

(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2007-25788

(P2007-25788A)

(43) 公開日 平成19年2月1日(2007.2.1)

(51) Int. Cl.	F I	テーマコード (参考)
G06F 17/28 (2006.01)	G06F 17/28 U	5B075
G06F 17/30 (2006.01)	G06F 17/30 170A	5B091

審査請求 未請求 請求項の数 22 O L (全 19 頁)

(21) 出願番号	特願2005-203157 (P2005-203157)	(71) 出願人	301022471 独立行政法人情報通信研究機構 東京都小金井市貫井北町4-2-1
(22) 出願日	平成17年7月12日 (2005.7.12)	(74) 代理人	100130498 弁理士 佐野 禎哉
特許法第30条第1項適用申請有り 進藤三佳、内元清貴、井佐原均、「コーパス・シソーラスに基づいた英語形容詞の意味拡張の調査・分析」、第19回ことば工学研究会資料、p. 23-33、社団法人人工知能学会、2005年3月5日発行 進藤三佳、内元清貴、井佐原均、「感覚・知覚領域に起源を持つ英語形容詞の意味拡張の調査・分析」、言語処理学会第11回年次大会 発表論文集、p1107-1110、言語処理学会、2005年3月15日発行	(72) 発明者	内元 清貴 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内	
	(72) 発明者	進藤 三佳 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内	
	(72) 発明者	井佐原 均 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内	
	Fターム(参考) 5B075 ND03 NK35 UU05		

最終頁に続く

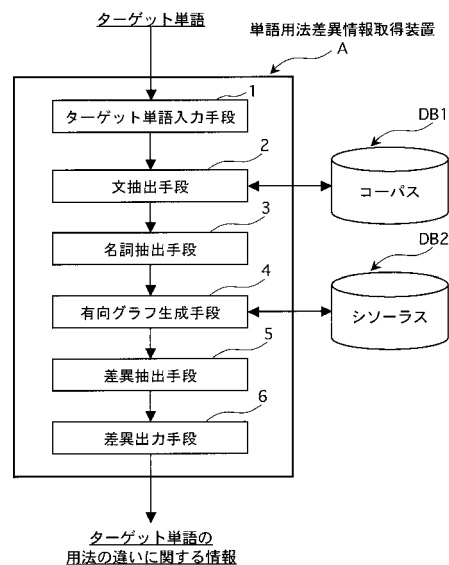
(54) 【発明の名称】 単語用法差異情報取得プログラム及び同装置

(57) 【要約】

【課題】同義語や類義語についてどのような意味の語と一緒に使われることが多いかという用法の違いに関する汎用性の高い情報を、高精度で自動的に得られるようにする。

【解決手段】複数の同義語、類義語の入力を受け付ける手段1、入力単語を含む文データをコーパスから抽出する手段2、この文データを構文解析してそれらに含まれる入力単語と文法的関係にある名詞を抽出する手段3、シソーラスからこの名詞とその上位概念のノードを抽出して有向グラフを入力単語ごとに生成する手段4、複数の有向グラフを比較して異なるノードを抽出する手段5、抽出した有向グラフの差異のノードを入力単語の用法の違いに関する情報として出力する手段6、を備えた装置Aを構成し、またコンピュータを同装置Aとして機能させるプログラムを用いる。

【選択図】 図1



【特許請求の範囲】

【請求項1】

同一又は類似の意味を有する複数のターゲット単語について、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続したコンピュータに、各ターゲット単語の用法の違いに関する情報を抽出し出力させるためのプログラムであって、当該コンピュータに、
複数のターゲット単語の入力を受け付けるターゲット単語入力ステップと、
前記コーパスにアクセスして、前記ターゲット単語入力ステップで受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出ステップと、
前記文抽出ステップで抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出ステップと、
前記シソーラスにアクセスして、前記名詞抽出ステップで抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成ステップと、
前記有向グラフ生成ステップで生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出ステップと、
前記差異抽出ステップで抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力ステップと、
を実行させることを特徴とする単語用法差異情報取得プログラム。

10

20

【請求項2】

前記差異抽出ステップにおいて、前記コンピュータに、
各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行させる、請求項1に記載の単語用法差異情報取得プログラム。

【請求項3】

前記ターゲット単語入力ステップで受け付けたターゲット単語が3つ以上の場合、
前記差異抽出ステップにおいて、前記コンピュータに、
特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行させる、請求項1又は2に記載の単語用法差異情報取得プログラム。

30

【請求項4】

前記コンピュータに、
前記名詞抽出ステップにおいて、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行させ、
前記有向グラフ生成ステップにおいて、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行させ、
前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行させる、請求項1乃至3の何れかに記載の単語用法差異情報取得プログラム。

40

【請求項5】

前記名詞抽出ステップにおいて、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用し、
前記コンピュータに、
前記有向グラフ生成ステップにおいて、生成する有向グラフにおいて前記名詞に対応する

50

ノードに前記頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行させ、前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された比較対象となる2つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行させる、請求項4に記載の単語用法差異情報取得プログラム。

【請求項6】

前記差異抽出ステップにおいて、前記コンピュータに、
比較対象となる2つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を遞減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行させる、請求項5に記載の単語用法差異情報取得プログラム。

10

20

【請求項7】

前記頻度に関するデータとして、前記頻度比率に代えて、頻度の値自体を適用している請求項5又は6に記載の単語用法差異情報取得プログラム。

【請求項8】

前記コンピュータに、
前記差異抽出ステップにおいて、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノードをさらに抽出する処理を実行させ、
前記差異出力ステップにおいて、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項4乃至7の何れかに記載の単語用法差異情報取得プログラム。

30

【請求項9】

前記コンピュータに、
前記差異出力ステップにおいて、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項1乃至8の何れかに記載の単語用法差異情報取得プログラム。

【請求項10】

前記コンピュータに、
前記差異出力ステップにおいて、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項1乃至8の何れかに記載の単語用法差異情報取得プログラム。

40

【請求項11】

前記ターゲット単語入力ステップにおいて入力受付可能なターゲット単語の品詞を、形容詞又は動詞に制限している、請求項1乃至10の何れかに記載の単語用法差異情報取得プログラム。

【請求項12】

プログラムに従って作動するコンピュータにより構成され、入力された同一又は類似の意味を有する複数のターゲット単語について、各ターゲット単語の用法の違いに関する情報を抽出し出力させる単語用法差異情報取得装置であって、前記コンピュータは、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続しており、

50

複数のターゲット単語の入力を受け付けるターゲット単語入力手段と、
前記コーパスにアクセスして、前記ターゲット単語入力手段で受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出手段と、
前記文抽出手段で抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出手段と、
前記シソーラスにアクセスして、前記名詞抽出手段で抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成手段と、
前記有向グラフ生成手段で生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出手段と、
前記差異抽出手段で抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力手段と、
を具備してなることを特徴とする単語用法差異情報取得装置。

10

【請求項 13】

前記差異抽出手段において、
各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行する、請求項 12 に記載の単語用法差異情報取得装置。

【請求項 14】

前記ターゲット単語入力手段で受け付けたターゲット単語が 3 つ以上の場合、
前記差異抽出手段において、特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行する、請求項 12 又は 13 に記載の単語用法差異情報取得装置。

20

【請求項 15】

前記名詞抽出手段において、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行し、
前記有向グラフ生成手段において、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行し、
前記差異抽出手段において、前記有向グラフ生成手段で生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行する、請求項 12 乃至 14 の何れかに記載の単語用法差異情報取得装置。

30

【請求項 16】

前記名詞抽出手段において、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用し、
前記有向グラフ生成手段において、生成する有向グラフにおいて前記名詞に対応するノードに前記頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行し、
前記差異抽出手段において、前記有向グラフ生成手段で生成した重み付けが施された比較対象となる 2 つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行する、請求項 15 に記載の単語用法差異情報取得装置。

40

【請求項 17】

50

前記差異抽出手段において、比較対象となる2つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を減減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行する、請求項16に記載の単語用法差異情報取得装置。

10

【請求項18】

前記頻度に関するデータとして、前記頻度比率に代えて、頻度の値自体を適用している請求項16又は17に記載の単語用法差異情報取得装置。

【請求項19】

前記差異抽出手段において、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノードをさらに抽出する処理を実行し、

前記差異出力手段において、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項15乃至18の何れかに記載の単語用法差異情報取得装置

。

【請求項20】

前記差異出力手段において、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項12乃至19の何れかに記載の単語用法差異情報取得装置。

20

【請求項21】

前記差異出力手段において、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項12乃至19の何れかに記載の単語用法差異情報取得装置。

【請求項22】

前記ターゲット単語入力手段において入力受付可能なターゲット単語の品詞を、形容詞又は動詞に制限している、請求項12乃至21の何れかに記載の単語用法差異情報取得装置

30

。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、複数の類義語の用法の相違を自動的に解析するためのプログラム及び装置に関するものである。

【背景技術】**【0002】**

複数の同義語や類義語を文中（発話文、記述文とも）で正確に使い分けることは、その言語を外国語として学習する者が難しいと感じるだけでなく、その言語を母国語として日常的に使っている者にとっても存外難しいものである。近年では、ワードプロセッシングソフトウェア（ワープロソフト）、外国語学習ソフトウェア翻訳ソフトウェア等が日常的に使用されるようになっており、これらソフトウェアには様々な入力・編集・出力支援機能が備えられていることがあるが、ユーザが類義語を用例の違いによつて的確に使い分けられるように自動的に峻別したり指摘したりすることは実現されていない。

40

【0003】

一つの試みとして、単語同士の共起の度合いによつて、その単語がどのような単語と共に使われやすいか、という言語学的な研究がなされている（非特許文献1参照）。この研究では、入力文を構文解析し、一文に出現する構文的に關係のある単語同士について、その際に偶然性を排除する処理を施したうえで共起のスコアを計測し、そのスコアをソート

50

することによって高スコアの単語同士は構文的に関係が深いと推定される、というものである。この場合、複数の類義語をターゲットとしてそれぞれ共起スコアが高い単語を抽出すれば、どの単語にはどのような用例があるかを推測することは可能である。

【非特許文献1】Stefan Th. Griesand Anatol Stefanowitsch, "Extending collocation analysis: A corpus-based perspective on alternations", *International Journal of Corpus Linguistics*, 9:1, 2004

【発明の開示】

【発明が解決しようとする課題】

【0004】

しかしながら、上記文献に記載の方法では、具体的な単語同士が一文に出現しているか否かという情報のみを以て、その単語同士に関係があると推定しているため、共起相手の単語が異なれば、ターゲットとする単語を用いることが正しい用法であるのかは不明である。換言すれば、上記文献に記載の方法では、その文においてターゲットとする単語をその類義語に置換しても正しい用法であるといえるのか、どのような単語をターゲットとすれば正しい用法となるのか判断することができない。そのため、如何なる文であれば複数の同義語のうちどれを用いるべきか、というような精度の高い情報を得ることが求められている。

【0005】

本発明は、このような課題を解決するためになされたものであり、同義語や類義語についてどのような意味の語と一緒に使われることが多いかという用法の違いに関する汎用性の高い情報を、高精度で自動的に得られるようにすることを目的としている。

【課題を解決するための手段】

【0006】

すなわち本発明の単語用法差異情報取得プログラムは、同一又は類似の意味を有する複数のターゲット単語について、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続したコンピュータに、各ターゲット単語の用法の違いに関する情報を抽出し出力させるためのプログラムである。そして、このコンピュータに、複数のターゲット単語の入力を受け付けるターゲット単語入力ステップと、コーパスにアクセスして、ターゲット単語入力ステップで受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出ステップと、文抽出ステップで抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出ステップと、シソーラスにアクセスして、名詞抽出ステップで抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成ステップと、有向グラフ生成ステップで生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出ステップと、前記差異抽出ステップで抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力ステップと、を実行させることを特徴としている。

【0007】

ここで、前記入力ステップで入力されるターゲット単語の数は、2つ以上であれば特に制限されることはないが、本発明の趣旨からそれらの単語は何れも同一又は類似の意味を有している同義語又は類義語であるとする。前記名詞抽出ステップにおける構文解析処理は、一般に知られている形態素解析ソフトウェアや構文解析ソフトウェアの機能を利用することができる。また、同ステップにおいて抽出される名詞が、「ターゲット単語と文法的関係にある」とは、その名詞とターゲット単語とが、修飾・被修飾関係や述語・項関係（例えば主述関係等）を意味し、上記構文解析処理によって得られる情報からこれらの関係を特定することができる。さらに、前記差異抽出ステップにおいて抽出される「異なるターゲット単語の有向グラフ間において異なるノード」とは、1つ以上のノードであれば

その数は特に制限のない限り限定されず、比較される有向グラフ同士で異なると判断されるノードのうち最上位のノードのみを示す場合があり得、その最上位のノード以下のノードを含む場合もあり得る。このようなプログラムは、単独のプログラムとして利用できることはもちろんであるが、他のプログラム、例えば文書入力プログラムや翻訳プログラム、言語学習プログラム等の一部に組み込んで利用することもできる。なお、入力される複数のターゲット単語は、同一であっても許容される。すなわち、入力される複数の同義語又は類義語には、同一の単語も含まれる。例えば、同一のターゲット単語が入力された場合、それらターゲット単語同士で、修飾・被修飾の関係にある名詞と主述関係にある名詞とを比較して、どのような意味の名詞と文法的関係があるときにはどちらの関係を使いやすいか、といった用法の違いを得ることも可能である。このことは、例えば、ある形容詞がターゲット単語である場合、意味的關係にある名詞が人ならば叙述用法が使われることが多い、というような情報を得ることも有用であることを意味する。 10

【0008】

このような本発明のプログラムを利用することで、これまでは得られなかった同義語や類義語の用法の違いに関する情報を、各ターゲット単語間の有向グラフ上におけるノードの差異として得ることができ、単語の正しい用法を学習したり、入力文の自動校正に利用するなどの場面で役立てることができる。特にコーパスやシソーラスは、一般に知られているものを適宜利用することができるが、それらの規模がある程度大きいほど、また格納されるデータの分野に偏りが少ないほど、出力されるターゲット単語の用法の違いに関する情報の信頼度を向上することができる。なお、差異出力ステップにおける情報の出力は、ディスプレイ等の表示装置に情報を表示させる態様、プリンタ等の印刷装置に情報を印刷させる態様、他のコンピュータに情報を送信する態様など、種々の態様を適宜選択することができる。 20

【0009】

上述のような本発明のプログラムは、例えば、前記差異抽出ステップにおいて、コンピュータに、各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行させるものとすることができる。このようにすることで、ターゲット単語ごとに生成された有向グラフの異なる部分のノードを容易に得ることができる。また同時に、共通する部分のノードを得ることも可能となる。 30

【0010】

特に、ターゲット単語入力ステップで受け付けたターゲット単語が3つ以上の場合は、前記差異抽出ステップにおいて、コンピュータに、特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行させることによって、3つ以上の有向グラフの比較を容易なものとするすることができる。

【0011】

また、上述した本発明の何れかの態様においては、コンピュータに、前記名詞抽出ステップにおいて、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行させ、前記有向グラフ生成ステップにおいて、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行させ、前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行させるようにすることが好適である。このような重み付き有向グラフを用いれば、得られる情報に優劣を付けて適宜必要なものだけを取り出すことを容易にすることができる。 40

【0012】

特にこの場合、前記名詞抽出ステップにおいて、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用することが適している。こうした場合、コンピュータに、前記有向グラフ生成ステップにおいて、生成する有向グラフにおいて名詞に対応するノードに頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行させ、差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された比較対象となる2つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行させるとよい。このようにすることで、上記の重み付けを、頻度比率によって信頼性の高いものとする

10

【0013】

さらにこの場合には、前記差異抽出ステップにおいて、前記コンピュータに、比較対象となる2つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を逡減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行させることも可能である。このようにすることで、得られる膨大な差異に関する情報から、その差異が急激に変化するノードを特定して、より適切な差異点を見つけ出し、同義語や類義語の用法が大きく異なるノードを適切に特定することができる。なお、頻度比率の逡減処理は、頻度比率が0になるまで適宜の数値ごとに行ってもよいし、機械学習を導入して入力された単語によって低減率を適宜調整しても良い。また、共通するノードの割合と前回の工程で得られた共通するノードの割合の比較処理は、両者の差やその絶対値によって行ってもよいし、比によって行ってもよい。

20

【0014】

なお、上記の重み付け以降の処理においては、頻度に関するデータとして、頻度比率に代えて、頻度の値自体を適用してもよいのはいうまでもない。

30

【0015】

さらに、頻度（頻度比率を適用する場合を含む）に基づく処理においては、コンピュータに、前記差異抽出ステップにおいて、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノード（異なるノード全部 or 最上位の異なるノードに限定可）をさらに抽出する処理を実行させ、前記差異出力ステップにおいて、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行させるようにしてもよい。ここで、頻度に基づく重みの大きい方から順に所定数のノードを抽出する際には、異なるノード全部を抽出してもよいし、異なるノードのうち最上位の異なるノードに限定して抽出することも妨げない。

40

【0016】

以上の発明の各態様においては、コンピュータに、前記差異出力ステップにおいて、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行させる場合に、用法が異なると判断される最上位の概念を特定して、ユーザに理解しやすい有用な情報を得ることができる。

【0017】

また同様に、コンピュータに、前記差異出力ステップにおいて、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行させることも可能であり、この場合は、用法が異

50

なる部分と共通する部分の境界を適切に与えることができる。なおこの場合、共通するノードのうち最下位のノードを上述のように頻度で重み付けして所定数だけ出力しても良いし、当該最下位のノードを全て出力しても構わない。

【0018】

以上に述べた本発明においては、特に前記ターゲット単語入力ステップにおいて入力受付可能なターゲット単語の品詞を、同義語や類義語の使い分けが難しい形容詞又は動詞に制限している場合に、特に有用である。

【0019】

また、本発明に係る単語用法差異情報取得装置は、上述したようなプログラムに従って作動するコンピュータにより構成され、入力された同一又は類似の意味を有する複数のターゲット単語について、各ターゲット単語の用法の違いに関する情報を抽出し出力させる単語用法差異情報取得装置である。このコンピュータは、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続してある。そして、複数のターゲット単語の入力を受け付けるターゲット単語入力手段と、コーパスにアクセスして、ターゲット単語入力手段で受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出手段と、文抽出手段で抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出手段と、シソーラスにアクセスして、前記名詞抽出手段で抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成手段と、有向グラフ生成手段で生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出手段と、差異抽出手段で抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力手段と、を具備してなることを特徴とするものであり、上述したプログラムにより動作するコンピュータとして、上述と同様の作用効果が得られる。

【0020】

このような装置は、一般的には、コンピュータが備えるハードディスクデバイス等の記憶装置に前記プログラムを格納しておき、必要に応じてメモリに当該プログラムを読み出してやCPU（中央演算装置）による処理を行い、各種出入力デバイスを作動させる態様が採用されるが、例えばネットワークサーバに前記プログラムを格納しておいて、そのサーバにアクセスした端末コンピュータがプログラムに従って作動するような態様を採用しても良い。

【0021】

以下、詳述しないが、上記の各態様のプログラムによって作動する本発明の単語用法差異情報取得装置は、対応するプログラムと基本的には同様の作用効果が得られるものである。

【発明の効果】**【0022】**

本発明によれば、単に同義語や類義語がどのような単語と共起して文中で用いられるかという情報を得るのではなく、複数の同義語や類義語についてどの語がどのような意味を表す文において用いるのが適切か、という単語間での用法の違いを自動的に精度良く得ることが可能である。したがって、個々の単語に特化した情報ではなく、ターゲットとした単語が用いられる文について汎用的な情報が得られることとなり、その応用範囲も広いといえる。そして、本発明を単独で用いる場合には、同義語や類義語の用法の違いに関する情報が得られるので、語学学習や単語の用法チェックに利用することができ、また例えば本発明のプログラム又は装置を、外国語学習ソフトウェア、文書入力ソフトウェア、翻訳ソフトウェアやそれらを組み込んだ装置に用いる場合には、単に文法的な誤りを抽出、指摘する程度にとどまらず、用例に基づいた入力・編集支援として活用することが可能である。

【発明を実施するための最良の形態】

【0023】

以下、本発明の一実施形態を、図面を参照して説明する。

【0024】

この実施形態は、本発明の単語用法差異情報取得プログラム（以下、「本プログラム」と呼ぶ場合がある）に従って作動する単語用法差異情報取得装置A（以下、「本装置A」と呼ぶ場合がある）である。本装置Aを実現するコンピュータは、一例としてごく一般的なパーソナルコンピュータで足りる。そこで本実施形態では、当該コンピュータとして、汎用パーソナルコンピュータを採用している。

【0025】

斯かるパーソナルコンピュータのハードウェア構成はごく一般的なものであるため詳述しないが、通常は、中央演算装置（CPU）等のプロセッサ、メインメモリ（RAM）、ハードディスクドライブ（HDD）等の補助記憶デバイス等を通信線（バス線等）で接続した態様で備えており、これらがシステムコントローラやI/Oコントローラ等に制御されて連携して動作する。さらにパーソナルコンピュータは、入力デバイスとしてキーボードやマウス等のポインティングデバイス、出力デバイスの一形態として情報を画像や映像として表示するディスプレイ及びグラフィックチップ等の表示制御デバイス、外部機器とのデータ授受を行うための通信デバイス等を備えているものとする。そして、通常はHDD等にオペレーションソフトウェア（OS）に加えて本プログラムがインストールされており、当該パーソナルコンピュータを本装置Aとして機能させる場合には、適宜RAMに読み出した本プログラムに基づくCPUの処理により各部を動作させる。この処理工程で生成されるデータや外部から入力又は取得されたデータは、一時的にRAM等に蓄えられて以降の処理に利用される。

【0026】

特に本実施形態では、前記パーソナルコンピュータに、例文データベースであるコーパスDB1と、語と語の上位下位概念の関係が記述されたデータベースであるシソーラスDB2を、前記通信デバイスを通じて接続しており（図1参照）、必要に応じてこれらにアクセスして情報を検索し読み出してくることができるようにしている。ここで、本実施形態では、コーパスDB1の一例として、大規模コーパスとしてよく知られている「BNC ; (The British National Corpus ; URL : <http://www.natcorp.ox.ac.uk/>)」をインターネットを通じて利用し、シソーラスDB2の一例として、語の意味的な上位下位関係をネットワーク構造のデータベースとして持つ大規模シソーラスとして知られる「WordNet 2.0」を利用するものとするが、本発明にはこれら以外のコーパスやシソーラスを用いることができるのはいうまでもない。

【0027】

さて、本装置Aは、本プログラムに従って作動することで、図1に概略的な機能構成図を示すように、ターゲット単語入力手段1、文抽出手段2、名詞抽出手段3、有向グラフ生成手段4、差異抽出手段5、差異出力手段6の各機能を有することになる。以下、本装置Aにおける各手段の機能と処理手順を、適宜具体例を示しながら説明する。特に具体例としては、説明の理解を容易にするため、入力されるターゲット単語として、英語の「keen」と「eager」の2単語を採用する。両単語は、辞書においてほぼ同義の意味が記載されており、(「keen : very interested, eager or wanting (to do) something very much」, 「eager : wanting much to do or have esp. something interesting or enjoyable」, CIDE (The Cambridge International Dictionary of English) による)、共に「強く欲求する」という意味を表す形容詞であり、両者の文の意味に応じた正確な使い分けは難しく、特に英語を母国語としない者にとってはその困難性は顕著である。なお、以下の説明で言及する図面のうち、図2, 図3, 図5, 図7は、本装置Aによる情報処理手順を示す概略的なフローチャートである。

10

20

30

40

50

【0028】

本装置 A による大まかな情報処理手順は、図 2 に示すようなものである。すなわち、ユーザによるキーボード等の入力デバイスの操作により、ターゲット単語入力手段 1 がターゲット単語（例えば上記「keen」と「eager」）の入力を受け付ける（S1）。そして、文抽出手段 2 がコーパス DB 1 からそのターゲット単語を含む文データを抽出する（S2）。次に、名詞抽出手段 3 が、各文データを構文解析し、ターゲット単語と文法的関係にある名詞を抽出する（S3）。さらに、有向グラフ生成手段が、各ターゲット単語ごとに、それらの名詞をキーワードとしてシソーラス DB 2 を検索して有向グラフを生成する（S4）。そして、差異抽出手段 5 が、各有向グラフ（ターゲット単語の数に対応する。上記例のようにターゲット単語が 2 つの場合は 2 つの有向グラフ）を比較し、各ターゲット単語の用法の違いに関する情報を抽出し（S5）、最後に差異出力手段 6 が、抽出されたこの用法の違いに関する情報を、前記ディスプレイに表示するなどして出力する（S6）。以下に、各手段におけるデータ処理の内容等を詳細に説明する。

10

【0029】

まず、ターゲット単語入力手段 1 は、入力を受け付け可能な単語の品詞を、形容詞又は動詞に制限しているものとする。これを実現するべくパーソナルコンピュータにターゲット単語入力手段 1 の機能を付与するには、例えば本プログラムの一部に、品詞解析プログラムを組み込んでおくことができる。すなわち、入力された単語で辞書データベース等を検索し、その単語の品詞が形容詞又は動詞であれば入力を受け付け、それら以外であれば入力の受け付けを拒否するようにすることができる。

20

【0030】

また、文抽出手段 2 では、入力を受け付けたターゲット単語でコーパス DB 1 を検索して、ターゲット単語を含む文データを全て抽出するものとする。これは、ターゲット単語を含む限定用法、叙述用法の両方の文についてできるだけ数多く且つ万遍なく収集することで、最終的に出力する「用法の違いに関する情報」の信頼性を向上させるためである。

【0031】

名詞抽出手段 3 による処理工程 S3 では、具体的には例えば図 3 に示すような処理が実行される。なお、以下の工程は、各ターゲット単語ごとに且つ各文データについて実行される。すなわち、まず、コーパス DB 1 から抽出された文データについて、まず文データを形態素解析する（S31）。この形態素解析処理では、具体的には、a. 文データを単語毎に区切り、b. 各単語から原形を抽出して、c. 各単語に品詞データを付与する。次に、形態素解析された文データを構文解析することにより、文データが有する文の統語構造を取得し（S32）、ターゲット単語と所定の文法関係（修飾・被修飾関係、又は主述関係等の述語・項関係）にある名詞データを特定し、これを抽出する（S33）。以上の処理を各ターゲット単語ごとに且つ各文データについて実行することで、抽出された名詞データには、各ターゲット単語ごとに、当該名詞が文データに出現した回数を示す頻度データが付与される。

30

【0032】

ここで構文解析処理について詳述すると、単語に関する品詞データの他にも、シソーラス（上記シソーラス DB 2 と同じでなくてもよい）を検索して得られる上位概念の情報、格フレーム辞書等の辞書を検索して得られる述語と項との関係を表すデータ（例えば動詞の場合、その目的格にはどのような意味の名詞を取り得るか、という情報）、必要に応じて設定された処理ルールや機械学習により設定された処理ルール、適宜のデータベースから抽出した文法データ（例えば、S->NP VP のような句構造文法に基づく文法規則）や共起データ等のデータを適宜利用して処理することで、名詞句や動詞句を特定したり、句と句の間に文法関係があるものを特定したりすることで、文の統語構造を取得する。斯かる構文解析処理には、一般に知られている適宜の構文解析プログラムを利用することができるが、本実施形態では、一例として、各語彙間の文法関係を解析できるパーサ（Parser）である RASP（Robust Accurate Statistical Parsing）を利用するものとする。図 4 に、コーパス DB 1 から抽出された「keen」と「eager」をそれぞれ含む文例（限

40

50

定的用法、叙述用法（主格補語）の各々1つずつ）と、これら文例のRASPによる出力例を示す。RASPによる出力では、形態素解析結果（上段括弧内）と、その文の統語構造（下段太字部分）とが得られる。形態素解析結果では、単語（原形）が||で分けられており、各単語について、当該文中で出現した順番（数字）と品詞データが付与されている。文の統語構造は、文法関係が|ncmod|（限定用法、修飾・被修飾関係）や|ncsubj|-|xcomp|（叙述用法、主語・述語関係）として表され、それらの右側に当該関係にある単語が並べられる（当然ながら、一方はターゲット単語である）。

【0033】

特に本実施形態では、RASPによる構文解析処理のルールとして、次のような設定を行うものとする。すなわち、

『|ncmod|については、RASPにより|ncmod|として抽出されるもののうち、形容詞と名詞の関係のみ、|ncmod|の関係として抽出した。ここで、形容詞は品詞がJJで始まる単語とし、名詞は品詞がNN、PN、VVG、PPのいずれかで始まる単語とする。また、品詞がNPで始まる単語は複合語の一部であることが多いため、品詞NPで始まる単語が、形容詞と|ncmod|の関係にあり、かつ、名詞とも|ncmod|の関係にある場合には、その品詞NPで始まる単語を介して形容詞と名詞の間にも|ncmod|の関係があるものとし、その形容詞と名詞を|ncmod|の関係にあるものとして抽出する。形容詞と名詞が前置詞で関係付けられている場合は、前置詞が「of」の場合のみ|ncmod|の関係として抽出する。|ncsubj|-|xcomp|については、『名詞と|ncsubj|の関係にある単語が、形容詞と|xcomp|の関係にある場合、その名詞と形容詞を|ncsubj|-|xcomp|の関係にあるものとして抽出する。ただし、このとき名詞と形容詞を仲介している単語は、次の単語に限定する。

< be, find, become, make, seem, appear, feel,
look, sound, smell, taste, remain, keep, stay, come, end-up, get, go, grow,
prove, turn, turn out, wind-up, burn, lie, loom, play, plead, rest, stand,
stand-up, blush, fall, fall-down, freeze, run, slam, spring, wax >

ただし、形容詞と名詞が前置詞で関係付けられている場合は抽出しない。また、形容詞の右が「that」の場合(it is [形容詞] that ...の場合)も抽出しない。』。なお、以上の他の文法関係の例には、|dobj|-|xcomp|（叙述用法、動詞・目的格関係）を挙げることができる。その場合のRASPにおける構文解析処理のルールとしては、次のようなものとする。すなわち、『名詞と|dobj|の関係にある単語が、形容詞と|xcomp|の関係にある場合、その名詞と形容詞を|dobj|-|xcomp|の関係にあるものとして抽出する。品詞NPで始まる単語が、ある単語と|dobj|の関係にあり、かつ、名詞とも|ncmod|の関係にある場合には、その品詞NPで始まる単語を|ncmod|の関係にある名詞と置き換えたのち、上記の|dobj|-|xcomp|の関係を抽出する。ただし、このとき名詞と形容詞を仲介している単語は、次の単語に限定する。

< hold, keep, leave, call, confess, profess,
pronounce, report, like, prefer, want, wish, believe, consider, deem, find,
hold, imagine, judge, presume, rate, reckon, suppose, think, drive, get, make,
prove, render, send, turn, certify, declare, proclaim >

ただし、形容詞と名詞が前置詞で関係付けられている場合は抽出しない。また、形容詞の右がthatの場合(it is JJ that ...の場合)も抽出しない。』

【0034】

次に、有向グラフ生成手段4において検索するシソーラスDB2について簡単に説明する。本実施形態で利用するシソーラスDB2である「WordNet 2.0」は、ネットワーク構造をもつ有向グラフ状のデータ構造を有しており、この有向グラフの各ノードは、単語あるいは意味カテゴリ（概念）を表す。ノード間に意味的な上位下位の関係がある場合には、そのノード間にリンクが張られている。全てのノードは少なくとも一つの他のノードとの間にリンクが張られており、最上位のノードは「Root」、最下位のノードは単語である（図6参照、最下位のノードには@が付してある）。

【0035】

10

20

30

40

50

さて、有向グラフ生成手段4による処理工程S4では、具体的には例えば図5に示すような手順で処理が実行される。まず、シソーラスDB2にアクセスして、上述した名詞抽出手段3で抽出されたターゲット単語と対をなす各名詞データで検索し、当該名詞がネットワーク（有向グラフ）上でどのノードの位置に現れるかを特定する（S41）。そして、特定したノードとROOTノード（最上位）とを結ぶ最短のパス（リンク及びノードを介した経路）を抽出する（S42）。次に、ターゲット単語毎に、抽出された全名詞データについて前工程で抽出したパスを重ね合わせることで、有向グラフを生成する（S43）。この工程で生成された有向グラフは、元のシソーラスDB2の有向グラフの一部を切り取った形になっており、これがターゲット単語の意味拡張を表していると考え、パスを重ね合わせたときの各ノードの重なりが多いほどそのノードの持つ意味カテゴリがコーパスDB1に高い頻度で出現していると考えられる。そこで、次の工程として、各ノードの重なり数をROOTノードでの重なり数で割ることによって、各ノードの頻度比率（出現比率）を算出する（正規化）（S44）。ROOTノードの頻度比率は1であり、各単語のノードの頻度比率は、コーパスDB1での当該単語の出現比率を同じターゲット単語と共に抽出された全名詞の出現頻度で割り算した値となる。以上の処理を経て、各ノードに頻度比率が付与された有向グラフ（以下、「重み付き有向グラフ」と呼ぶ）が生成される（S45）。すなわち、この重み付き有向グラフは、ターゲット単語の意味拡張をコーパスDB1での共起した名詞の頻度で重み付けされたものである。図6に、ターゲット単語「keen」及び「eager」について生成された重み付き有向グラフの例を示す。

10

20

【0036】

次に、差異抽出手段5による処理工程S5では、具体的には例えば図7に示すような手順で処理が実行される。まず、工程S45で生成された対比されるべき2つの重み付き有向グラフから、同じノードを抽出し、両重み付き有向グラフを重ね合わせる（S51）。この処理により、両重み付き有向グラフ間で共通するノードと異なるノードが大まかに分けられる。図8に、この重ね合わせ処理のイメージ図を示す。同図中、印がノードを示し、2つのノード間を結ぶ直線がリンクを示している。ここで、共通するノードのうち有向グラフ上で意味的に最下位にあるノード（ボトムノード）と、差異のノードのうち有向グラフ上で意味的に最上位にあるノード（トップノード）の境界が、2つのターゲット単語の意味拡張の共通部分と差異部分の境目であり、二つの語彙の意味的な違いを顕著に表している部分であるといえる。2つのターゲット単語の用法の違いを粗く分けるときには、前記トップノード又はそれ以下を、差異ノードに決定して差異抽出手段5による処理工程S5を終了することもできるが、共通するノードであっても頻度比率の違いが著しい場合があり、特にその違いが顕著な場合は差異ノードとして扱う方が好ましいといえる場合がある。

30

【0037】

そこで、次の工程として、それら同じノードについて頻度比率の比Cを計算する（S52）。これら同じノードは通常、それぞれの重み付き有向グラフにおいて相互に異なる頻度比率を有している。ただし、この比Cの計算では、その値が1以下となるように、頻度比率の値が大きい方を分母とする。次に、算出された比Cを所定の閾値Cxと比較することで、その比Cの算出に用いられたノードを、暫定的な差異ノードと共通ノードの何れかに分ける。ここで閾値Cxは、初期値において1以下の適宜の正数を採用することができるが、ボトムノード近辺における頻度比率の比の値を採用することが好ましい。具体的な処理としては、比Cが閾値Cx以上の場合は、それらのノードを暫定的な差異ノード、それ以外の場合（比Cが閾値Cx未満）は当該ノードを暫定的な共通ノードとする（S53）。この処理で暫定的な差異ノードとしたノードについては、工程S51で元々異なるノードであった部分に組み入れて、それら差異ノードからなる有向グラフからトップノードを全て抽出する（S54）。そして、各ターゲット単語について、それらトップノードの集合を頻度比率の高いものから順にソートし（S55）、上位所定数（N個）のトップノードを抽出する（S56）。この抽出数Nは任意の値であるが、通常は10から100の

40

50

間の適宜の値を採用すればよく、好ましくは20前後の値で十分であり、本実施形態では $N = 20$ とする。次に、これら N 個のノードのうち、両ターゲット単語の有向グラフにおいて共通するノードの値 P_c^n が所定値 C_p 以上か否かを判定する (S57)。この C_p の値には、適宜の値を採用することができるが、好ましくは0.3から0.7の間の値とするのがよく、さらに好ましくは0.5前後であり、本実施形態では $C_p = 0.5$ とする。ここで、共通するノードの値 P_c^n が所定値 C_p 未満であれば、一旦、比 C の値が0か否かを判断し (S57a)、 $C = 0$ でなければ (S57; No)、閾値 C_x の値を一定値下げ、工程S53に戻る (S57b)。ここで、工程S57bにおける閾値 C_x の変化値は、適宜の値 (例えば0.1) を採用することができる。他方、 $C = 0$ であれば (S57; Yes)、共通するノードの値 P_c^n が初めて C_p 以上となる比 C の値を採用して (S57c)、本工程S5の処理を終了する。また、共通するノードの値 P_c^n が所定値 C_p 以上の場合は (S57; Yes)、共通するノードの値 P_c^n と前回の工程S53から工程S57の処理サイクルにおいて得られた共通するノードの値 P_c^{n-1} とを比較し、その比較した値が一定値 Q 以上か否か、すなわちその比較した値が急激に変化したか否かを判定する (S58) (但し、1回目の処理サイクルは除く)。この比較処理は、両方の値の差 (差の絶対値; $|P_c^n - P_c^{n-1}|$) を採用してもよいし、両方の値の比 (P_c^n / P_c^{n-1}) を採用してもよい。本実施形態では差の絶対値を採用しており、判断基準となる値 Q を $4 / 20$ (0.2) としている。但しこの値 Q は任意であり、例えば機械学習により適宜最適化を図ることができる。そして、この比較処理の結果の値が一定値 Q 未満であれば (S58; No)、工程S57に戻るが、一定値 Q 以上であれば (S58; Yes)、共通するノードであっても頻度比率の違いが著しいと考えることができるので、工程S53で暫定的な差異ノードとしたノードを差異ノードに決定し、他方、工程S53暫定的な共通ノードとしたノードを共通ノードに決定する (S59)。以上を以て、差異抽出手段5による処理工程S5を終了する。

10

20

30

40

【0038】

以上のようにして、対比されるべき2つの重み付き有向グラフで共通ノードと差異ノードに分けられることになる。なお、次の差異出力手段6による処理に備えて、比較されるターゲット単語の用法の違いに関する情報として、工程S59で決定した差異ノードのトップノードのみを全て又は一部 (頻度比率が上位の一部又はランダムに一部) 抽出しておいたり、このトップノードを含む全差異ノードを抽出しておいたり、トップノードを含む上位所定数 (又はランダムに所定数) の差異ノードを抽出しておいたり、トップノードとそれから有向グラフ上を辿ることができる最下位の差異ノード (名詞を示すノード) を抽出しておくことができる。また、差異ノードのトップノードを決定することで、その直上位のノードを共通ノードの最下位ノード (ボトムノード) として抽出することができるので、これを用法の違いに関する情報 (当該情報の裏返しの意味として、比較されるターゲット単語の意味の共通する最下限概念) 又はその一部として利用することができる。本実施形態では、ターゲット単語の用法の違いに関する情報として、頻度比率が高い方から所定数の差異ノードのトップノードと、各トップノードから辿れる最下位の単語ノードの所定数、またそれらに加えて共通ノードのボトムノードの頻度比率が低い方から所定数とその直上位のノードを抽出しておき、差異出力手段6により、これらのノードを整理して出力するものとする。

【0039】

図9にターゲット単語が「keen」と「eager」の場合における画面又は印刷出力例を示す。同図上欄は「keen」と「eager」の共通ノードを示し、下欄は差異ノードを示す。上欄中、右欄は共通ノードのボトムノード (単語) であり、左欄はその直上位概念のノードである。下欄中、左欄は「keen」と「eager」の別を示し、それらのそれぞれについて中欄は差異ノードのトップノード、右欄は各トップノードの下位に当たる単語ノードである。このような出力例からユーザは、「keen」については、競争を前提とした活動 (sportsman, player) や専門的な職業に従事する主体 (gardener) や熱狂的で欲求の程度が極端に強い主体 (supporter, fan) 等が出現するような意味の

50

文で用いられる傾向が強く、本来の意味である「痛み、刃の鋭さ、獯猛さ」といったものに含まれる「はなはだしさやネガティブ性」を温存しているのに対して、「eager」については、好んである行動をしたがるという自然でポジティブな主体（audience, buyer, volunteer）等が出現するような意味の文で用いられる傾向が強く、刺激の甚だしさよりも味覚等の感覚が持つ「自然な欲求」を温存している、という2つの単語の用法の違いが分かる。

【0040】

なお、本発明は上述した実施形態に限定されるものではない。例えば、入力されるターゲット単語は3つ以上とすることが可能である。その場合、例えば各ターゲット単語について有向グラフを生成しておき、1つ以外の有向グラフを全て重ね合わせて有向グラフを生成し、それと残りの有向グラフを比較することで、あるターゲット単語と他のターゲット単語の用法の違いに関する情報を得ることができる。また、全てのターゲット単語の有向グラフを全て重ね合わせることで、各ターゲット単語の用法の違いの概略的な情報を得ることができる。その他、有向グラフに頻度による重み付けを行わない態様や、頻度比率の変わりに頻度の値そのものを利用する態様などの種々の変更が可能である。また、本発明のプログラムや装置を、他の文書入力プログラム、言語学習プログラム、翻訳プログラム等の各種プログラムやそれらに従って作動する装置の一部に応用することも可能である。さらに、各部の具体的構成や処理工程についても上記実施形態に限られるものではなく、本発明の趣旨を逸脱しない範囲で種々変形が可能である。

10

【図面の簡単な説明】

20

【0041】

【図1】本発明の一実施形態に係る単語用法差異情報取得装置の概略的な機能構成図

【図2】同装置の単語用法差異情報取得プログラムに基づく概略的な全処理工程図

【図3】図2の処理工程の一部（S3）を詳細に示す処理工程図

【図4】同装置に入力されるターゲット単語に基づいてコーパスから抽出される文例とその構文解析例を示す図

【図5】図2の処理工程の一部（S4）を詳細に示す処理工程図

【図6】図2の処理工程S4により生成される重み付き有向グラフ例を示す図

【図7】図2の処理工程の一部（S5）を詳細に示す処理工程図

【図8】図7の処理工程S5においてなされる有向グラフの重ね合わせ処理を示すイメージ図

30

【図9】図2の処理工程S6による出力例を示す図

【符号の説明】

【0042】

A ... 単語用法差異情報取得装置

DB1 ... コーパス

DB2 ... シソーラス

1 ... ターゲット単語入力手段

2 ... 文抽出手段

3 ... 名詞抽出手段

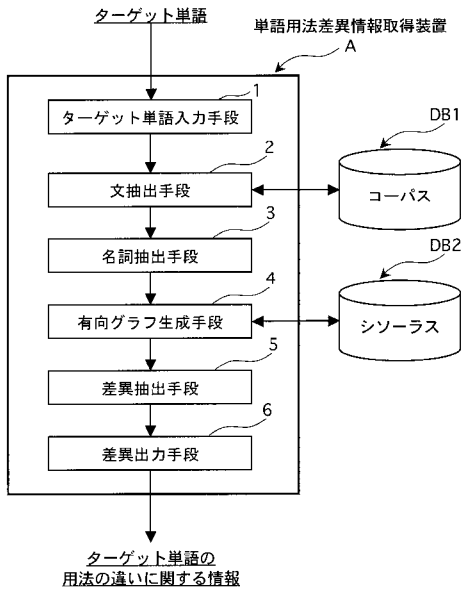
4 ... 有向グラフ生成手段

5 ... 差異抽出手段

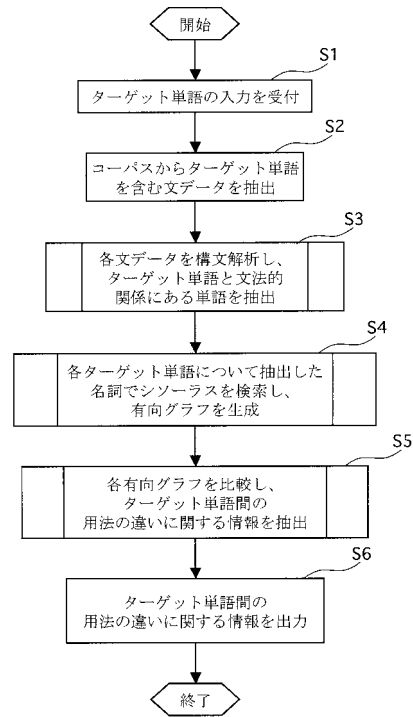
6 ... 差異出力手段

40

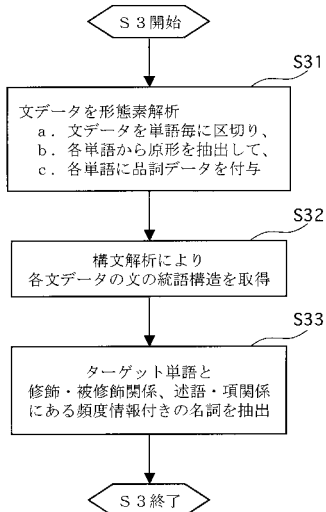
【 図 1 】



【 図 2 】



【 図 3 】



【 図 4 】



【 図 9 】

[keenとeagerの差異ノードと共通ノードの出力例]

Common nodes and nouns included in each node		
PEOPLE(keen: 2.6%, eager 3.5%)	people(keen: 28, eager: 9)	
INSTITUTION(keen: 1.8%, eager 1.8%)	company(keen: 14, eager: 5)	
Different nodes and nouns included in each node		
K E E N	CONTESTANT (4.1%)	sportsman(21), player(17), golfer(13), rival(4) etc.
	EXPERT (3.4%)	gardener(33), observer (6) etc.
	ADVOCATE (2.8%)	supporter(27), fan(8), enthusiast(3), admirer(2) etc.
	PEER, COMPEER (1.8%)	member(16), colleague(2), participant(2) etc.
	REGION (1.4%)	country(5), germany(2), russia(2) etc.
E A G E R	GATHERING (3.5%)	audience(8), crowd(2) etc.
	USER (2.5%)	buyer(6), customer(5), consumer(2) etc.
	ANIMAL, BEAST (2.4%)	dog(4), horse(2), elephant(1) etc.
	SERVICEMAN (2.4%)	volunteer(4), man(3) etc.
ENROLEE (2.4%)	student(10), pupil(3), undergraduate(1) etc.	

フロントページの続き

Fターム(参考) 5B091 AA15 AB17 CA02 CA05 CA12 CC05 CC16