数に 1 を加算し、次の対訳用例文に対する処理に移る。もしも登録されていなければ、次のような処理を実行する。

【0065】

すなわち用例文・書換パタン照合部 54 は、抽出された原文の可変部文字列を登録語辞書検索部 62 に与える。登録語辞書検索部 62 は、この可変部文字列を見出しに含む登録語を翻訳システムの辞書 36 内で検索し、当該文字列が登録されているか否かに関する情報と、登録されている場合にはそれら登録語に関する情報として検索された情報とを全て可変部解析部 56 に与える。

【0066】

可変部解析部 56 は、登録語辞書検索部 62 によって可変部文字列が翻訳システムの辞書 36 に登録されていないという情報が返された場合には、当該可変部文字列に未登録を示す情報を付加し、処理中の書換パタンに付随する可変部情報として書換パタン記憶部 52 に追加登録する。登録されている場合には、登録語辞書検索部 62 から与えられる情報をこの可変部文字列に付加し、処理中の書換パタンに付随する可変部情報として書換パタン記憶部 52 に登録する。

【0067】

ある書換パタンについて全ての対訳用例文の照合が終了すると、用例文・書換パタン照

10

20

30

40

50

合部 5 4 は、書換ボタン記憶部 5 2 から次の書換ボタンを読出し、同様の処理を実行する。この処理を全ての書換ボタンについて完了すると、可変部解析部 5 6 は未登録語登録情報生成部 5 8 に対して処理が終了したことを示す信号を与え、動作を終了する。

【 0 0 6 8 】

未登録語登録情報生成部 5 8 は、可変部解析部 5 6 から処理が終了したことを示す信号を受信すると、まず未登録語リストを初期化する。さらに未登録語登録情報生成部 5 8 は、書換ボタン記憶部 5 2 を走査し、未登録であることを示す情報が付加された可変部情報を持つ書換ボタンを順に読出す。未登録語登録情報生成部 5 8 は、読出した書換ボタンの可変部情報を利用して、読込んだ可変部情報のうちの可変部文字列についての未登録語情報を以下のようにして生成する。すなわち、未登録語登録情報生成部 5 8 は、入力された対訳文のうち、書換ボタンの訳文ボタンに含まれる可変部に相当する文字列のうち、原文において可変部文字列が占めていた可変部に対応する部分の文字列を可変部文字列に対する訳語として特定する。また未登録語登録情報生成部 5 8 は、処理中の書換ボタンに付随する可変部情報に含まれる品詞分類又は意味分類についてそれぞれ和集合をとることで、処理中の可変部文字列に関する品詞分類及び意味分類に関する情報を作成する。書換ボタンに付随する可変部情報の数がしきい値より大きい場合には、出現数が一定値以上の可変部情報のみを用いて同様に品詞分類及び意味分類に関する情報を作成する。未登録語登録情報生成部 5 8 は、こうして作成した未登録語情報を未登録語リストに追加する。

【 0 0 6 9 】

用例文・書換ボタン照合部 5 4 は、書換ボタンに複数の未登録を示す情報が付された可変部情報が存在する場合には、それらの各々について上記した処理を繰返す。

【 0 0 7 0 】

未登録語登録情報生成部 5 8 はさらに、次の書換ボタンを書換ボタン記憶部 5 2 から読出し、同様の処理を繰返す。全ての書換ボタンについての処理を終了すると、辞書登録部 6 0 に対して未登録語リストを渡すとともに、未登録語登録情報生成処理が完了したことを示す信号を辞書登録部 6 0 に与える。

【 0 0 7 1 】

辞書登録部 6 0 は、未登録語登録情報生成部 5 8 から未登録語登録情報生成処理が完了したことを示す信号が与えられると、未登録語登録情報生成部 5 8 から受け取った未登録語リストの先頭から順に処理し、可変部文字列と、その訳語、品詞分類、意味分類等の情報を翻訳システムの辞書 3 6 に追加していく。未登録語リストに含まれる全ての未登録語情報についてこの処理を終了すると、未登録語自動登録装置 3 0 は動作を終了する。

【 0 0 7 2 】

[具体的動作例]

上記した未登録語自動登録装置 3 0 の動作について、具体的な例を用いて説明する。ここでは、書換ボタンとして「最近～という新技術が話題になっている」という原文ボタンを持つ書換ボタン（これを「第 1 の書換ボタン」と呼ぶ。）が翻訳システムの記憶部 3 2 から抽出され書換ボタン記憶部 5 2 に登録されたものとする。また対訳用例文群 3 4 には、この書換ボタンにマッチする対訳用例文として、「最近ユビキタスという新技術が話題になっている。」という原文を持つ対訳用例文（これを第 1 の対訳用例文という。）と、「最近インターネットという新技術が話題になっている。」という原文を持つ対訳用例文（これを第 2 の対訳用例文と呼ぶ。）とが含まれているものとする。また前者では「ユビキタス」という語に対応する訳文の可変部には「ubiquitous computing」が含まれているものとする。また、「ユビキタス」については翻訳システムの辞書 3 6 には登録されておらず、「インターネット」については登録されているものとする。なお、「インターネット」については、品詞分類「普通名詞」、意味分類「技術用語」等の情報が翻訳システムの辞書 3 6 に登録されているものとする。

【 0 0 7 3 】

用例文・書換ボタン照合部 5 4 は、書換ボタン記憶部 5 2 の書換ボタンを読込み、対訳用例文群 3 4 に含まれる対訳用例文群と照合して原文の可変部文字列を抽出する。もしも

抽出された可変部文字列を含む可変部情報が処理中の書換パタンの可変部情報として書換パターン記憶部 5 2 に登録済みであれば、用例文・書換パターン照合部 5 4 は当該可変部情報の出現数に 1 を加算し、次の対訳文に処理を進める。また可変部情報が書換パターン記憶部 5 2 に登録済みでなければ、以下のような処理を行なう。

【 0 0 7 4 】

用例文・書換パターン照合部 5 4 が上記した第 1 の書換パターンを読込んだ場合、まず対訳用例文群 3 4 に含まれる第 1 の対訳用例文がこの書換パタンの原文パターンにマッチする。したがって用例文・書換パターン照合部 5 4 はこの対訳用例文から「ユビキタス」という語を可変部文字列として抽出する。可変部解析部 5 6 はこの語を含む可変部情報が書換パターン記憶部 5 2 に第 1 の書換パターンに付随するものとして登録されているか否かを判定する。ここでは登録されていないものとする。この場合可変部解析部 5 6 は、登録語辞書検索部 6 2 を用いて翻訳システムの辞書 3 6 を検索するが、対応する語は登録されておらず、情報は得られない。したがって可変部解析部 5 6 は、この可変部文字列に未登録を示す情報を付加し、書換パターン記憶部 5 2 に、第 1 の書換パターンに付随する可変部情報として追加登録する。

10

【 0 0 7 5 】

用例文・書換パターン照合部 5 4 はさらに、第 1 の書換パターンにマッチする対訳用例文として、上記した第 2 の対訳用例文を対訳用例文群 3 4 から抽出する。可変部解析部 5 6 がこの可変部文字列について登録語辞書検索部 6 2 を用いて翻訳システムの辞書 3 6 を検索した結果、既に述べたようにこの語が登録されており、その情報が登録語辞書検索部 6 2 から可変部解析部 5 6 に戻される。この情報には、品詞分類「普通名詞」、意味分類「技術用語」等の情報が含まれている。可変部解析部 5 6 は、これら情報を可変部文字列「インターネット」とともに、第 1 の書換パターンに付随する可変部情報として書換パターン記憶部 5 2 に追加する。

20

【 0 0 7 6 】

このようにして、他にも第 1 の書換パターンにマッチする対訳用例文が多数抽出され、可変部文字列について翻訳システムの辞書 3 6 に登録されているか否かが調べられる。辞書 3 6 に登録されている語については、辞書から得られた情報が第 1 の書換パターンに付随する可変部情報として書換パターン記憶部 5 2 に追加されていく。また、既に書換パターン記憶部 5 2 に登録された可変部情報については、対応する可変部文字列が対訳用例文群 3 4 から用例文・書換パターン照合部 5 4 により抽出されるたびに出現数が加算されていく。

30

【 0 0 7 7 】

用例文・書換パターン照合部 5 4 が全ての書換パターンについて処理を終了した後、未登録語登録情報生成部 5 8 による処理が開始される。未登録語登録情報生成部 5 8 は書換パターン記憶部 5 2 を走査し、未登録を示す情報が付加された可変部情報を抽出する。ここでは未登録語登録情報生成部 5 8 が上記した「ユビキタス」という語を抽出したものとする。未登録語登録情報生成部 5 8 は、この語が抽出された元の対訳用例文に対し、第 1 の書換パターンに含まれる原文パターンと訳文パターンとをあてはめ、「ユビキタス」に対応する訳文の可変部文字列として「ubiquitous computing」という文字列を特定する。さらに可変部解析部 5 6 は、「ユビキタス」に対する品詞分類及び意味分類等として、書換パターン記憶部 5 2 において第 1 の書換パターンに付随する可変部情報として登録されているものに含まれる品詞分類又は意味分類等の和集合を求め、その結果を「ユビキタス」に対する品詞分類及び意味分類とする。もしも第 1 の書換パターンに付随する可変部情報の数が所定のしきい値より大きければ、未登録語登録情報生成部 5 8 はそれら可変部情報のうちで出現数の大きなもののみを用いて、同様に品詞分類又は意味分類等の和集合を求めてその結果を「ユビキタス」に対する品詞分類及び意味分類とする。ここでは、「インターネット」という語について翻訳システムの辞書 3 6 から得られた品詞分類及び意味分類が「ユビキタス」という語の品詞分類及び意味分類として生成されたものとする。

40

【 0 0 7 8 】

未登録語登録情報生成部 5 8 は、このようにして得られた可変部文字列「ユビキタス」

50

と、対応する訳語「ubiquitous computing」、品詞分類「普通名詞」、意味分類「技術用語」等を含む未登録語情報を生成し、未登録語リストに追加する。

【0079】

この未登録語リストを得た辞書登録部60は、翻訳システムの辞書36に、「ユビキタス」という見出しを追加する。その訳語は「ubiquitous computing」であり、品詞分類は「普通名詞」を含み、意味分類は「技術用語」を含む。

【0080】

〔実施の形態の効果〕

以上のようにこの実施の形態によれば、対訳用例文群34を準備しておくことで、翻訳システムの辞書36に全く登録されていない語であっても自動的に追加登録していくことができる。しかも、見出しとなる文字列だけでなく、その訳語、品詞分類、意味分類等の情報が自動的に生成され、登録される。したがって、従来必要であった未登録語のピックアップ、訳語の決定、意味分類の付与等という作業が不要になる。その結果、非常に少ない労力で翻訳システムの辞書36を常に最新に保つことができる。

【0081】

こうして得られた辞書36は、翻訳システムに限らず、自然言語処理に関する処理において汎用的に利用できる。その結果、自然言語処理に利用する全ての辞書において未登録である完全な未登録語を自動的に辞書に登録可能な未登録語自動登録装置を提供できる。また、この実施の形態により、自然言語処理を行なうシステム、たとえば自動翻訳システム等の利用効率が向上するという効果が得られる。

【0082】

〔可能な変形例〕

なお、本実施の形態では、図1に示す辞書登録部60は自動的に翻訳システムの辞書36に対し追加登録を行なっている。もしも完全に自動的に追加登録した場合に不要な見出しが翻訳システムの辞書36に多く追加されるおそれがあれば、辞書登録部60による見出しの追加時に、ユーザに対して登録の可否を尋ねるダイアログを提示し、ユーザによる指示があつて初めて翻訳システムの辞書36に見出しを追加するようにしてもよい。この場合でも、登録すべき見出しの抽出、訳語の特定、品詞分類、意味分類等の情報を手作業で収集する必要はなく、従来に比して翻訳システムの辞書36の保守を効率的に行なうことができる。また、このように少ない労力で翻訳システムの辞書36の保守ができることから、翻訳システムの辞書36を使用する自然言語処理システム全体の性能を常に一定に保つことができ、新たな語の出現等が多数あつてもよい精度で自然言語処理を実行させることができる。

【0083】

なお、本実施の形態では書換ボタンとして図2に示すようなものを使用している。しかし書換ボタンはこのような例には限定されない。上記したように、書換ボタンとしては、原文及び訳文において、固定部の文字列と、固定部に対する可変部の位置情報及び原文と訳文との可変部の対応情報さえ特定できればよい。上記したものと別の書換ボタンの例を図9に示す。

【0084】

図9に示す例では、書換ボタンを構造化している。すなわち、この書換ボタン390は、原文の固定部の数を示す情報（原文固定部数）と、各原文固定部の文字列（原文固定部1, 2, 3）と、訳文の固定部の数を示す情報（訳文固定部数）と、各訳文固定部の文字列（訳文固定部1, 2）と、可変部の数（可変部数）と、各可変部の位置を示す情報（可変部1位置、可変部2位置）とを含む。

【0085】

このように書換ボタンを構造化しても上記した実施の形態と同様の処理を実現できる。なお、図9に示す例において、各部の名称は必ずしも書換ボタン中に含まれていなくてもよい。各部の間を所定の分離記号で分離し、データの位置でそれらの意味を特定するようにしてもよい。

【 0 0 8 6 】

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内の全ての変更を含む。

【図面の簡単な説明】

【 0 0 8 7 】

【図 1】本発明の一実施の形態に係る未登録語自動登録装置 3 0 のブロック図である。

【図 2】書換パタンの一例を示す図である。

【図 3】未登録語自動登録装置 3 0 をコンピュータで実現する場合のコンピュータプログラムの概略フローチャートである。 10

【図 4】図 3 のステップ 1 2 0 で実行される可変部抽出・可変部情報作成処理の前半のフローチャートである。

【図 5】可変部情報抽出・可変部情報作成処理の後半のフローチャートである。

【図 6】図 3 のステップ 1 2 2 で実行される辞書更新処理のフローチャートである。

【図 7】図 1 に示す未登録語自動登録装置 3 0 を実現するコンピュータシステム 3 3 0 の外観を示す図である。

【図 8】図 7 に示すコンピュータシステム 3 3 0 のブロック図である。

【図 9】書換パタンの他の一例を示す図である。

【符号の説明】 20

【 0 0 8 8 】

3 0 未登録語自動登録装置

3 2 翻訳システムの記憶部

3 4 対訳用例文群

3 6 翻訳システムの辞書

5 0 書換パタン抽出部

5 2 書換パタン記憶部

5 4 用例文・書換パタン照合部

5 6 可変部解析部

5 8 未登録語登録情報生成部 30

6 0 辞書登録部

6 2 登録語辞書検索部

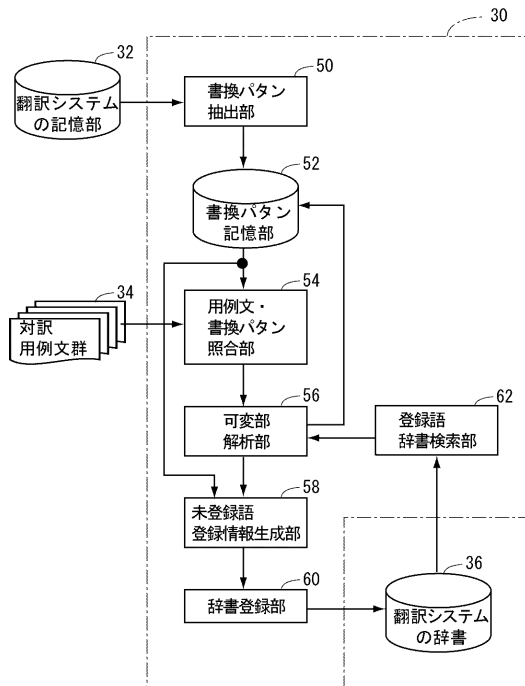
8 0 , 3 9 0 書換パタン

9 0 原文パタン

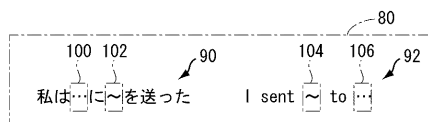
9 2 訳文パタン

1 0 0 , 1 0 2 , 1 0 4 , 1 0 6 可変部

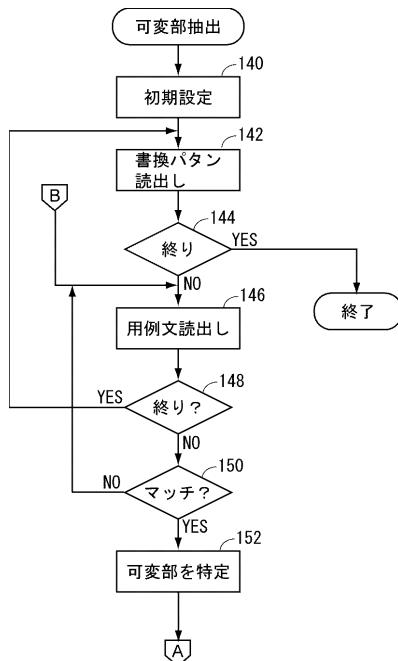
【図 1】



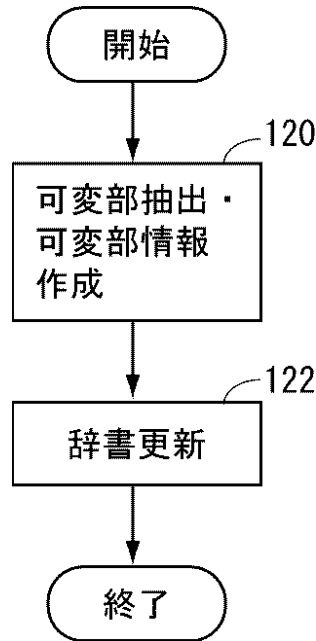
【図 2】



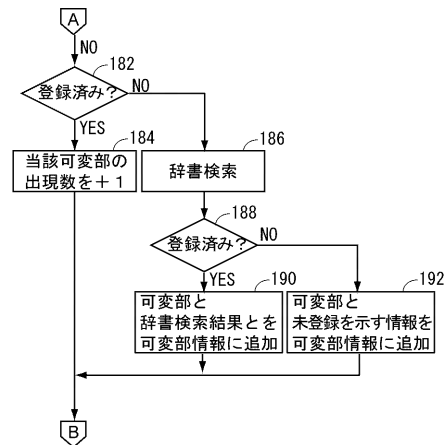
【図 4】



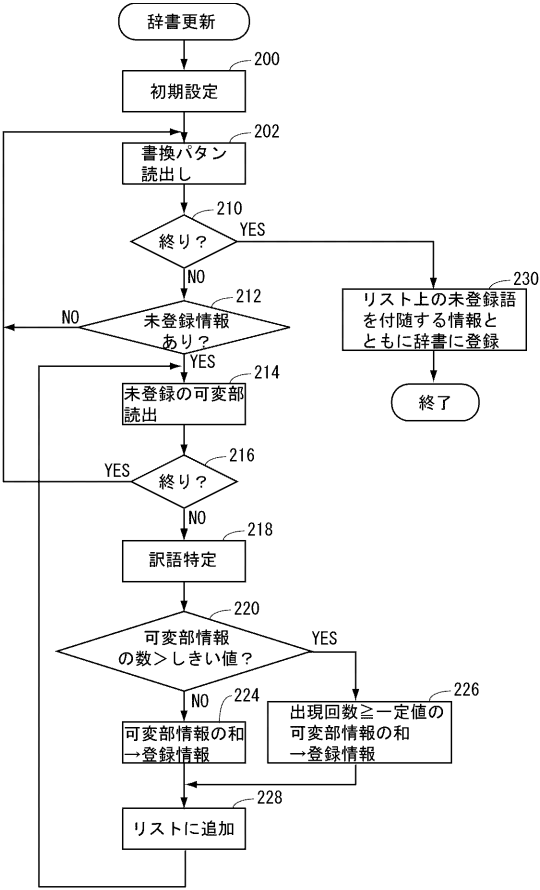
【図 3】



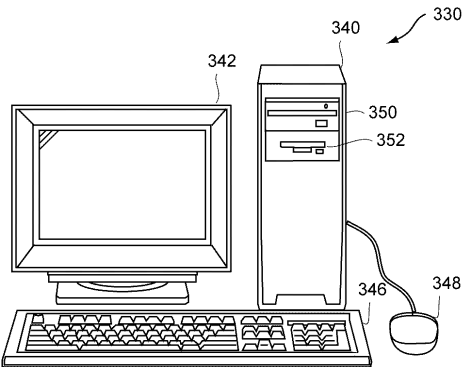
【図 5】



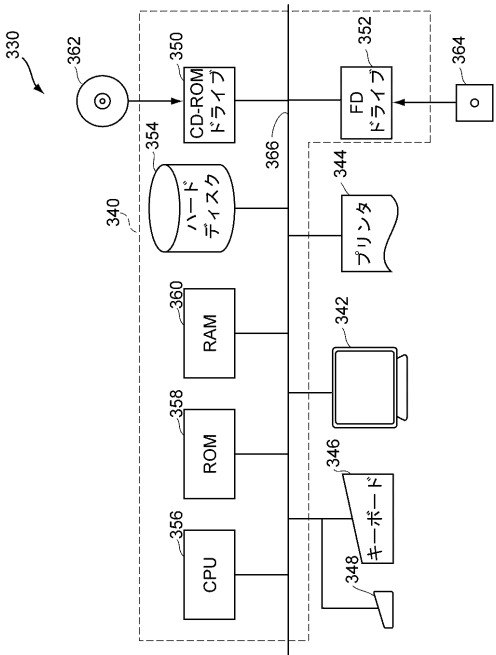
【図 6】



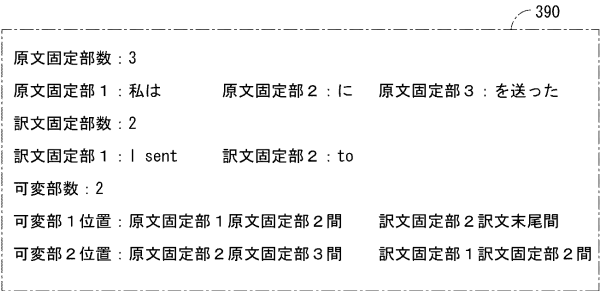
【図 7】



【図 8】



【図 9】



フロントページの続き

(56)参考文献 特開2004-348514(JP,A)
特開平06-052221(JP,A)

(58)調査した分野(Int.Cl., DB名)
G06F 17/27 - 17/28