

(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2011-34262
(P2011-34262A)

(43) 公開日 平成23年2月17日(2011.2.17)

(51) Int.Cl.	F 1	テーマコード (参考)
G06F 17/30 (2006.01)	G06F 17/30 230Z	5B075
	G06F 17/30 170J	
	G06F 17/30 419A	
	G06F 17/30 210D	

審査請求 未請求 請求項の数 10 O L (全 38 頁)

(21) 出願番号 特願2009-178673 (P2009-178673)
(22) 出願日 平成21年7月31日 (2009.7.31)

特許法第30条第1項適用申請有り 平成21年3月2日 言語処理学会発行の「言語処理学会第15回年次大会発表論文集」に発表

(特許庁注：以下のものは登録商標)

1. QRコード

(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

(74) 代理人 100115749
弁理士 谷川 英和

(72) 発明者 山田 一郎
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 鳥澤 健太郎
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 風間 淳一
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

最終頁に続く

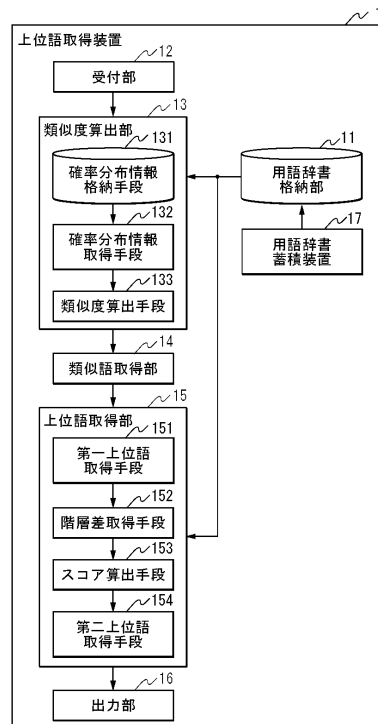
(54) 【発明の名称】 データ作成装置、上位語取得装置、データ作成方法、およびプログラム

(57) 【要約】

【課題】従来、大規模な上位語と下位語のデータベースを構築できなかった。

【解決手段】2以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納し得る用語辞書格納部と、上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付部と、用語対情報受付部が受け付けた用語対情報が有する上位語と、用語辞書格納部に格納されている1以上の各用語との類似度を算出する類似度算出部と、類似度算出部が算出した上位k(kは1以上の整数)の類似度に対応するk個の類似用語を取得する類似語取得部と、受け付けた用語対情報が有する下位語の上位語として、k個の類似用語を選択して、用語辞書格納部に下位語を蓄積する下位語付加部とを具備するデータ作成装置により、大規模な上位語と下位語のデータベースを構築できる。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

2 以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納し得る用語辞書格納部と、
上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付部と、
前記用語対情報受付部が受け付けた用語対情報が有する上位語と、前記用語辞書格納部に格納されている 1 以上の各用語との類似度を算出する類似度算出部と、
前記類似度算出部が算出した上位 k (k は 1 以上の整数) の類似度に対応する k 個の類似用語を取得する類似語取得部と、
前記受け付けた用語対情報が有する下位語の上位語として、前記 k 個の類似用語を選択して、前記用語辞書格納部に前記下位語を蓄積する下位語付加部とを具備するデータ作成装置。

10

【請求項 2】

前記下位語付加部は、
前記受け付けた用語対情報が有する下位語の上位語として、前記用語対情報が有する上位語と同一の文字列を有する類似用語を選択して、前記用語辞書格納部に前記下位語を蓄積する請求項 1 記載のデータ作成装置。

【請求項 3】

受け付けられた用語を 2 以上の文字列に分割し、最後尾の文字列を含む 1 以上の文字列を有する上位語を取得する上位語生成部をさらに具備し、
前記用語対情報受付部は、
前記上位語生成部が取得した上位語と、前記受け付けられた用語である下位語との対の情報である用語対情報を受け付ける請求項 1 または請求項 2 記載のデータ作成装置。

20

【請求項 4】

前記類似度算出部は、
前記用語対情報受付部が受け付けた用語対情報が有する下位語と、前記用語辞書格納部に格納されている 1 以上の各用語との類似度も算出し、
前記類似語取得部は、
前記上位語との類似度、および前記下位語との類似度の両方の類似度を用いて、 k 個の類似用語を取得する請求項 1 から請求項 3 いずれか記載のデータ作成装置。

30

【請求項 5】

請求項 1 から請求項 4 いずれか記載のデータ作成装置が構築した用語辞書格納部を用いる上位語取得装置であり、
前記上位語取得装置は、
前記用語辞書格納部と、
上位概念の用語を取得する対象となる用語である対象語を受け付ける受付部と、
前記受付部が受け付けた対象語と、前記用語辞書格納部に格納されている 1 以上の各用語との類似度を算出する類似度算出部と、
前記類似度算出部が算出した上位 k (k は 1 以上の整数) の類似度に対応する k 個の下位語を取得する類似語取得部と、
前記類似語取得部が取得した k 個の各下位語に対応する上位語を取得し、前記類似語取得部が取得した類似度をパラメータとする増加関数であるスコア算出の演算式に、前記類似度を代入し、対象語の上位語としての相応しさを示すスコアを、前記上位語ごとに算出し、少なくとも、前記スコアが最も高い上位語を取得する上位語取得部と、
前記上位語取得部が取得した上位語を出力する出力部とを具備する請求項 1 から請求項 4 いずれか記載の上位語取得装置。

40

【請求項 6】

前記上位語取得部は、
前記類似語取得部が取得した k 個の各下位語に対応する上位語を取得する第一上位語取得手段と、

50

前記第一上位語取得手段が取得した各上位語に対して、当該上位語と当該上位語に対応する下位語との階層差を取得する階層差取得手段と、
 前記類似語取得部が取得した類似度をパラメータとする増加関数であり、前記階層差取得手段が取得した階層差をパラメータとする減少関数であるスコア算出の演算式に、前記類似度と前記階層差とを代入し、対象語の上位語としての相応しさを示すスコアを、前記上位語ごとに算出するスコア算出手段と、
 少なくとも、前記スコア算出手段が算出したスコアが最も高い上位語を取得する第二上位語取得手段とを具備する請求項 5 記載の上位語取得装置。

【請求項 7】

前記類似度算出部は、
 用語と、動詞と助詞とを有する文字列を 1 以上有する 1 以上の各クラスに属する確率とを対応付けた情報である確率分布情報を、用語ごとに格納し得る確率分布情報格納手段と、
 前記受付部が受け付けた対象語の確率分布情報、および前記用語辞書格納部に格納されている 1 以上の各用語の確率分布情報を取得する確率分布情報取得手段と、
 前記対象語の確率分布情報、および前記各用語の確率分布情報を用いて、前記対象語と前記各用語の類似度を算出する類似度算出手段とを具備する請求項 5 または請求項 6 記載の上位語取得装置。

10

【請求項 8】

前記用語辞書を、前記用語辞書格納部に蓄積する用語辞書蓄積装置をさらに具備する請求項 5 から請求項 7 いずれか記載の上位語取得装置であって、
 前記用語辞書蓄積装置は、
 上位語を抽出するための定義文のパターンを示す情報である定義文パターン情報を、1 以上格納している定義文パターン情報格納部と、
 用語を説明する文章群であり、用語ごとに、定義文と、カテゴリと、用語の階層関係を特定する情報である階層関係定義情報と上位語と下位語とを有する用語説明文章群から、前記 1 以上の定義文パターン情報のうちのいずれか一の定義文パターン情報を適用して、前記対象語を有する対象語の定義文を取得する定義文取得部と、
 前記定義文取得部が取得した定義文から、前記適用された一の定義文パターン情報に従って、前記対象語の上位語の候補である第一の上位語候補と前記対象語の対である第一用語対候補を取得する第一用語対候補取得部と、
 前記用語説明文章群から、前記対象語のカテゴリを前記対象語の第二の上位語候補として、前記第二の上位語候補と前記対象語の対である第二用語対候補を取得する第二用語対候補取得部と、
 階層関係定義情報を 1 以上格納し得る階層関係定義情報格納部と、
 前記階層関係定義情報を用いて、上位語と下位語との対である 1 以上の第三用語対候補を取得する第三用語対候補取得部と、
 前記第一用語対候補を有する文または文の一部と、前記第二用語対候補を有する文または文の一部と、前記第三用語対候補を有する文または文の一部とから、言語処理した結果である 1 以上の素性を取得し、前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれの素性ベクトルを構成する素性ベクトル構成部と、
 前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれについて、対応する素性ベクトルを、サポートベクターマシンを用いて、前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれが、上位語と下位語の関係にあるか否かを判断する機械学習部と、
 前記機械学習部が、上位語と下位語の関係にあると判断した用語対候補が有する上位語および下位語を、前記用語辞書格納部に蓄積する用語対蓄積部とを具備する請求項 5 から請求項 7 いずれか記載の上位語取得装置。

20

30

40

【請求項 9】

記憶媒体に、
 2 以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納

50

しており、

用語対情報受付部、類似度算出部、類似語取得部、および下位語付加部とにより実現されるデータ作成方法であって、

前記用語対情報受付部により、上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付ステップと、

前記類似度算出部により、前記用語対情報受付ステップで受け付けられた用語対情報が有する上位語と、前記記憶媒体に格納されている 1 以上の各用語との類似度を算出する類似度算出ステップと、

前記類似語取得部により、前記類似度算出ステップで算出された上位 k (k は 1 以上の整数) の類似度に対応する k 個の類似用語を取得する類似語取得ステップと、

前記下位語付加部により、前記受け付けた用語対情報が有する下位語の上位語として、前記 k 個の類似用語を選択して、前記用語辞書格納部に前記下位語を蓄積する下位語付加ステップとを具備するデータ作成方法。

【請求項 10】

記憶媒体に、

2 以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納しており、

コンピュータを、

上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付部と、

前記用語対情報受付部が受け付けた用語対情報が有する上位語と、前記記憶媒体に格納されている 1 以上の各用語との類似度を算出する類似度算出部と、

前記類似度算出部が算出した上位 k (k は 1 以上の整数) の類似度に対応する k 個の類似用語を取得する類似語取得部と、

前記受け付けた用語対情報が有する下位語の上位語として、前記 k 個の類似用語を選択して、前記用語辞書格納部に前記下位語を蓄積する下位語付加部として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、大規模な上位語と下位語のデータベースを作成するデータ作成装置等に関するものである。

【背景技術】

【0002】

従来、文字列のパターンを用いて、ある用語の上位語を取得する技術があった（非特許文献 1 から非特許文献 5 など参照）。ここで、文字列のパターンとは、「<下位語>のようなく上位語」などである。そして、これらのパターンを種として、半自動的に、または自動的に新しいパターンを取得する技術も存在する（例えば、非特許文献 1、非特許文献 2 等を参照）。これらの文字列のパターンを用いる方法は、対象語と上位語との共起を必要とする。

【0003】

また、従来、文字列のパターンを用いる方法以外の方法として、クラスタリングベースの方法がある。この方法は、用語間の類似度または HTML ドキュメントの階層関係を用いて自動的に構築された用語クラスのための共通上位語を取得する（例えば、非特許文献 6、非特許文献 7、非特許文献 8 など参照）。

【0004】

さらに、文字列のパターンの方法とクラスタリングベースの方法との両方を用いて、上位語を取得する技術があった（非特許文献 9 など参照）。

【先行技術文献】

【非特許文献】

【0005】

10

20

30

40

50

【非特許文献1】M.Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th Conference on Computational Linguistics (COLING), pp. 539-545.

【非特許文献2】P.Pantel, D.Ravichandran and E.Hovy. 2004a. Towards Terascale Knowledge Acquisition. In Proceedings of the 20th International Conference on Computational Linguistics.

【非特許文献3】R. Snow, D. Jurafsky and A. Y. Ng. 2005. Learning Syntactic Patterns for Automatic Hypernym Discovery. NIPS 2005.

【非特許文献4】R.Snow, D.Jurafsky, A.Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 801-808.

【非特許文献5】M.Ando, S.Sekine and S.Ishizaki. 2003. Automatic Extraction of Hyponyms from Newspaper Using Lexicosyntactic Patterns. IPSJ SIG Notes, 2003-NL-157, pp. 77-82 (in Japanese).

【非特許文献6】S. A. Caraballo. 1999. Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text. In Proceedings of the Conference of the Association for Computational Linguistics (ACL).

【非特許文献7】P. Pantel and D. Ravichandran. 2004b. Automatically Labeling Semantic Classes. In Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference.

【非特許文献8】K. Shinzato and K. Torisawa. 2004. Acquiring Hyponymy Relations from Web Documents. In Proceedings of HLT-NAACL, pp. 73-80.

【非特許文献9】O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld and A. Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence, 165(1):91-134.

【発明の概要】

【発明が解決しようとする課題】

【0006】

しかしながら、従来の文字列のパターンを用いる方法では、大規模な上位語と下位語のデータベースを構築できなかったという課題があった。

【0007】

つまり、文字列のパターンを用いる方法では、同一文における対象語と上位語との共起が必要であるため、大規模な上位語と下位語のデータベースを構築できない。

【0008】

また、クラスタリングベースの方法では、特定のクラスに属する対象語群やリスト構造にある対象語群に対して、同一の上位語が与えられる。このことは、クラスの粒度やリスト構造が不適切である場合、適切な上位語が抽出できない、という課題が生じる。

【0009】

さらに、文字列のパターンの方法とクラスタリングベースの方法との両方を用いる方法でも、大規模な上位語と下位語のデータベースを構築できない、または適切な上位語が抽出できない、という課題が生じ得る。

【課題を解決するための手段】

【0010】

本第一の発明のデータ作成装置は、2以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納し得る用語辞書格納部と、上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付部と、用語対情報受付部が受け付けた用語対情報が有する上位語と、用語辞書格納部に格納されている1以上の各用語との類似度を算出する類似度算出部と、類似度算出部が算出した上位 k (k は1以上の整数)の類似度に対応する k 個の類似用語を取得する類似語取得部と、受け付けた用語対情報が有す

10

20

30

40

50

る下位語の上位語として、 k 個の類似用語を選択して、用語辞書格納部に下位語を蓄積する下位語付加部とを具備するデータ作成装置である。

【0011】

かかる構成により、大規模な上位語と下位語のデータベースを構築できる。

【0012】

また、本第二の発明のデータ作成装置は、第一の発明に対して、下位語付加部は、受け付けた用語対情報が有する下位語の上位語として、用語対情報が有する上位語と同一の文字列を有する類似用語を選択して、用語辞書格納部に下位語を蓄積するデータ作成装置である。

【0013】

かかる構成により、大規模な上位語と下位語のデータベースを構築できる。

【0014】

また、本第三の発明のデータ作成装置は、第一または第二の発明に対して、受け付けられた用語を2以上の文字列に分割し、最後尾の文字列を含む1以上の文字列を有する上位語を取得する上位語生成部をさらに具備し、用語対情報受付部は、上位語生成部が取得した上位語と、受け付けられた用語である下位語との対の情報である用語対情報を受け付けるデータ作成装置である。

【0015】

かかる構成により、大規模な上位語と下位語のデータベースを構築できる。

【0016】

また、本第四の発明のデータ作成装置は、第一から第三いずれかの発明に対して、類似度算出部は、用語対情報受付部が受け付けた用語対情報が有する下位語と、用語辞書格納部に格納されている1以上の各用語との類似度をも算出し、類似語取得部は、上位語との類似度、および下位語との類似度の両方の類似度を用いて、 k 個の類似用語を取得するデータ作成装置である。

【0017】

かかる構成により、大規模な上位語と下位語のデータベースを構築できる。

【0018】

また、本第五の発明の上位語取得装置は、第一から第四いずれかの発明のデータ作成装置が構築した用語辞書格納部を用いる上位語取得装置であり、前記上位語取得装置は、前記用語辞書格納部と、上位概念の用語を取得する対象となる用語である対象語を受け付ける受付部と、前記受付部が受け付けた対象語と、前記用語辞書格納部に格納されている1以上の各用語との類似度を算出する類似度算出部と、前記類似度算出部が算出した上位 k (k は1以上の整数) の類似度に対応する k 個の下位語を取得する類似語取得部と、前記類似語取得部が取得した k 個の各下位語に対応する上位語を取得し、前記類似語取得部が取得した類似度をパラメータとする増加関数であるスコア算出の演算式に、前記類似度を代入し、対象語の上位語としての相応しさを示すスコアを、前記上位語ごとに算出し、少なくとも、前記スコアが最も高い上位語を取得する上位語取得部と、前記上位語取得部が取得した上位語を出力する出力部とを具備する請求項1から請求項4いずれか記載の上位語取得装置である。

【0019】

かかる構成により、大規模な上位語と下位語のデータベースを構築できる。

【0020】

また、本第六の発明の上位語取得装置は、第五の発明に対して、上位語取得部は、類似語取得部が取得した k 個の各下位語に対応する上位語を取得する第一上位語取得手段と、第一上位語取得手段が取得した各上位語に対して、上位語と上位語に対応する下位語との階層差を取得する階層差取得手段と、類似語取得部が取得した類似度をパラメータとする増加関数であり、階層差取得手段が取得した階層差をパラメータとする減少関数であるスコア算出の演算式に、類似度と階層差とを代入し、対象語の上位語としての相応しさを示すスコアを、上位語ごとに算出するスコア算出手段と、少なくとも、スコア算出手段が算

10

20

30

40

50

出したスコアが最も高い上位語を取得する第二上位語取得手段とを具備する上位語取得装置である。

【0021】

かかる構成により、大規模な上位語と下位語のデータベースを、精度高く構築できる。

【0022】

また、本第七の発明の上位語取得装置は、第五または第六の発明に対して、類似度算出部は、用語と、動詞と助詞とを有する文字列を1以上有する1以上の各クラスに属する確率とを対応付けた情報である確率分布情報を、用語ごとに格納し得る確率分布情報格納手段と、受付部が受け付けた対象語の確率分布情報、および用語辞書格納部に格納されている1以上の各用語の確率分布情報を取得する確率分布情報取得手段と、対象語の確率分布情報、および各用語の確率分布情報を用いて、対象語と各用語の類似度を算出する類似度算出手段とを具備する上位語取得装置である。

10

【0023】

かかる構成により、大規模な上位語と下位語のデータベースを、さらに精度高く構築できる。

【0024】

また、本第八の発明の上位語取得装置は、第五から第七いずれかの発明に対して、用語辞書を、用語辞書格納部に蓄積する用語辞書蓄積装置をさらに具備するデータ作成装置であって、用語辞書蓄積装置は、上位語を抽出するための定義文のパターンを示す情報である定義文パターン情報を、1以上格納している定義文パターン情報格納部と、用語を説明する文章群であり、用語ごとに、定義文と、カテゴリと、用語の階層関係を特定する情報である階層関係定義情報と上位語と下位語とを有する用語説明文章群から、1以上の定義文パターン情報のうちのいずれか一の定義文パターン情報を適用して、対象語を有する対象語の定義文を取得する定義文取得部と、定義文取得部が取得した定義文から、適用された一の定義文パターン情報に従って、対象語の上位語の候補である第一の上位語候補と対象語の対である第一用語対候補を取得する第一用語対候補取得部と、用語説明文章群から、対象語のカテゴリを対象語の第二の上位語候補として、第二の上位語候補と対象語の対である第二用語対候補を取得する第二用語対候補取得部と、階層関係定義情報を1以上格納し得る階層関係定義情報格納部と、階層関係定義情報を用いて、上位語と下位語との対である1以上の第三用語対候補を取得する第三用語対候補取得部と、第一用語対候補を有する文または文の一部と、第二用語対候補を有する文または文の一部と、第三用語対候補を有する文または文の一部とから、言語処理した結果である1以上の素性を取得し、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれの素性ベクトルを構成する素性ベクトル構成部と、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれについて、対応する素性ベクトルを、サポートベクターマシンを用いて、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれが、上位語と下位語の関係にあるか否かを判断する機械学習部と、機械学習部が、上位語と下位語の関係にあると判断した用語対候補が有する上位語および下位語を、用語辞書格納部に蓄積する用語対蓄積部とを具備する上位語取得装置である。

20

30

【0025】

かかる構成により、大規模な用語辞書を構築できる。

40

【発明の効果】

【0026】

本発明による上位語取得装置によれば、大規模な上位語と下位語のデータベースを構築できる。

【図面の簡単な説明】

【0027】

【図1】実施の形態1における上位語取得装置のブロック図

【図2】同上位語取得装置のブロック図

【図3】同上位語取得装置の動作について説明するフローチャート

50

【図 4】同類似度を算出するアルゴリズムの例を説明するフローチャート

【図 5】同用語辞書蓄積装置の動作について説明するフローチャート

【図 6】同用語辞書の例を示す図

【図 7】同確率分布管理表を示す図

【図 8】同対象語と類似する k 個の共通下位語を取得した場合の概念図

【図 9】同実験結果を示す図

【図 10】同取得できた上位語の例を示す図

【図 11】同用語説明文章群（ウィキペディア）の例を示す図

【図 12】同用語説明文章群（ウィキペディア）の元になるデータを示す図

【図 13】同階層関係定義情報管理表を示す図

10

【図 14】同抽出した階層構造を示す図

【図 15】同上位語候補から文字列を取り除くための除外パターンを示す図

【図 16】同素性の例を説明する図

【図 17】実施の形態 2 におけるデータ作成装置が構成する大規模な辞書の概念を示す図

【図 18】同データ作成装置のブロック図

【図 19】同データ作成装置の動作について説明するフローチャート

【図 20】同コンピュータシステムの概観図

【図 21】同コンピュータシステムのブロック図

【発明を実施するための形態】

【0028】

20

以下、上位語取得装置等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。

（実施の形態 1）

【0029】

本実施の形態において、2 階層以上に階層化された用語辞書に含まれる各用語と対象語との類似度を算出し、類似度が上位の k （ k は 1 以上の整数）個の下位語を取得し、当該類似度を用いて、前記 k 個の下位語から m （ m は 1 以上の整数）個の下位語を取得し、当該 m 個の下位語の上位語を用語辞書から取得し、出力する上位語取得装置について説明する。

30

【0030】

また、本実施の形態において、例えば、用語辞書には 3 階層以上に階層化された上位語と下位語とが格納され、 m 個の上位語を抽出する際に、 k 個の下位語に対するスコアを算出し、当該スコアを用いて m 個の上位語を取得する上位語取得装置について説明する。なお、上位語取得装置は、通常、上位語と下位語との階層差が大きいほど、スコアが小さくなる演算式を用いてスコアを算出し、スコアの高い上位語を出力する。

【0031】

また、本実施の形態において、動詞と助詞とを有する文字列を 1 以上有する 1 以上の各クラスに、用語が属する確率である確率分布情報を、用語ごとに格納しており、各用語が各クラスに属する確率分布を有し、当該確率分布を用いて、各用語と対象語との類似度を算出する上位語取得装置について説明する。

40

【0032】

さらに、本実施の形態において、階層化された用語辞書を自動構築する用語辞書蓄積装置について説明する。

【0033】

図 1 は、本実施の形態における上位語取得装置 1 のブロック図である。図 2 は、上位語取得装置 1 を構成する用語辞書蓄積装置 17 のブロック図である。なお、上位語取得装置 1 は、用語辞書蓄積装置 17 を有しなくても良い。かかる場合、用語辞書格納部 11 の用語辞書は、図示しない手段により、予め準備されている。

【0034】

50

上位語取得装置 1 は、用語辞書格納部 1 1、受付部 1 2、類似度算出部 1 3、類似語取得部 1 4、上位語取得部 1 5、出力部 1 6、用語辞書蓄積装置 1 7 を具備する。

【 0 0 3 5 】

類似度算出部 1 3 は、確率分布情報格納手段 1 3 1、確率分布情報取得手段 1 3 2、類似度算出手段 1 3 3 を具備する。

【 0 0 3 6 】

上位語取得部 1 5 は、第一上位語取得手段 1 5 1、階層差取得手段 1 5 2、スコア算出手段 1 5 3、第二上位語取得手段 1 5 4 を具備する。

【 0 0 3 7 】

用語辞書蓄積装置 1 7 は、定義文パターン情報格納部 1 7 1、定義文取得部 1 7 2、第一用語対候補取得部 1 7 3、第二用語対候補取得部 1 7 4、階層関係定義情報格納部 1 7 5、第三用語対候補取得部 1 7 6、素性ベクトル構成部 1 7 7、機械学習部 1 7 8、用語対蓄積部 1 7 9 を具備する。

10

【 0 0 3 8 】

用語辞書格納部 1 1 は、用語辞書を格納し得る。用語辞書とは、2 以上の用語を有し、かつ、2 以上の用語間の階層関係を管理している情報である。用語辞書は、例えば、上位語と下位語の対の情報の集合である。用語辞書格納部 1 1 は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。用語辞書格納部 1 1 の用語辞書は、後述する用語辞書蓄積装置 1 7 により蓄積されることは好適である。ただし、用語辞書が記憶される過程は問わない。例えば、記録媒体を介して用語辞書が用語辞書格納部 1 1 で記憶されるようになってよく、通信回線等を介して送信された用語辞書が用語辞書格納部 1 1 で記憶されるようになってよく、あるいは、入力デバイスを介して入力された用語辞書が用語辞書格納部 1 1 で記憶されるようになってよくよい。

20

【 0 0 3 9 】

受付部 1 2 は、対象語を受け付ける。対象語とは、上位概念の用語を取得する対象となる用語である。ここで、対象語の受け付け方法は問わない。対象語は、ユーザからの手入力により受け付けられても良いし、プログラムから渡されても良いし、記憶媒体から読み出されるなどしても良い。例えば、対象語は、Web ページから自動的に取得されても良い。例えば、図示しない手段が Web ページから名詞または名詞句を取得し、当該名詞または名詞句が用語辞書に存在するか存在しないかを判断し、存在しない場合に、当該名詞または名詞句を対象語としても良い。対象語の入力手段は、テンキーやキーボードやマウスやメニュー画面によるもの等、何でも良い。受付部 1 2 は、テンキーやキーボード等の入力手段のデバイスドライバや、メニュー画面の制御ソフトウェア等で実現され得る。

30

【 0 0 4 0 】

類似度算出部 1 3 は、受付部 1 2 が受け付けた対象語と、用語辞書格納部 1 1 に格納されている 1 以上の各用語との類似度を算出する。対象語と用語との類似度を算出するアルゴリズムは問わない。算出方法の例は後述する。類似度算出部 1 3 は、通常、MPU やメモリ等から実現され得る。類似度算出部 1 3 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

40

【 0 0 4 1 】

確率分布情報格納手段 1 3 1 は、用語ごとに、確率分布情報を格納し得る。確率分布情報とは、用語が、1 以上の各クラスに属する確率の分布（集合）の情報である。確率分布情報は、ベクトルを構成し得る。クラスとは、動詞と助詞との組を 1 以上有する情報群、または、動詞と助詞との組を抽象化したものを 1 以上有する情報群である。クラスとは、例えば、同じ名詞と共起しやすい動詞と助詞とを有する文字列の集合である。クラスは、適宜、隠れクラスという。ここで、用語は、用語を識別する情報でも良い。動詞と助詞とを有する文字列とは、動詞と助詞とが分離していても良い。分離するための文字は、スペース、コンマなど問わない。確率分布情報格納手段 1 3 1 は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。確率分布情報格納手段 1 3 1 に確率分

50

布情報が記憶される過程は問わない。

【0042】

確率分布情報取得手段132は、受付部12が受け付けた対象語の確率分布情報、および用語辞書格納部11に格納されている1以上の各用語の確率分布情報を取得する。確率分布情報取得手段132は、通常、MPUやメモリ等から実現され得る。確率分布情報取得手段132の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

【0043】

類似度算出手段133は、対象語の確率分布情報、および各用語の確率分布情報を用いて、対象語と各用語の類似度を算出する。類似度算出手段133は、例えば、対象語の確率分布情報であるベクトルと、用語の確率分布情報であるベクトルとの距離を類似度として算出する。つまり、類似度算出手段133は、同様の動詞および助詞と共起しやすい名詞や名詞句は、類似している可能性が高い、という特性を利用している。類似度算出手段133は、例えば、以下のように類似度を算出しても良い。例えば、単語と単語の間の類似度をあらかじめ人手で定めた辞書や規則を図示しない記憶媒体に格納しておき、当該辞書または規則を用いて、類似度算出手段133は、単語と単語の間の類似度を取得しても良い。類似度算出手段133は、通常、MPUやメモリ等から実現され得る。類似度算出手段133の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

10

【0044】

類似語取得部14は、類似度算出部13が算出した1以上の上位 k (k は1以上の整数)の類似度に対応する k 個の下位語を取得する。上位 k とは、類似度が上位から一定数の下位語でも良いし、類似度が上位から一定割合の下位語でも良いし、類似度が一定値以上のものでも良い。結果として、上位 k の類似度に対応する下位語を取得すれば良い。類似語取得部14は、通常、MPUやメモリ等から実現され得る。類似語取得部14の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

20

【0045】

上位語取得部15は、類似語取得部14が取得した k 個の各下位語に対応する上位語を取得し、上位語から m (m は1以上の整数)個の上位語を、用語辞書格納部11から取得する。また、上位語取得部15が、 k 個の上位語から m 個の上位語を選択するアルゴリズムは問わない。上位語取得部15は、後述する数式1を用いて上位語らしさを示すスコアを算出し、当該スコアを用いて上位語を選択することが好適である。また、上位語取得部15は、類似語取得部14が取得した k 個の各下位語に対応する上位語を取得し、類似語取得部14が取得した類似度をパラメータとする増加関数であるスコア算出の演算式に、取得した類似度を代入し、対象語の上位語としての相応しさを示すスコアを、上位語ごとに算出し、少なくとも、スコアが最も高い上位語を取得しても良い。ただし、上位語取得部15は、ランダムに m 個の上位語を選択しても良い。上位語取得部15は、通常、MPUやメモリ等から実現され得る。上位語取得部15の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

30

40

【0046】

第一上位語取得手段151は、類似語取得部14が取得した k 個の各下位語に対応する上位語を取得する。第一上位語取得手段151は、通常、MPUやメモリ等から実現され得る。第一上位語取得手段151の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

【0047】

階層差取得手段152は、第一上位語取得手段151が取得した各上位語に対して、上位語と上位語に対応する下位語との階層差を取得する。階層差取得手段152は、用語辞

50

書格納部 11 に格納されている用語辞書から、2つの用語の階層差を取得する。用語辞書は、2以上の用語間の階層関係を管理しているので、2つの用語の階層差を取得し得る。階層差取得手段 152 は、通常、用語とその親の用語との階層差を「1」として取得し、用語とその孫の用語との階層差を「2」として取得する。階層差取得手段 152 は、通常、MPU やメモリ等から実現され得る。階層差取得手段 152 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【0048】

スコア算出手段 153 は、対象語の上位語としての相応しさを示すスコアを、上位語ごとに算出する。具体的には、スコア算出手段 153 は、スコア算出の演算式に、類似度と階層差とを代入し、対象語の上位語としての相応しさを示すスコアを算出する。この演算式は、通常、類似語取得部 14 が取得した類似度が大きければ大きいほど、スコアが大きくなる関数である。つまり、演算式は、通常、類似度をパラメータとする増加関数である。また、この演算式は、通常、階層差取得手段 152 が取得した階層差が大きければ大きいほど、スコアが小さくなる関数である。つまり、演算式は、通常、階層差をパラメータとする減少関数である。さらに具体的には、スコア算出手段 153 は、例えば、以下の数式 1 により、上位語ごとのスコアを算出する。なお、スコア算出手段 153 は、演算式の情報を含め格納している。

10

【数 1】

$$\text{score}(n_{\text{hyper}}) = \sum_{n_{\text{hypo}} \in \text{Desc}(n_{\text{hyper}}) \cap \text{ksimilar}(n_{\text{trg}})} d^{r(n_{\text{hyper}}, n_{\text{hypo}})-1} \times \text{sim}(n_{\text{trg}}, n_{\text{hypo}}),$$

20

【0049】

数式 1 において、 n_{trg} は、対象語である。Desc(n_{hyper}) は、上位語 (n_{hyper}) の下位語（子孫の用語であり、2階層以上、下位の用語も含む）を示す。また、ksimilar(n_{trg}) は、対象語 (n_{trg}) と類似する k 個の用語の集合である。 $r(n_{\text{hyper}}, n_{\text{hypo}})$ は、上位語 (n_{hyper}) と下位語 (n_{hypo}) との木構造中の階層の差を示す。上位語と下位語が親子関係にある場合、 $r(n_{\text{hyper}}, n_{\text{hypo}})$ は 1 である。 d は、階層の差に対するペナルティを算出するための定数であり、0 より大きく 1 未満である。 $d^{r(n_{\text{hyper}}, n_{\text{hypo}})-1}$ は、階層の深さに依存するペナルティである。 $\text{sim}(n_{\text{trg}}, n_{\text{hypo}})$ は、2つの用語 (n_{trg} と n_{hypo}) の類似度を示す。類似度は、類似度算出部 13 が算出した類似度である。

30

【0050】

スコア算出手段 153 は、通常、MPU やメモリ等から実現され得る。スコア算出手段 153 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【0051】

第二上位語取得手段 154 は、少なくとも、スコア算出手段 153 が算出したスコアが最も高い上位語を取得する。第二上位語取得手段 154 は、スコアが最も高い一つの上位語を取得することは好適であるが、スコアが上位の一定数の上位語、スコアが上位の一定割合の上位語、またはスコアが一定以上の複数の上位語を取得しても良い。第二上位語取得手段 154 は、通常、MPU やメモリ等から実現され得る。第二上位語取得手段 154 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

40

【0052】

出力部 16 は、上位語取得部 15 が取得した上位語を出力する。ここで、出力とは、ディスプレイへの表示、プロジェクターを用いた投影、プリンタへの印字、音出力、外部の装置への送信、記録媒体への蓄積、他の処理装置や他のプログラムなどへの処理結果の引渡しなどを含む概念である。出力部 16 は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。出力部 16 は、出力デバイスのドライバース

50

フトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【 0 0 5 3 】

用語辞書蓄積装置 17 は、用語辞書を、用語辞書格納部 11 に蓄積する。用語辞書蓄積装置 17 は、通常、MPU やメモリ等から実現され得る。用語辞書蓄積装置 17 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 0 5 4 】

定義文パターン情報格納部 171 は、1以上の定義文パターン情報を格納している。定義文パターン情報とは、上位語を抽出するための定義文のパターンを示す情報である。定義文パターン情報は、例えば、「とは* <上位語>。」「は、* <上位語>の一つ。」「は、* <上位語>の代表的なものである。」などである。定義文パターン情報格納部 171 は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。定義文パターン情報格納部 171 に定義文パターン情報が記憶される過程は問わない。

【 0 0 5 5 】

定義文取得部 172 は、用語説明文章群から、1以上の定義文パターン情報のうちのいずれか一の定義文パターン情報を適用して、対象語を有する対象語の定義文を取得する。例えば、定義文取得部 172 は、定義文パターン情報「とは* <上位語>。」を適用し、用語説明文章群から、対象語「紅茶」の定義文「紅茶とは、摘み取った茶の葉と芽を乾燥させ、もみ込んで完全発酵させた茶葉。」を取得する。ここで、用語説明文章群とは、用語を説明する文章群であり、用語ごとに、定義文と、カテゴリと、上位下位情報とを有する情報である。上位下位情報とは、階層関係定義情報と上位語と下位語とを有する情報である。階層関係定義情報は、用語の階層関係を特定する情報である。階層関係定義情報の具体例については、後述する。用語説明文章群は、例えば、フリー百科事典「ウィキペディア (Wikipedia)」である。用語説明文章群は、用語辞書蓄積装置 17 が保持していても良いし、ネットワーク上の外部の装置（例えば、インターネット上のサーバ装置）が保持していても良い。定義文取得部 172 は、通常、MPU やメモリ等から実現され得る。定義文取得部 172 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 0 5 6 】

第一用語対候補取得部 173 は、定義文取得部 172 が取得した定義文から、適用された一の定義文パターン情報に従って、対象語の上位語の候補である第一の上位語候補（上記の例の場合「茶葉」）を取得する。そして、第一用語対候補取得部 173 は、第一の上位語候補と対象語の対である第一用語対候補（上記の例の場合（茶葉，紅茶））を取得する。第一用語対候補取得部 173 は、通常、MPU やメモリ等から実現され得る。第一用語対候補取得部 173 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【 0 0 5 7 】

第二用語対候補取得部 174 は、用語説明文章群から、対象語のカテゴリを対象語の第二の上位語候補として取得する。そして、第二用語対候補取得部 174 は、当該第二の上位語候補と対象語の対である第二用語対候補を取得する。第二用語対候補取得部 174 は、例えば、用語説明文章群から、[[Category: <上位語>]]のパターンに合致する文字列を取得し、当該文字列から<上位語>を取得する。この<上位語>が、第二の上位語候補である。第二用語対候補取得部 174 は、例えば、対象語「紅茶」を説明した用語説明文章群から、文字列[[Category: 茶]]を取得し、当該文字列から上位語「茶」を第二の上位語候補として取得する。そして、第二用語対候補取得部 174 は、第二用語対候補（茶，紅茶）を得る。第二用語対候補取得部 174 は、通常、MPU やメモリ等から実現され得る。第二用語対候補取得部 174 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア（専用回路）で

10

20

30

40

50

実現しても良い。

【0058】

階層関係定義情報格納部175は、1以上の階層関係定義情報を格納し得る。階層関係定義情報とは、用語の階層関係を特定する情報である。階層関係定義情報は、例えば、「=+ title =+」「;title:def」「#+title」「*+title」などである。titleは、見出しを示す。また、+は直前の記号(=や#など)が連続して出現し得ることを示す。階層関係定義情報格納部175は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。階層関係定義情報格納部175に階層関係定義情報が記憶される過程は問わない。

【0059】

第三用語対候補取得部176は、階層関係定義情報を用いて、上位語と下位語との対である1以上の第三用語対候補を取得する。第三用語対候補取得部176は、例えば、対象語「紅茶」を説明する用語説明文章群から、「=ブレンドティー・・・;チャイ・・・」を取得し、用語の階層情報「紅茶-ブレンドティー-チャイ」を取得する。そして、第三用語対候補取得部176は、第三用語対候補(紅茶,ブレンドティ)(ブレンドティー,チャイ)(紅茶,チャイ)を取得する。第三用語対候補取得部176は、通常、MPUやメモリ等から実現され得る。第三用語対候補取得部176の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

【0060】

素性ベクトル構成部177は、第一用語対候補を有する文または文の一部と、第二用語対候補を有する文または文の一部と、第三用語対候補を有する文または文の一部とから、言語処理した結果である1以上の素性を取得し、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれの素性ベクトルを構成する。第一用語対候補を有する文または文の一部とは、第一用語対候補のみであってもよい。同様に、第二用語対候補を有する文または文の一部は、第二用語対候補のみであってもよい。第三用語対候補を有する文または文の一部は、第三用語対候補のみであってもよい。なお、上記の言語処理とは、形態素解析やパターンマッチングなどである。

【0061】

また、素性ベクトル構成部177が取得する素性は、例えば、以下の8種類がある。第一は、素性「POS」である。素性「POS」は、上位語と下位語がそれぞれ特定の品詞を含む場合、または上位語と下位語の末尾の形態素の品詞が特定の品詞である場合に発火する素性である。発火するとは、例えば、値が「1」であり、発火しないとは、例えば、値が「0」である。第二の素性「MORPH」は、上位語と下位語がそれぞれ特定の形態素を含むとき発火する素性である。第三の素性「EXP」は、上位語と下位語がそれぞれ特定の文字列に一致するときに発火する素性である。第四の素性「ATTR」は、上位語と下位語がそれぞれ属性語に一致するときに発火する素性である。属性語とは、事物の様々な観点を表現した語で、上位語または下位語になりにくい語である。属性語について、例えば、「紅茶」の属性語は、「生産量」や「産地」である。素性ベクトル構成部177は、1以上の属性語を予め保持している、とする。なお、用語に対して属性語が存在しない場合もあり得る。かかる場合、素性ベクトル構成部177は、上位語と下位語が属性語に一致しない、こととなる。また、第五の素性「LCHAR」は上位語と下位語の末尾の1文字が同じであるとき発火する素性である。第五の素性「LCHAR」は、例えば、「高校/公立校」のように、末尾の1文字が同じであるとき発火する。第六の素性「PAT」は、階層構造から取り出した上位下位関係候補の上位語が、特定の文字列パターンにマッチするときに発火する素性である。ここでの特定の文字列パターンは、例えば、「代表的なX」「代表X」「主要なX」「主なX」「主要X」「基本的なX」「基本X」「著名なX」「大きなX」「他のX」「一部X」「代表的X」「基本的X」「著名X」「一部のX」「Xの一覧」「X一覧」「X詳細」「Xリスト」「Xの詳細」などである。ここで、Xは「上位語」である。なお、素性ベクトル構成部177は、1以上の特定の文字列パタ

10

20

30

40

50

ーンを予め保持している、とする。第七の素性「L A Y E R」は、階層構造で上位語と下位語に付与されていた修飾記号の種類に応じて発火する素性である。第八の素性「D I S T」は、抽出元の階層構造中での上位語候補と下位語候補の距離（辺の数）に応じて発火する素性である。上記の素性を取得する技術は、公知の言語処理技術であるので、詳細な説明を省略する。

【0062】

また、素性ベクトル構成部177は、第一用語対候補または第二用語対候補を有する文または文の一部から取得する素性は、例えば、素性「P O S」「M O R P H」「E X P」「A T T R」「L C H A R」の5つを用いることは好適である。さらに、素性ベクトル構成部177は、第三用語対候補を有する文または文の一部から取得する素性は、例えば、

10

【0063】

素性ベクトル構成部177は、通常、M P Uやメモリ等から実現され得る。素性ベクトル構成部177の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはR O M等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【0064】

機械学習部178は、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれについて、対応する素性ベクトルを、機械学習器に入力して、第一用語対候補と第二用語対候補と第三用語対候補のそれぞれが、上位語と下位語の関係にあるか否かを判断する。機械学習器を実現するアルゴリズムの例は、例えば、サポートベクターマシンである。また、機械学習器を実現するアルゴリズムの例は、例えば、決定木である。また、機械学習部178は、学習データを1以上（通常、多数）格納している。学習データとは、素性ベクトルと判定結果の対の情報である。判定結果とは、用語対候補が上位語と下位語の関係にあるか否か示す情報である。機械学習部178は、通常、M P Uやメモリ等から実現され得る。機械学習部178の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはR O M等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

20

【0065】

用語対蓄積部179は、機械学習部178が、上位語と下位語の関係にあると判断した用語対候補が有する上位語および下位語を、用語辞書格納部11に蓄積する。用語対蓄積部179は、通常、M P Uやメモリ等から実現され得る。用語対蓄積部179の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはR O M等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

30

【0066】

次に、上位語取得装置1の動作について、図3のフローチャートを用いて説明する。図3のフローチャートにおいて、用語辞書蓄積装置17の動作は含めない。

【0067】

（ステップS301）受付部12は、対象語を受け付けたか否かを判断する。対象語を受け付ければステップS302に行き、対象語を受け付けなければステップS301に戻る。

40

【0068】

（ステップS302）類似度算出部13は、カウンタ*i*に1を代入する。

【0069】

（ステップS303）類似度算出部13は、用語辞書格納部11に、*i*番目の用語が存在するか否かを判断する。*i*番目の用語が存在すればステップS304に行き、*i*番目の用語が存在しなければステップS307に行く。

【0070】

（ステップS304）類似度算出部13は、対象語と、*i*番目の用語との類似度を算出する。類似度を算出するアルゴリズムの例は、図4のフローチャートを用いて説明する。

【0071】

50

(ステップ S 3 0 5) 類似度算出部 1 3 は、ステップ S 3 0 4 で算出した類似度を、 i 番目の用語と対にして、バッファ等の記憶媒体に一時蓄積する。

【0072】

(ステップ S 3 0 6) 類似度算出部 1 3 は、カウンタ i を 1、インクリメントする。ステップ S 3 0 3 に戻る。

【0073】

(ステップ S 3 0 7) 類似語取得部 1 4 は、ステップ S 3 0 5 で蓄積された用語のうち、類似度が上位 k 個 (k は 1 以上の整数) の用語 (ここでは、下位語である。) を取得する。

【0074】

(ステップ S 3 0 8) 上位語取得部 1 5 は、カウンタ i に 1 を代入する。

【0075】

(ステップ S 3 0 9) 上位語取得部 1 5 は、条件「 $i \leq k$ 」であるか否かを判断する。条件を満たせばステップ S 3 1 0 に行き、条件を満たさなければステップ S 3 1 8 に行く。

【0076】

(ステップ S 3 1 0) 上位語取得部 1 5 は、カウンタ j に 1 を代入する。

【0077】

(ステップ S 3 1 1) 上位語取得部 1 5 の第一上位語取得手段 1 5 1 は、用語辞書格納部 1 1 に、 i 番目の用語 (下位語) に対する j 番目の上位語が存在するか否かを判断する。 j 番目の上位語が存在すればステップ S 3 1 2 に行き、 j 番目の上位語が存在しなければステップ S 3 1 7 に行く。

【0078】

(ステップ S 3 1 2) スコア算出手段 1 5 3 は、下位語と上位語との階層差を用いて、階層差ペナルティを算出する。階層差ペナルティとは、下位語と上位語との階層差が大きくなればなるほど、値が大きくなるものである。階層差ペナルティは、例えば、数式 1 の右辺の前段部 ($\times \text{sim}(n_{\text{hyper}}, n_{\text{hypo}})$) を除いた部分である。なお、スコア算出手段 1 5 3 は、階層差取得手段 1 5 2 が取得した 2 つの用語の階層差を用いて、階層差ペナルティを算出する。

【0079】

(ステップ S 3 1 3) スコア算出手段 1 5 3 は、用語の類似度を読み出す。

【0080】

(ステップ S 3 1 4) スコア算出手段 1 5 3 は、ステップ S 3 1 2 で算出した階層差ペナルティと、ステップ S 3 1 3 で読み出した用語の類似度とを用いて、スコアを算出する。スコア算出手段 1 5 3 は、例えば、数式 1 により、スコアを算出する。

【0081】

(ステップ S 3 1 5) 第二上位語取得手段 1 5 4 は、ステップ S 3 1 4 で算出されたスコアと、上位語とを対にして、バッファ等に一時蓄積する。

【0082】

(ステップ S 3 1 6) 上位語取得部 1 5 は、カウンタ j を 1、インクリメントする。ステップ S 3 1 1 に戻る。

【0083】

(ステップ S 3 1 7) 上位語取得部 1 5 は、カウンタ i を 1、インクリメントする。ステップ S 3 0 9 に戻る。

【0084】

(ステップ S 3 1 8) 第二上位語取得手段 1 5 4 は、ステップ S 3 1 5 で蓄積した上位語のうち、最大のスコアに対応する上位語を取得する。

【0085】

(ステップ S 3 1 9) 出力部 1 6 は、ステップ S 3 1 8 で取得された上位語を出力する。ステップ S 3 0 1 に戻る。

10

20

30

40

50

【0086】

なお、図3のフローチャートにおいて、ステップS303で取得する用語の集合を予め絞り込んで良い。

【0087】

さらに、図3のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。

【0088】

次に、ステップS304の類似度算出処理について、図4のフローチャートを用いて説明する。

【0089】

(ステップS401)確率分布情報取得手段132は、対象語の確率分布情報(通常、ベクトル)を、確率分布情報格納手段131から取得する。

【0090】

(ステップS402)確率分布情報取得手段132は、i番目の用語の確率分布情報(通常、ベクトル)を、確率分布情報格納手段131から取得する。

【0091】

(ステップS403)類似度算出手段133は、ステップS401で取得された確率分布情報と、ステップS402で取得された確率分布情報との距離を算出する。

【0092】

(ステップS404)類似度算出手段133は、ステップS403で算出した距離を用いて、類似度を算出する。上位処理にリターンする。

【0093】

なお、図4のフローチャートにおいて、2つの確率分布情報(ベクトル)の距離を算出する方法は問わない。

【0094】

次に、用語辞書蓄積装置17の動作について、図5のフローチャートを用いて説明する。

【0095】

(ステップS501)用語辞書蓄積装置17は、カウンタiに1を代入する。

【0096】

(ステップS502)用語辞書蓄積装置17は、用語説明文章群(例えば、いわゆるウィキペディア)に、i番目の用語の説明が存在するか否かを判断する。i番目の用語の説明が存在すればステップS503に行き、存在しなければ処理を終了する。ここで、「用語の説明が存在する」とは、用語が用語説明文章群の見出し語となっていなくてもかまわない。

【0097】

(ステップS503)用語辞書蓄積装置17は、i番目の用語の説明を、用語説明文章群から取得する。なお、用語の説明の区切りは、例えば、予め決められた文字コード(例えば、改ページコードなど)や、文字列などによる。その他、用語の説明の区切りの検出方法は問わない。

【0098】

(ステップS504)定義文取得部172は、ステップS503で取得されたi番目の用語の説明から、1以上の定義文パターン情報のうちのいずれか一の定義文パターン情報を適用して、対象語を有する対象語の定義文を取得する。なお、定義文パターン情報は、定義文パターン情報格納部171に格納されている。

【0099】

(ステップS505)第一用語対候補取得部173は、ステップS504で取得された定義文から、適用された一の定義文パターン情報に従って、対象語の上位語の候補である第一の上位語候補を取得する。そして、第一用語対候補取得部173は、第一の上位語候補と、i番目の用語とからなる第一用語対候補を構成し、バッファ等に一時格納する。

10

20

30

40

50

【0100】

(ステップS506)第二用語対候補取得部174は、ステップS503で取得された*i*番目の用語の説明から、用語のカテゴリを取得する。なお、カテゴリは複数でも良い。また、このカテゴリは、用語の第二の上位語候補である。

【0101】

(ステップS507)第二用語対候補取得部174は、ステップS506で取得した第二の上位語候補と、*i*番目の用語とからなる第二用語対候補を構成し、バッファ等に一時格納する。なお、第二用語対候補は、第二の上位語候補の数だけ格納される。

【0102】

(ステップS508)第三用語対候補取得部176は、階層関係定義情報格納部175から、階層関係定義情報を読み出す。

10

【0103】

(ステップS509)第三用語対候補取得部176は、ステップS508で読み出した階層関係定義情報を用いて、*i*番目の用語の説明から、上位語と下位語との対である1以上の第三用語対候補を構成し、バッファ等に一時格納する。

【0104】

(ステップS510)素性ベクトル構成部177は、カウンタ*j*に1を代入する。

【0105】

(ステップS511)素性ベクトル構成部177は、*j*番目の用語対候補(第一用語対候補、第二用語対候補、または第三用語対候補)が、バッファ等の中に存在するか否かを判断する。*j*番目の用語対候補が存在すればステップS512に行き、存在しなければステップS517に行く。

20

【0106】

(ステップS512)素性ベクトル構成部177は、用語対候補の種類に応じた素性ベクトルを構成する。用語対候補の種類とは、第一用語対候補、第二用語対候補、または第三用語対候補のいずれかである。つまり、素性ベクトル構成部177は、用語対候補(上位語と下位語)に対して、例えば、形態素解析やパターンマッチングなどの言語処理を行い、上述した8つの素性を取得し、8つの値からなるベクトル(素性ベクトル)を構成する。

【0107】

(ステップS513)機械学習部178は、サポートベクターマシン(SVM)に、素性ベクトルを渡し、SVMを実行して、判定結果を取得する。

30

【0108】

(ステップS514)用語対蓄積部179は、ステップS513で取得された判定結果が、「用語対候補が上位語と下位語の関係にある(例えば、「1」)」との判定結果である場合はステップS515に行き、「用語対候補が上位語と下位語の関係にない(例えば、「0」)」との判定結果である場合は、ステップS516に行く。

【0109】

(ステップS515)用語対蓄積部179は、*j*番目の用語対候補を、上位語および下位語として、用語辞書格納部11に蓄積する。

40

【0110】

(ステップS516)素性ベクトル構成部177は、カウンタ*j*を1、インクリメントする。ステップS511に戻る。

【0111】

(ステップS517)素性ベクトル構成部177は、カウンタ*i*を1、インクリメントする。ステップS502に戻る。

【0112】

なお、図5のフローチャートにおいて、SVMは公知技術であるので、詳細な説明を省略する。

【0113】

50

以下、本実施の形態における上位語取得装置 1 の具体的な動作について説明する。

【0114】

今、用語辞書格納部 11 は、図 6 に示すような、2 以上の用語間の階層関係を管理している。図 6 において、用語間の階層関係をツリー状に表現しているが、内部データの構造は問わない。用語辞書は、上位語と下位語の対の集合などでも良い。用語辞書は、上位語と下位語とをリンクで接続しているデータ構造でも良い。図 6 において、「自動車 A」「自動車 B」などは、具体的な車種を示す名称である、とする。また、階層関係とは、概念の上位、下位の階層を示す。図 6 の用語辞書は、例えば、ウィキペディアから取得された約 95,000 の上位語と、1,200,000 の下位語を有する。

【0115】

また、確率分布情報格納手段 131 は、例えば、図 7 に示す確率分布管理表を保持している。確率分布管理表は、用語ごとに、確率分布情報を有する情報である。確率分布情報は、上述したように、用語が、1 以上の各クラスに属する確率の分布の情報である。例えば、確率分布情報 (0.1, 0.05, 0, 0.2, ...) における「0.1」は、用語が第一のクラスの動詞と助詞と共起する確率を示す。確率分布情報 (0.1, 0.05, 0, 0.2, ...) における「0.05」は、用語が第二のクラスの動詞と助詞と共起する確率を示す。

【0116】

かかる場合、まず、受付部 12 は、対象語「自動車 E」を受け付けた、とする。そして、上位語取得装置 1 は、以下に説明する処理を行って、「自動車 E」の上位語を出力する、とする。以下、その処理について説明する。

【0117】

まず、類似度算出部 13 は、図 6 の用語辞書の用語と対象語「自動車 E」との類似度を算出する。以下、類似度を算出する方法の例を 2 つ説明する。第一を RVD と呼び、第二を CVD と呼ぶ。

【0118】

RVD および CVD において、類似度算出部 13 は、以下のように用語を選択する。まず、類似度算出部 13 は、用語 (名詞 n) と共起する $\langle v, rel \rangle$ の数 (種類の数であり、出現頻度ではない) を用いて用語を選択する。さらに具体的には、例えば、類似度算出部 13 は、用語と共起する $\langle v, rel \rangle$ の種類数を全てカウントし、実験では、その数の上位一定数の用語 (例えば、100 万語) を選択する。なお、名詞 n と助詞 rel からなる文節が、動詞 v を含む文節を修飾するとき、「名詞 n (用語) が $\langle v, rel \rangle$ と共起する」とする。

【0119】

そして、類似度算出部 13 は、図 6 の用語辞書 (ウィキペディアから取得された用語辞書) から、28,015 の上位語 (共通上位語という) と、175,022 の下位語 (共通下位語という) とを抽出した、とする。これらの共通上位語は、対象語「自動車 E」の上位語の候補となる。また、これらの共通下位語は、適切な上位語を特定するための手がかりとなる。共通上位語とは、用語辞書から抽出した上位下位関係の中の上位語と、類似度算出部 13 の対象とした用語の集合の中で共通する用語である。共通下位語とは、同様に用語辞書から抽出した上位下位関係の中の下位語と、類似度算出部 13 の対象とした用語の集合の中で共通する用語である。

【0120】

ウェブから取得した用語であり、ウィキペディアから取得された用語辞書に存在しない潜在的な対象語は、約 810,000 存在する。これらの対象語のうち、類似度算出部 13 は、形態素解析に失敗する用語や句を除く処理を行う。その結果、類似度算出部 13 は、約 670,000 の用語を取得できた。そのうちの 하나가、対象語「自動車 E」とする。

【0121】

類似度算出部 13 の確率分布情報取得手段 132 は、「自動車 E」の確率分布情報を、

10

20

30

40

50

図 7 に示す確率分布管理表から読み出す。

【 0 1 2 2 】

次に、確率分布情報取得手段 1 3 2 は、図 6 の用語辞書の上記の共通下位語の集合から（なお、用語辞書のすべての用語を用いても良い。）、順に用語を選択し、各用語の確率分布情報を、図 7 に示す確率分布管理表から読み出す。

【 0 1 2 3 】

次に、類似度算出手段 1 3 3 は、対象語「自動車 E」の確率分布情報、および各用語の確率分布情報を用いて、対象語「自動車 E」と各用語の類似度を算出する。

【 0 1 2 4 】

RVD は、以下のようなアルゴリズムである。RVD において、類似度算出手段 1 3 3 は、 $\langle v, rel, n \rangle$ を用いる。「v」は動詞、「n」は名詞である。「rel」は「v」と「n」の関係を示す。日本語において、関係「rel」は、名詞または名詞句「n」に後続し、動詞「v」に係る助詞である。そして、類似度算出手段 1 3 3 は、 $\langle v, rel, n \rangle$ を 2 つの部分に分割する。第一は $\langle v, rel \rangle$ である。第二は、 $\langle n \rangle$ である。次に、類似度算出手段 1 3 3 は、 $\langle v, rel \rangle$ の組の発生の条件付き確率「 $P(\langle v, rel \rangle | n)$ 」を取得する。「 $P(\langle v, rel \rangle | n)$ 」は、名詞句 n の文法的なコンテキストの確率分布である。そして、類似度算出手段 1 3 3 は、確率分布の距離を算出する。なお、確率分布の距離を算出す方法は種々ある。類似度算出手段 1 3 3 は、例えば、「Jensen-Shannon divergence」を用いる。「Jensen-Shannon divergence」は、2 つの確率分布「 $P(\cdot | n_1)$ 」「 $P(\cdot | n_2)$ 」の距離「 $D_{JS}(P(\cdot | n_1) || P(\cdot | n_2))$ 」を、以下の数式 2 のように算出する。

【 数 2 】

$$D_{JS}(P(\cdot | n_1) || P(\cdot | n_2)) \\ = \frac{1}{2}(D_{KL}(P(\cdot | n_1) || \frac{P(\cdot | n_1) + P(\cdot | n_2)}{2}) \\ + D_{KL}(P(\cdot | n_2) || \frac{P(\cdot | n_1) + P(\cdot | n_2)}{2})),$$

【 0 1 2 5 】

数式 2 において、 D_{KL} は、「the Kullback-Leibler divergence」を示す。 D_{KL} は、数式 3 で示される。

【 数 3 】

$$D_{KL}(P(\cdot | n_1) || P(\cdot | n_2)) = \sum P(\cdot | n_1) \log \frac{P(\cdot | n_1)}{P(\cdot | n_2)}.$$

【 0 1 2 6 】

そして、類似度算出手段 1 3 3 は、例えば、「 $D_{JS}(P(\cdot | n_1) || P(\cdot | n_2))$ 」を用いて、2 つの用語「 n_1 」「 n_2 」の類似度「 $sim(n_1, n_2)$ 」を、数式 4 により算出する。

【 数 4 】

$$sim(n_1, n_2) = 1 - D_{JS}(P(\cdot | n_1) || P(\cdot | n_2)).$$

【 0 1 2 7 】

ここで、類似度は、0 から 1 までの値をとり得る。類似する度合いが大きければ、類似度は 1 に近づき、類似する度合いが小さければ、類似度は 0 に近づく。

【 0 1 2 8 】

なお、以下のように、図示しない手段により、図 7 に示すような確率分布管理表を構築した。つまり、1, 0 0 0, 0 0 0 の名詞句と、1 0 0, 0 0 0 の動詞と助詞のセットを

10

20

30

40

50

用いて、確率「 $P(\langle v, rel \rangle | n)$ 」をウェブコーパス (Shinzatoらが発表した以下のコーパス「K. Shinzato, D. Kawahara, C. Hashimoto and S. Kurohashi. 2008. A Large-Scale Web Data Collection as A Natural Language Processing Infrastructure. In the 6th International Conference on Language Resources and Evaluation (LREC).」) から取得した。

【0129】

なお、確率「 $P(\langle v, rel \rangle | n)$ 」は、以下の数式5により算出できた。

【数5】

$$P(\langle v, rel \rangle | n) = \frac{\log(f(\langle v, rel, n \rangle)) + 1}{\sum_{\langle v, rel \rangle \in D} \log(f(\langle v, rel, n \rangle)) + 1} \quad \text{if } f(\langle v, rel, n \rangle) > 0, \quad 10$$

【0130】

また、数式5において、 \log を使っているが、 \log を使わなくても良い。よって、数式5は、「 $P(\langle v, rel \rangle | n) = (f(\langle v, rel, n \rangle) + 1) / (f(\langle v, rel, n \rangle) + 1)$ 」でも良い。

【0131】

数式5において、「 $f(\langle v, rel, n \rangle)$ 」は、 $\langle v, rel, n \rangle$ の出現頻度である。また、 D は、 $\{\langle v, rel \rangle | f(\langle v, rel, n \rangle) > 0\}$ として定義されるセットである。また、「 $f(\langle v, rel, n \rangle) = 0$ 」の場合、「 $P(\langle v, rel \rangle | n)$ 」は、「0」である。 20

【0132】

次に、CVDについて説明する。CVDにおいて、「EM-based clustering」というクラス分類方法により、名詞を分類する方法を用いる。CVDにおいて、 $\langle v, rel, n \rangle$ の組の出現確率は、以下の数式6で示される。

【数6】

$$P(\langle v, rel, n \rangle) =_{\text{def}} \sum_{a \in A} P(\langle v, rel \rangle | a) P(n | a) P(a), \quad 30$$

【0133】

数式6において、「 a 」は $\langle v, rel \rangle$ の組および「 n 」の隠れクラスを示す。数式6において、確率「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」および「 $P(a)$ 」が直接的に算出できない。隠れクラス「 a 」が与えられたコーパスから取得できないからである。

【0134】

「EM-based clustering」は、与えられたコーパスから、これらの確率（「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」および「 $P(a)$ 」）を推定する。「EM-based clustering」は「Eステップ」と「Mステップ」の2つのステップからなる。「Eステップ」において、確率「 $P(\langle v, rel \rangle | a)$ 」が算出される。「Mステップ」において、「Eステップ」における結果を用いて、最大尤度になるまで、「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」および「 $P(a)$ 」が更新される。 40

【0135】

以上の処理により、各 $\langle v, rel \rangle$ 、 n 、および a に対して、確率「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」および「 $P(a)$ 」が算出される。

【0136】

そして、「 $P(a | n)$ 」は、以下の数式7により算出される。

【数 7】

$$P(a|n) = \frac{P(n|a)P(a)}{\sum_{a \in A} P(n|a)P(a)}$$

【0137】

「 $P(a|n)$ 」は、 n のクラスを決定するために用いられる。例えば、最大の「 $P(a|n)$ 」を有するクラスが、 n が属するクラスである。類似する $\langle v, rel \rangle$ の組と共起する名詞句は、同じクラスに属する傾向がある。

【0138】

以上により、対象語と各用語との類似度が算出され、用語ごとに類似度が一時蓄積された。 10

【0139】

次に、類似語取得部 14 は、蓄積された用語のうち、類似度が上位 k 個 (k は 1 以上の整数) の用語 (ここでは、下位語である。) を取得する。ここで、類似語取得部 14 は、対象語「自動車 E」について、共通下位語の集合から、「自動車 A」「自動車 B」「自動車 C」「自動車 D」を取得した、とする。

【0140】

次に、上位語取得部 15 が、上記で算出した類似度を用いて、適切な上位語を取得する処理について説明する。

【0141】

まず、上位語取得部 15 の第一上位語取得手段 151 は、類似語取得部 14 が取得した k 個の各下位語に対応する上位語を取得する。つまり、第一上位語取得手段 151 は、上位語「小型車」「乗り物」「自動車」を取得する。 20

【0142】

次に、階層差取得手段 152 は、第一上位語取得手段 151 が取得した各上位語に対して、上位語と上位語に対応する下位語との階層差を取得する。つまり、「自動車 A」「自動車 B」「自動車 C」について、親である「自動車」との階層差「1」、祖父母である「乗り物」との階層差「2」を取得する。また、「自動車 D」について、親である「小型車」に対する階層差「1」と、祖父母である「自動車」に対する階層差「2」、さらに親である「乗り物」に対する階層差「3」を取得する。 30

【0143】

そして、スコア算出手段 153 は、上記の数式 1 を用いて、対象語の上位語としての相応しさを示すスコアを、上位語ごとに算出する。

【0144】

スコアの算出の結果、各上位語は、対象語に対するスコアを有することとなる。そして、次に、第二上位語取得手段 154 は、少なくとも、スコア算出手段 153 が算出したスコアが最も高い上位語を取得する。ここでは、第二上位語取得手段 154 は、スコアが最も高い上位語「自動車」を取得した、とする。

【0145】

そして、出力部 16 は、対象語「自動車 E」に対する上位語「自動車」を出力する。なお、図 8 において、波線の矩形で囲まれた用語は、対象語「自動車 E」と類似する k 個の共通下位語である。また、実線の矩形で囲まれた用語は、共通上位語である。さらに、矢印は、階層差を示す。 40

(実験結果)

【0146】

以下、実験の結果について説明する。まず、実験結果において、比較対象の 3 つのベースラインアプローチについて説明する。

【0147】

第一のベースラインアプローチは、対象語に最も類似する下位語の親の用語を取得する方法である。このアプローチにおいて、共通下位語と対象語との類似度を算出する。そし 50

て、共通下位語の中で、最も対象語に類似する共通下位語を取得する。そして、最も対象語に類似する共通下位語の親の用語を、上位語として取得する。第一のベースラインアプローチにおいて、共通下位語が複数の親の用語を持つ場合、複数の上位語が取得され得る。なお、共通下位語と対象語との類似度を算出する場合、例えば、上述した数式 2、数式 3、および数式 4 を用いても良い。なお、最も類似する下位語との Jensen-Shannon divergence (数式 2 の値) が一定値 以上のものは、上位語を抽出しないものとする。

【0148】

第二のベースラインアプローチは、共通上位語と対象語との類似度を算出し、当該類似度を用いる方法である。類似度は、上下関係の適切さとして利用される。

【0149】

第三のベースラインアプローチは、共通上位語の子の用語群との類似度の平均値を用いる方法である。このアプローチでは、上位語 (n_{hyper}) の子の用語によって、確率 $P_{\text{child}}(\cdot | n_{\text{hyper}})$ が、以下の数式 8 により定義される。

【数 8】

$$P_{\text{child}}(\cdot | n_{\text{hyper}}) = \frac{\sum_{n_{\text{hypo}} \in \text{Ch}(n_{\text{hyper}})} P(\cdot | n_{\text{hypo}}) P(n_{\text{hypo}})}{\sum_{n_{\text{hypo}} \in \text{Ch}(n_{\text{hyper}})} P(n_{\text{hypo}})},$$

【0150】

数式 8 において、 $\text{Ch}(n_{\text{hyper}})$ は、上位語 (n_{hyper}) のすべての子のセットである。そして、上位語 (n_{hyper}) の確率分布と対象語 (n_{hypo}) の確率分布との類似度が計算される。そして、最も類似度が大きい確率分布に対応する上位語 (n_{hyper}) が上位語として取得される。

【0151】

第三のベースラインアプローチにおいて、用語辞書 (ウィキペディア関係データベース) が不正確な上下関係を含み得るので、上位語の確率分布の信頼性は低くなる、と考えられる。そのため、本アプローチにおいて、子の用語の数が閾値より多い場合にだけ、上位語として使用した。

【0152】

以下、3つのベースラインアプローチと、本実施の形態における上位語取得装置 1 による方法とを比較した実験結果について説明する。この実験は、約 67 万の名詞句から抽出した各対象語の上位語をウィキペディア関係データベースから取得する場合の実験である。

【0153】

上位語取得装置 1 による方法では、いくつかのパラメータが存在する。そして、実験において、ランダムに、694 の単語を選択し、パラメータを最適化した。そして、手作業により、694 の単語の上位語を決定した。また、パラメータを最適化したので、上記の 4 つの方法において、最良の性能を実現できた。

【0154】

RVD において、パラメータは、類似度の高い選択される用語の数「 $k = 100$ 」、類似度の閾値「 0.05 」、および階層差を用いたペナルティの定数値「 $d = 0.6$ 」である。また、CVD において、パラメータは、類似度の高い選択される用語の数「 $k = 200$ 」、類似度の閾値「 0.3 」、および階層差を用いたペナルティの定数値「 $d = 0.6$ 」である。

【0155】

さらに、第三のベースラインアプローチにおいて、パラメータは、子の用語の数の閾値であり、その閾値を「20」とした。つまり、19 以下の子の用語の数しか有さない場合、その用語は、上位語には採用しなかった。

【0156】

10

20

30

40

50

上記の調整したパラメータを用いて、以下のように実験を行った。なお、実験において、 k 個の類似する用語を取得する場合に、階層差を用いたペナルティを算出するための定数「 $d = 0$ 」の場合にも実験を行った。この実験は、「 k - similar words (CVD, $d = 0$)」である。また、この実験は、 k 個の類似する用語について、それらの用語の親の用語のみを考慮することを意味する。

【0157】

また、実験において、評価データ（スコアを算出する対象の用語）を、類似度が上位1,000個のセット、10,000のセット、100,000のセット、670,000のセット、の4つのセットとした。

【0158】

そして、各セットから、ランダムに200個のサンプルを抽出し、各種の方法で抽出された上位語が、正しい上位語であるか否かを判断した。判断において、「<対象語>は<上位語>の一種である」「<対象語>は<上位語>の一例である」などの文が妥当と考えられる場合、上下関係は正しいと判断した。

【0159】

また、第一のベースラインアプローチにおいて、対象語の上位語が複数、存在する場合、いずれかの上位語が、用意された上位語と一致する場合、正解である、とした。

【0160】

その実験結果を、図9に示す。図9において、表内の各数値（0より大きく1以下の数値）は、正しい上位語を抽出した割合を示す。

【0161】

図9において、「 k - similar words」の方法（上位語取得装置1の方法）は、他のベースラインアプローチと比較して、極めて良好な評価を得た。特に、「 k - similar words (CVD)」の方法は、「 k - similar words (CVD, $d = 0$)」「 k - similar words (RVD)」と比較しても、類似度が上位1,000語のセットの場合を除いて、優れている。このことは、用語間の階層構造を用いること、および確率分布の類似度を用いてクラスタリングするプロセスは、有効であることを示している。

【0162】

また、上記により、CVDによる方法は、「the Fisher's exact test」による1%の有意水準における他のベースラインの方法と比較して、著しい優位性を確認できた。

【0163】

また、第二のベースラインアプローチの精度が最も悪かった。不当に抽出された上位語の中に対象語と同様のレベルであった単語があった。例えば、単語「セメント工場」は単語「ドライクリーニング工場」の上位語として抽出された。これは、類似度だけを使用することによって、その単語が上位語か、同レベルの単語であるかを判断するのは、難しいことを示している。

【0164】

また、対象語に最も類似する下位語の上位語を取得する第一のベースラインアプローチと、上位語の子の用語群の類似度を用いる第三のベースラインアプローチは、用語辞書（ウィキペディア関係データベース）における雑音のため、精度があがらなかった。さらに、第一と第三のベースラインアプローチにおいて、不適切な上位語が抽出された。対照的に、「 k - similar words」の方法（上位語取得装置1の方法）は、雑音に対して強健であった。さらに、「 k - similar words」の方法は、スコアを算出するために、すべての子孫の用語を使用するので、適切な上位語を抽出できた。

【0165】

さらに、図10は、「 k - similar words (CVD)」の方法で取得できた上位語の例を示す。

【0166】

また、「 k - similar words (CVD)」を用いてランダムに抽出した3

10

20

30

40

50

00セットの上位語/下位語の組について、文字列のパターンマッチングに基づく従来の方法で、これらの上位語/下位語の組を抽出できるかどうか調査した。その結果、300組のうち、243組の上位語/下位語を抽出できなかった。これは、上位語取得装置1が非常に多数の上位語/下位語の組を抽出できることを示している。

【0167】

次に、用語辞書蓄積装置17が、用語辞書格納部11に用語辞書を蓄積する処理の具体例について説明する。

【0168】

今、用語辞書蓄積装置17は、図11に示す用語説明文章群(ウィキペディア)の情報を保持している。なお、用語説明文章群は、通常、図示しないインターネット上のサーバ装置から取得された情報である。図11は、用語説明文章群のブラウザ表示の内容である。また、図11の表示の元になるデータを図12に示す。

10

【0169】

また、定義文パターン情報格納部171は、「とは* <上位語>。」「は、* <上位語>の一つ。」「は、* <上位語>の代表的なものである。」などの定義文パターン情報を格納している。

【0170】

さらに、階層関係定義情報格納部175は、図13に示す階層関係定義情報管理表を保持している。

【0171】

かかる状況において、まず、定義文取得部172は、定義文パターン情報「とは* <上位語>。」を、用語説明文章群内の文章に適用し、定義文「紅茶とは、摘み取った・・・発酵させた茶葉。」を取得する。

20

【0172】

次に、第一用語対候補取得部173は、定義文「紅茶とは、摘み取った・・・発酵させた茶葉。」から、適用された一の定義文パターン情報「とは* <上位語>。」に従って、対象語の上位語の候補である第一の上位語候補「茶葉」を取得する。そして、第一用語対候補取得部173は、第一の上位語候補と対象語の対である第一用語対候補(茶葉, 紅茶)を取得する。

【0173】

次に、第二用語対候補取得部174は、用語説明文章群から、[[Category: <上位語>]]のパターンに合致する文字列を取得し、当該文字列から<上位語>を取得する。この<上位語>が、第二の上位語候補である。第二用語対候補取得部174は、例えば、対象語「紅茶」を説明した用語説明文章群から、文字列「茶」と「喫茶文化」を取得する。そして、第二用語対候補取得部174は、2つの第二用語対候補(茶, 紅茶)(喫茶文化, 紅茶)を得る。

30

【0174】

次に、第三用語対候補取得部176は、図13の修飾記号に注目し、記事からtitleをノードとするグラフ構造として階層構造を抽出する。具体的には、第三用語対候補取得部176は、titleに付与されている修飾記号の優先度と長さに従ってノードの親子関係を決定することで階層構造を抽出する。例えば、第三用語対候補取得部176は、図11のページから、そのMediaWikiコード(図12)を元に図14のような階層構造を抽出する。定義文やカテゴリを上位下位関係の知識源とする場合、下位語は記事の見出し語に制限されるため、獲得できる下位語の数はWikipediaの記事数より少なくなるが、階層構造ではそのような制限が無い場合、より多様な上位下位関係が獲得できる。

40

【0175】

そして、第三用語対候補取得部176は、図14の階層構造から、(ブレンドティー, チャイ)や(紅茶, リプトン)などの第三用語対候補を抽出する。なお、第三用語対候補取得部176は、冗長な上位語を簡略化するため、図15のパターン(除外パターン)をもつ上位語候補からパターン中のX以外の部分を取り除く。ここで、Xは任意の文字列と

50

する。例えば、上位語「主な紅茶ブランド」はパターン「主なX」を適用することで、「紅茶ブランド」と置換される。

【0176】

次に、素性ベクトル構成部177は、第一用語対候補を有する文または文の一部と、第二用語対候補を有する文または文の一部に対して、用語対候補ごとに、図16に示す素性「POS」「MORPH」「EXP」「ATTR」「LCHAR」「LCHAR」の6つを取得する。そして、素性ベクトル構成部177は、用語対候補ごとに、6つの要素(値)を有する素性ベクトルを取得する。

【0177】

そして、素性ベクトル構成部177は、第三用語対候補を有する文または文の一部に対して、用語対候補ごとに、図16に示すすべての素性を取得する。そして、素性ベクトル構成部177は、用語対候補ごとに、8つの要素(値)を有する素性ベクトルを取得する。

10

【0178】

次に、機械学習部178は、用語対候補ごとに、素性ベクトルを、SVMに入力し、その結果得られたSVMのスコアが閾値以上の用語対候補を正しい上位語/下位語として獲得する。

【0179】

そして、用語対蓄積部179は、機械学習部178が、上位語と下位語の関係にあると判断した用語対候補が有する上位語および下位語を、用語辞書格納部11に蓄積する。

20

【0180】

以上の処理により、用語辞書蓄積装置17は、用語辞書を、用語辞書格納部11に蓄積できる。

【0181】

以上、本実施の形態によれば、大規模な上位語と下位語のデータベースを構築できる。また、本実施の形態によれば、精度の高い上位語と下位語のデータベースを構築できる。さらに、本実施の形態における用語辞書蓄積装置17により、大規模な用語辞書を構築できる。

【0182】

なお、本実施の形態において、上位語取得部15は、以下のように上位語を取得しても良い。例えば、上位語取得部15は、類似語取得部14が取得したk個の各下位語の上位語(最終的に出力する上位語の候補)を取得する。ここで、取得した上位語には、重複があり得る。そして、上位語取得部15は、最も出現頻度の多い上位語の候補(または、上位m個の用語)を、最終的に選択する上位語としても良い。例えば、上位語取得部15が取得した上位語の候補が「自動車」「自動車」「自動車」「自転車」「船」であった場合、上位語取得部15は、「自動車」を上位語として選択しても良い。

30

【0183】

また、本実施の形態において、上位語取得部15は、以下のように上位語を取得しても良い。例えば、上位語取得部15は、各上位語の候補に対して、類似度を加算し、その合計がもっと大きい上位語の候補(または、加算値の上位m個の用語)を、最終的な上位語として選択しても良い。例えば、取り出した上位語の候補の各々が、「自動車 類似度1.1」「自動車 類似度1.0」「自動車 類似度0.9」「自転車 類似度5」「船 類似度0.8」である場合、「自動車」の類似度の和が「3」であり、「自転車」の類似度の和が「5」であり、「船」の類似度の和が「0.8」であるので、上位語取得部15は、「自転車」を上位語とする。

40

【0184】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをCD-ROMなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における情報処理

50

装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、記憶媒体に、2以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納しており、コンピュータを、上位概念の用語を取得する対象となる用語である対象語を受け付ける受付部と、前記受付部が受け付けた対象語と、前記記憶媒体に格納されている1以上の各用語との類似度を算出する類似度算出部と、前記類似度算出部が算出した上位 k (k は1以上の整数)の類似度に対応する k 個の下位語を取得する類似語取得部と、前記類似語取得部が取得した k 個の各下位語に対応する上位語を取得し、当該上位語から m (m は1以上の整数)個の上位語を、前記記憶媒体から取得する上位語取得部と、前記上位語取得部が取得した上位語を出力する出力部として機能させるためのプログラムである。

10

【0185】

また、上記プログラムにおいて、前記上位語取得部は、前記類似語取得部が取得した k 個の各下位語に対応する上位語を取得する第一上位語取得手段と、前記第一上位語取得手段が取得した各上位語に対して、当該上位語と当該上位語に対応する下位語との階層差を取得する階層差取得手段と、前記類似語取得部が取得した類似度をパラメータとする増加関数であり、前記階層差取得手段が取得した階層差をパラメータとする減少関数であるスコア算出の演算式に、前記類似度と前記階層差とを代入し、対象語の上位語としての相応しさを示すスコアを、前記上位語ごとに算出するスコア算出手段と、少なくとも、前記スコア算出手段が算出したスコアが最も高い上位語を取得する第二上位語取得手段を具備するものとして機能するためのプログラムであることは好適である。

20

【0186】

また、上記プログラムにおいて、前記類似度算出部は、動詞と助詞とを有する文字列を1以上有する1以上の各クラスに、用語が属する確率である確率分布情報を、用語ごとに格納し得る確率分布情報格納手段と、前記受付部が受け付けた対象語の確率分布情報、および前記記憶媒体に格納されている1以上の各用語の確率分布情報を取得する確率分布情報取得手段と、前記対象語の確率分布情報、および前記各用語の確率分布情報を用いて、前記対象語と前記各用語の類似度を算出する類似度算出手段とを具備するものとして機能するためのプログラムであることは好適である。

【0187】

また、上記プログラムにおいて、前記用語辞書を、前記記憶媒体に蓄積する用語辞書蓄積装置をさらに具備する上位語取得装置であって、前記用語辞書蓄積装置は、上位語を抽出するための定義文のパターンを示す情報である定義文パターン情報を、1以上格納している定義文パターン情報格納部と、用語を説明する文章群であり、用語ごとに、定義文と、カテゴリと、用語の階層関係を特定する情報である階層関係定義情報と上位語と下位語とを有する用語説明文章群から、前記1以上の定義文パターン情報のうちのいずれか一の定義文パターン情報を適用して、前記対象語を有する対象語の定義文を取得する定義文取得部と、前記定義文抽出部が取得した定義文から、前記適用された一の定義文パターン情報に従って、前記対象語の上位語の候補である第一の上位語候補と前記対象語の対である第一用語対候補を取得する第一用語対候補取得部と、前記用語説明文章群から、前記対象語のカテゴリを前記対象語の第二の上位語候補として、前記第二の上位語候補と前記対象語の対である第二用語対候補を取得する第二用語対候補取得部と、階層関係定義情報を1以上格納し得る階層関係定義情報格納部と、前記階層関係定義情報を用いて、上位語と下位語との対である1以上の第三用語対候補を取得する第三用語対候補取得部と、前記第一用語対候補を有する文または文の一部と、前記第二用語対候補を有する文または文の一部と、前記第三用語対候補を有する文または文の一部とから、言語処理した結果である1以上の素性を取得し、前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれの素性ベクトルを構成する素性ベクトル構成部と、前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれについて、対応する素性ベクトルを、サポートベクターマシンを用いて、前記第一用語対候補と前記第二用語対候補と前記第三用語対候補のそれぞれが、上位語と下位語の関係にあるか否かを判断する機械学習部と、前記機

30

40

50

械学習部が、上位語と下位語の関係にあると判断した用語対候補が有する上位語および下位語を、前記記憶媒体に蓄積する用語対蓄積部とを具備するものとして機能するためのプログラムであることは好適である。

(実施の形態2)

【0188】

本実施の形態において、上位語と下位語とを有し、階層化されている用語辞書に、下位語をさらに付加するデータ作成装置について説明する。つまり、本実施の形態におけるデータ作成装置2は、図17に示すように、既に存在する用語辞書に、上位語または下位語を付加し、大規模な概念辞書を構築するものである。なお、データ作成装置2は、上位語取得装置1が構築した用語辞書を利用することは好適である。なお、階層化されている用語辞書は、例えば、日本語WordNetと言われているものである。また、例えば、付加される上位語と下位語の集合は、例えば、登録語数の多い構築中の概念辞書が好適である。

10

【0189】

図18は、本実施の形態におけるデータ作成装置2のブロック図である。データ作成装置2は、用語辞書格納部11、受付部12、上位語生成部21、用語対情報受付部22、類似度算出部23、類似語取得部24、下位語付加部25を備える。ここで、用語辞書格納部11は、上位語取得装置1が構築したものであることは好適である。

【0190】

用語対情報受付部22は、用語対情報を受け付ける。用語対情報とは、上位語と下位語との対の情報である。ここで、用語対情報の受け付け方法は問わない。ユーザからの手入力の受け付けでも良いし、プログラムから渡されても良いし、記憶媒体から読み出されるなどしても良い。用語対情報の一部(上位語のみ、下位語のみ、または全部)は、データ作成装置2の内部で生成されたものでも良い。例えば、用語対情報は、Webページから自動的に取得されても良い。用語対情報の入力手段は、テンキーやキーボードやマウスやメニュー画面によるもの等、何でも良い。用語対情報受付部22は、テンキーやキーボード等の入力手段のデバイスドライバーや、メニュー画面の制御ソフトウェア等で実現され得る。

20

【0191】

上位語生成部21は、受付部12が受け付けた用語を2以上の文字列に分割し、最後尾の文字列を含む1以上の文字列を有する上位語を取得する。例えば、用語「PCソフト」が受け付けられた場合、上位語生成部21は、用語「PCソフト」を「PC」と「ソフト」に分割し、最後尾の文字列を含む1以上の文字列「ソフト」を「PCソフト」の上位語とする。また、例えば、用語「スピードスケート長距離選手」が受け付けられた場合、上位語生成部21は、用語「スピードスケート長距離選手」を「スピード」「スケート」「長距離」「選手」に分割し、「スケート長距離選手」または/および「長距離選手」または/および「選手」を「スピードスケート長距離選」の上位語とする。また、ここで、ユーザに問い合わせ、ユーザの入力に応じて、不適切な用語(例えば、「長距離選手」)を除くようにしても良い。なお、用語の分割の方法は、例えば、形態素解析によるが、他の方法でも良い。他の方法とは、例えば、所定数の文字(例えば、1文字)の単位に、文字を分割する方法、漢字列とカタカナ列とひらがな列で文字を分割する方法などがある。上位語生成部21は、通常、MPUやメモリ等から実現され得る。上位語生成部21の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはROM等の記録媒体に記録されている。但し、ハードウェア(専用回路)で実現しても良い。

30

40

【0192】

類似度算出部23は、用語対情報受付部22が受け付けた用語対情報が有する上位語と、用語辞書格納部11に格納されている1以上の各用語との類似度を算出する。また、類似度算出部23は、用語対情報受付部22が受け付けた用語対情報が有する下位語と、用語辞書格納部11に格納されている1以上の各用語との類似度も算出することは好適である。類似度算出部23が2つの用語の類似度を算出する方法は、類似度算出部13が2

50

つの用語の類似度を算出する方法と同様であるので、詳細な説明を省略する。つまり、例えば、類似度算出部 23 は、確率分布情報格納手段 131、確率分布情報取得手段 132、類似度算出手段 133 を具備する。なお、類似度算出部 23 が 2 つの用語の類似度を算出するアルゴリズムは問わない。類似度算出部 23 は、通常、MPU やメモリ等から実現され得る。類似度算出部 23 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【0193】

類似語取得部 24 は、類似度算出部 23 が算出した上位 k (k は 1 以上の整数) の類似度に対応する k 個の類似用語を取得する。類似語取得部 24 は、上位語との類似度、および下位語との類似度の両方の類似度を用いて、 k 個の類似用語を取得することは好適である。類似語取得部 24 は、上位語との類似度、および下位語との類似度の和や平均等の値が上位の k 個の類似用語を取得することは好適である。類似語取得部 24 は、通常、MPU やメモリ等から実現され得る。類似語取得部 24 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

10

【0194】

下位語付加部 25 は、受け付けた用語対情報が有する下位語または上位語下位語対の上位語として、 k 個の類似用語を選択して、用語辞書格納部 11 に下位語を蓄積する。上位語下位語対とは、上位語と下位語との対である。下位語付加部 25 は、受け付けた用語対情報が有する下位語または上位語下位語対の上位語として、用語対情報が有する上位語と同一の文字列を有する類似用語を選択して、用語辞書格納部 11 に下位語を蓄積することは好適である。下位語付加部 25 は、用語辞書格納部 11 内で、上位語を特定できる態様で、下位語または上位語下位語対を用語辞書格納部 11 に蓄積する。つまり、「上位語として、 k 個の類似用語を選択して、下位語を蓄積する」とは、蓄積される下位語の上位語がどれか (2 以上でも良い) を特定できるように下位語を蓄積することである。上位語を特定できるような蓄積とは、例えば、下位語と上位語とのリンクの情報を、下位語と一緒に登録したり、下位語と対に上位語の ID を登録したりすることである。下位語付加部 25 は、通常、MPU やメモリ等から実現され得る。下位語付加部 25 の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは ROM 等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

20

30

【0195】

次に、データ作成装置 2 の動作について、図 19 のフローチャートを用いて説明する。図 19 のフローチャートにおいて、図 3 のフローチャートと同一のステップについて、説明を省略する。

【0196】

(ステップ S1901) 受付部 12 は、用語を受け付けたか否かを判断する。用語を受け付ければステップ S1902 に行き、用語を受け付けなければステップ S1902 に戻る。

【0197】

(ステップ S1902) 上位語生成部 21 は、受付部 12 が受け付けた用語を 2 以上の文字列に分割する。上位語生成部 21 は、例えば、形態素解析により、用語を 2 以上の部分の文字列に分割する。形態素解析は公知の技術であるの詳細な説明を省略する。

40

【0198】

(ステップ S1903) 上位語生成部 21 は、分割した文字列の中の、最後尾の文字列を含む 1 以上の文字列を有する上位語を取得する。ここで取得される上位語は、2 以上でも良い。なお、下記のステップにおいて、上位語は一つとして説明する。上位語が 2 以上である場合、ステップ S1904 以降の処理が繰り返し実行される。

【0199】

(ステップ S1904) 用語対情報受付部 22 は、ステップ S1903 で取得された上

50

位語と、ステップ S 1 9 0 1 で受け付けられた用語（下位語）との対の情報を取得する。

【0200】

（ステップ S 1 9 0 5）下位語付加部 2 5 は、ステップ S 3 0 7 で取得された k 個の用語を上位語として、用語（下位語）を蓄積する。ステップ S 1 9 0 1 に戻る。

【0201】

なお、図 1 9 のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。

【0202】

以下、本実施の形態におけるデータ作成装置 2 の具体的な動作について説明する。なお、用語辞書格納部 1 1 の各用語は、用語を識別する ID と同義語群を有する、とする。同義語群は、1 以上の用語の集合であり、1 以上の同義語である。また、例えば、類似度算出部 2 3 は、用語を識別する ID に対応付けて、確率分布情報を保持している、とする。

（具体例 1）

【0203】

今、受付部 1 2 は、上位語下位語対「PC ソフトーぶよぶよ通」を受け付けた、とする。そして、上位語生成部 2 1 は、上位語下位語対のうちの上位語「PC ソフト」を形態素解析し、「PC ソフト」から、分割された文字列「PC」と「ソフト」を得る。

【0204】

次に、上位語生成部 2 1 は、「PC ソフト」の上位語として、「ソフト」を得る。

【0205】

次に、用語対情報受付部 2 2 は、（上位語，下位語）の組である（ソフト，PC ソフト）を受け付ける。

【0206】

次に、類似度算出部 2 3 は、用語対情報のうちの上位語「ソフト」と一致する用語と ID を、用語辞書格納部 1 1 から取得する、とする。ここでは、類似度算出部 2 3 は、用語辞書格納部 1 1 内のすべての用語を対象に類似度を算出するのではなく、用語対情報が有する上位語「ソフト」と同一の文字列の用語のみの類似度を算出する、とする。つまり、類似度算出部 2 3 は、用語辞書格納部 1 1 から、3 つの用語「ID = 0 6 5 6 6 0 7 7，ソフト」、「ID = 0 3 3 2 5 9 4 1，ソフト」、および「ID = 0 7 6 1 4 5 0 0，ソフト」を選択する。なお、「ソフト」は多義性を有する用語である。つまり「ID = 0 6 5 6 6 0 7 7，ソフト」は、同義語として（ソフト，ソフトウェア，ソフトウェア，パッケージ）を有する。「ID = 0 3 3 2 5 9 4 1，ソフト」は、同義語として（ソフト，ソフト帽，フェドーラ，ホンブルグ帽，中折，中折れ，中折れフェルト帽，中折れ帽，中折れ帽子，中折帽，中折帽子）を有する。さらに、「ID = 0 7 6 1 4 5 0 0，ソフト」は、同義語として（アイス，アイスクリーム，クリーム，ソフト，氷菓）を有する。

【0207】

そして、類似度算出部 2 3 は、「ID = 0 6 5 6 6 0 7 7，ソフト」と用語対情報が有する上位語「ソフト」との類似度を 2 . 4 5 5 6 7 1 9 2 と算出した、とする。なお、類似度算出部 2 3 は、類似度算出部 1 3 と同様、例えば、確率分布情報を用いて、2 つの用語の類似度を算出する、とする。

【0208】

また、類似度算出部 2 3 は、「ID = 0 3 3 2 5 9 4 1，ソフト」と用語対情報が有する上位語「ソフト」との類似度を 0 と算出した、とする。

【0209】

さらに、類似度算出部 2 3 は、「ID = 0 7 6 1 4 5 0 0，ソフト」と用語対情報が有する上位語「ソフト」との類似度を 0 と算出した、とする。

【0210】

かかる場合、類似語取得部 2 4 は、最も類似度が大きい、「ID = 0 6 5 6 6 0 7 7，ソフト」を取得する。

【0211】

10

20

30

40

50

そして、下位語付加部 25 は、用語対情報が有する下位語「PCソフト」を、「ID = 06566077, ソフト」の子の用語（下位語）である、として、用語辞書格納部 11 に蓄積する。また、下位語付加部 25 は、上位語下位語対「PCソフトーぶよぶよ通」を、「ID = 06566077, ソフト」にぶら下げて（下位に）蓄積することはさらに好適である。

（具体例 2）

【0212】

今、受付部 12 は、「QRコード マイクロQRコード」を受け付けた、とする。そして、上位語生成部 21 は、「QRコード」を分割し、分割された文字列「QR」と「コード」を得る。そして、上位語生成部 21 は、「QRコード」の上位語として、「コード」を得る。

10

【0213】

次に、用語対情報受付部 22 は、（上位語，下位語）の組である（コード，QRコード）を受け付ける。

【0214】

次に、類似度算出部 23 は、用語対情報のうちの上位語「コード」を用語に含む同語群と ID とを、用語辞書格納部 11 から取得する、とする。そして、類似度算出部 23 は、「06355894（コード，記号，符号）」「06353934（コード，暗号，記号，符丁，符帳，符牒，略号）」「06254239（コード，暗号）」「06869951（コード，和音，和弦）」を得る。なお、前記は「ID（同義語群）」である。

20

【0215】

次に、類似度算出部 23 は、4つの各用語と、用語対情報のうちの上位語「コード」との類似度を、それぞれ「1.311352712」「1.029338」「0」「0」と算出した、とする。

【0216】

次に、下位語付加部 25 は、例えば、類似度が閾値（例えば、「1」）以上の用語「06355894（コード，記号，符号）」および「06353934（コード，暗号，記号，符丁，符帳，符牒，略号）」の下位語として、用語対情報が有する「QRコード マイクロQRコード」を、用語辞書格納部 11 に蓄積する。ここでは、下位語を複数の上位語の子の用語として蓄積した。つまり、親の用語は 2 以上でも良い。

30

（具体例 3）

【0217】

今、受付部 12 は、「SF映画 アルマゲドン」を受け付けた、とする。そして、上位語生成部 21 は、「SF映画」を分割し、分割された文字列「SF」と「映画」と「画」を得る。そして、上位語生成部 21 は、「SF映画」の上位語として、「画」を得た、とする。ここで、上位語生成部 21 は、最上位の用語「画」を取得したが、2 以上の上位語（ここでは、「映画」「画」）を取得しても良い。

【0218】

次に、類似度算出部 23 は、用語対情報のうちの上位語「画」を用語に含む同語群と ID とを、用語辞書格納部 11 から取得する、とする。そして、類似度算出部 23 は、7つの ID 等を取得した、とする。7つの ID 等は、「06613686（エクラン，シネマ，ピクチャー，フィルム，ムービー，ムーヴィー，映画，活動，活動写真，写真，電影）」「03234306（ドロ잉，画，画図，絵，絵画，絵図，図，図案，図画，図絵，図面，線描）」「03561345（さし画，さし絵，イラスト，イラストレーション，カット，ピクチャー，映像，画，画図，画像，画面，絵，絵とき，絵画，絵図，駒絵，小間絵，図，図解，図形，図説，図版，図面，素描，挿し画，挿し絵，挿画，挿絵，挿図，陽画）」「03876519（うつし絵，すがた絵，ピクチャー，ペイント，ペンキ，ペンキ塗り，影像，映し絵，映絵，画，画図，画像，画幅，画面，絵，絵画，絵図，絵像，彩画，彩絵，彩色，姿絵，写し絵，写絵，写真絵，書き絵，書絵，色絵，図画，図絵，丹青，描き絵，描画，描絵，描像）」「06277803（影像，画，画像，画面，絵

40

50

、 絵画、 像）」 「06799260（画）」 「07003119（画、 画図、 絵、 絵画、 絵図、 図、 図案、 図引き、 図画、 図絵、 描画、 描絵）」である。

【0219】

次に、類似度算出部23は、7つの各用語と、用語対情報のうちの上位語「画」との類似度を、それぞれ「2.2976732」「0」「0」「0」「0」「0」「0」と算出した、とする。

【0220】

次に、下位語付加部25は、例えば、類似度が最大の利用語「06613686（エクラ、 シネマ、 ピクチャー、 フィルム、 ムービー、 ムーヴィー、 映画、 活動、 活動写真、 写真、 電影）」の下位語として、用語対情報が有する「SF映画 アルマゲドン」を、用語辞書格納部11に蓄積する。

（具体例4）

【0221】

今、受付部12は、「ICカード トランセカード」を受け付けた、とする。そして、上位語生成部21は、「ICカード」を分割し、分割された文字列「IC」と「カード」を得る。そして、上位語生成部21は、「ICカード」の上位語として、「カード」を得た、とする。

【0222】

次に、用語対情報受付部22は、（上位語、下位語）の組である（カード、ICカード）を受け付ける。

【0223】

次に、類似度算出部23は、用語対情報のうちの上位語「カード」を利用語に含む同語群とIDとを、用語辞書格納部11から取得する、とする。そして、類似度算出部23は、「06627006（カード、メッセージカード、簡、手札、年賀状）」 「14800034（カード、札）」 「02962545（カード、歌留多、札）」 「03033986（カード、札）」 「06507941（カード、スコアカード、札）」を得る。

【0224】

次に、類似度算出部23は、5つの各用語と、用語対情報のうちの上位語「カード」との類似度を、それぞれ「0.317643」「0.315908」「0」「0」「0」と算出した、とする。

【0225】

次に、下位語付加部25は、例えば、類似度が上位4割の利用語「06627006（カード、メッセージカード、簡、手札、年賀状）」 「14800034（カード、札）」の下位語として、用語対情報が有する「ICカード トランセカード」を、用語辞書格納部11に蓄積する。ここでは、下位語を複数の上位語の子の利用語として蓄積した。つまり、親の利用語は2以上でも良い。

【0226】

以上、本実施の形態によれば、大規模な上位語と下位語のデータベースを構築できる。

【0227】

なお、本実施の形態におけるデータ作成装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、記憶媒体に、2以上の用語を有し、かつ、用語間の階層関係を管理している情報である用語辞書を格納しており、コンピュータを、上位語と下位語との対の情報である用語対情報を受け付ける用語対情報受付部と、前記用語対情報受付部が受け付けた用語対情報が有する上位語と、前記用語辞書格納部に格納されている1以上の各用語との類似度を算出する類似度算出部と、前記類似度算出部が算出した上位k（kは1以上の整数）の類似度に対応するk個の類似用語を取得する類似語取得部と、前記受け付けた用語対情報が有する下位語の上位語として、前記k個の類似用語を選択して、前記用語辞書格納部に前記下位語を蓄積する下位語付加部として機能させるためのプログラムである。

【0228】

10

20

30

40

50

また、上記プログラムにおいて、前記下位語付加部は、前記受け付けた用語対情報が有する下位語の上位語として、前記用語対情報が有する上位語と同一の文字列を有する類似用語を選択して、前記記憶媒体に前記下位語を蓄積するものとして機能するためのプログラムであることは好適である。

【0229】

また、上記プログラムにおいて、受け付けられた用語を2以上の文字列に分割し、最後尾の文字列を含む1以上の文字列を有する上位語を取得する上位語生成部をさらに具備し、前記用語対情報受付部は、前記上位語生成部が取得した上位語と、前記受け付けられた用語である下位語との対の情報である用語対情報を受け付けるものとして機能するためのプログラムであることは好適である。

10

【0230】

また、上記プログラムにおいて、前記類似度算出部は、前記用語対情報受付部が受け付けた用語対情報が有する下位語と、前記用語辞書格納部に格納されている1以上の各用語との類似度をも算出し、前記類似語取得部は、前記上位語との類似度、および前記下位語との類似度の両方の類似度を用いて、k個の類似用語を取得するものとして機能するためのプログラムであることは好適である。

【0231】

また、図20は、本明細書で述べたプログラムを実行して、上述した実施の形態の上位語取得装置等を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図20は、このコンピュータシステム340の概観図であり、図21は、コンピュータシステム340のブロック図である。

20

【0232】

図20において、コンピュータシステム340は、FDドライブ、CD-ROMドライブを含むコンピュータ341と、キーボード342と、マウス343と、モニタ344とを含む。

【0233】

図21において、コンピュータ341は、FDドライブ3411、CD-ROMドライブ3412に加えて、MPU3413と、CD-ROMドライブ3412及びFDドライブ3411に接続されたバス3414と、ブートアッププログラム等のプログラムを記憶するためのROM3415とに接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのRAM3416と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク3417とを含む。ここでは、図示しないが、コンピュータ341は、さらに、LANへの接続を提供するネットワークカードを含んでも良い。

30

【0234】

コンピュータシステム340に、上述した実施の形態の上位語取得装置等の機能を実行させるプログラムは、CD-ROM3501、またはFD3502に記憶されて、CD-ROMドライブ3412またはFDドライブ3411に挿入され、さらにハードディスク3417に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ341に送信され、ハードディスク3417に記憶されても良い。プログラムは実行の際にRAM3416にロードされる。プログラムは、CD-ROM3501、FD3502またはネットワークから直接、ロードされても良い。

40

【0235】

プログラムは、コンピュータ341に、上述した実施の形態の上位語取得装置等の機能を実行させるオペレーティングシステム(OS)、またはサードパーティプログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能(モジュール)を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム340がどのように動作するかは周知であり、詳細な説明は省略する。

50

【0236】

また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【0237】

また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【0238】

本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

10

【産業上の利用可能性】

【0239】

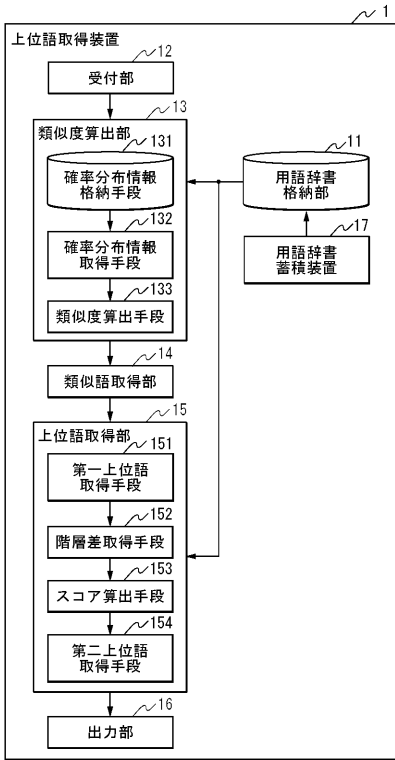
以上のように、本発明にかかるデータ作成装置は、大規模な上位語と下位語のデータベースを構築できるという効果を有し、データ作成装置等として有用である。

【符号の説明】

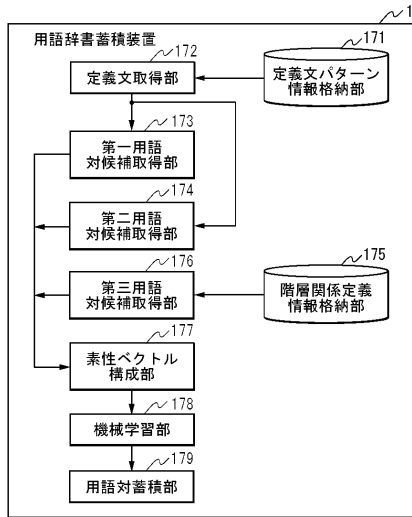
【0240】

- | | | |
|-----|--------------|----|
| 1 | 上位語取得装置 | |
| 2 | データ作成装置 | |
| 11 | 用語辞書格納部 | |
| 12 | 受付部 | 20 |
| 13 | 類似度算出部 | |
| 14 | 類似語取得部 | |
| 15 | 上位語取得部 | |
| 16 | 出力部 | |
| 17 | 用語辞書蓄積装置 | |
| 21 | 上位語生成部 | |
| 22 | 用語対情報受付部 | |
| 25 | 下位語付加部 | |
| 131 | 確率分布情報格納手段 | |
| 132 | 確率分布情報取得手段 | 30 |
| 133 | 類似度算出手段 | |
| 151 | 第一上位語取得手段 | |
| 152 | 階層差取得手段 | |
| 153 | スコア算出手段 | |
| 154 | 第二上位語取得手段 | |
| 171 | 定義文パターン情報格納部 | |
| 172 | 定義文取得部 | |
| 173 | 第一用語対候補取得部 | |
| 174 | 第二用語対候補取得部 | |
| 175 | 階層関係定義情報格納部 | 40 |
| 176 | 第三用語対候補取得部 | |
| 177 | 素性ベクトル構成部 | |
| 178 | 機械学習部 | |
| 179 | 用語対蓄積部 | |

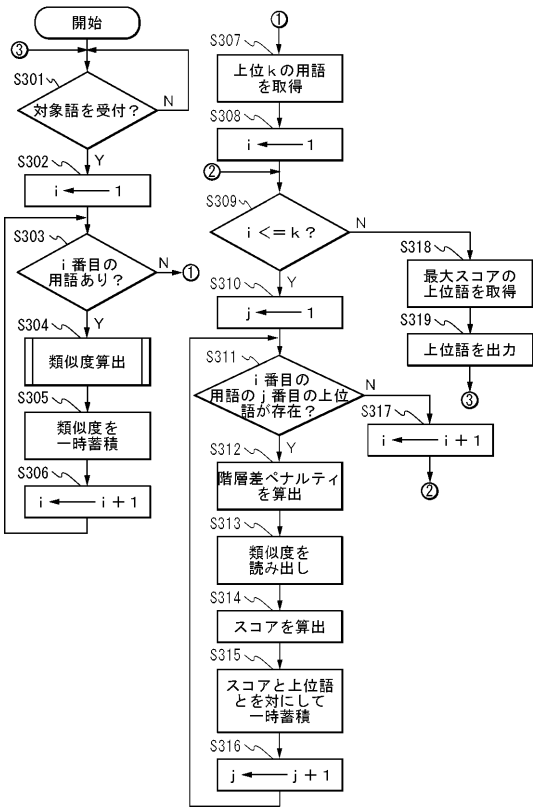
【 図 1 】



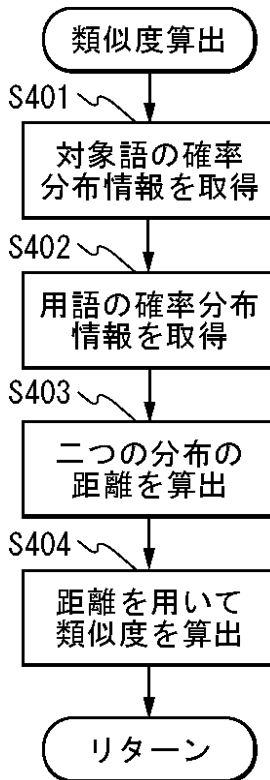
【 図 2 】



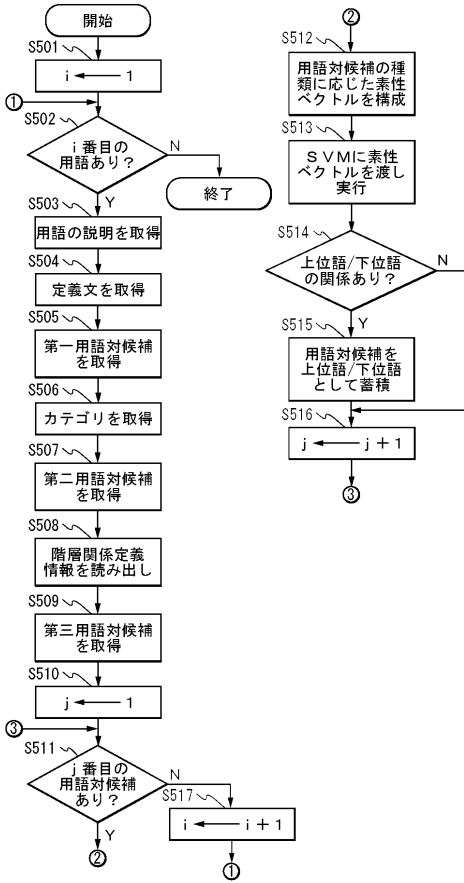
【 図 3 】



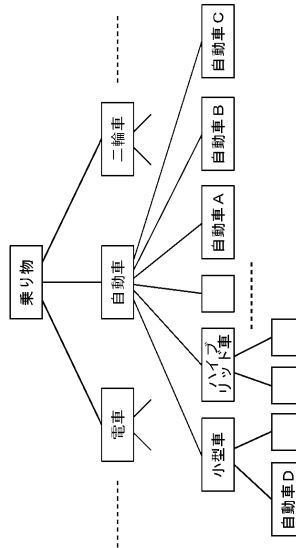
【 図 4 】



【図5】



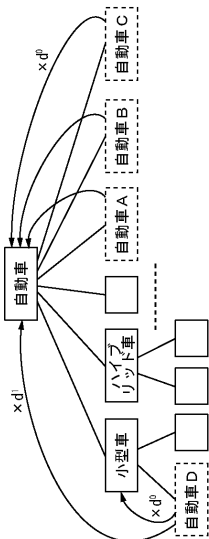
【図6】



【図7】

用語	確率分布情報
自動車	(0.1, 0.05, 0.0, 0.2, ……)
小型車	(0.08, 0.03, 0.02, 0.3, ……)
ハイブリッド車	(0.07, 0.04, 0.1, 0.15, ……)
⋮	⋮

【図8】



【図9】

	k-similar words (CVD)	k-similar words (RVD)	k-similar words (CVD, d=0)	Baseline approach 1 (CVD)	Baseline approach 2 (CVD)	Baseline approach 3 (CVD)
1,000	0.940	1.000	0.850	0.730	0.290	0.630
10,000	0.910	0.875	0.875	0.555	0.300	0.445
100,000	0.745	0.710	0.730	0.500	0.280	0.435
670,000	0.520	0.500	0.470	0.345	0.115	0.170

【 図 1 0 】

スコア	対象語	上位語
58.6	INDIVI	ブランド
54.3	クレオメ	花
34.4	UOKR	ゲーム
21.7	オーキッド	町
20.5	スマートフォツ	車
15.6	深川めし	料理
8.9	ジョン・パリー	作曲家
8.5	JVM	ソフトウェア
6.6	メタンガス	元素
5.4	メールセミナー	本
3.9	グロメツ	商品
3.1	スプリングバック	現象

【 図 1 1 】

紅茶

出典:フリー百科事典『ウィキペディア(Wikipedia)』

紅茶とは、摘み取った茶を乾燥させ、もみ込んで完全発酵させた茶葉。

主な紅茶ブランド [編集]

イギリス [編集]

- ・リプトン
- ・トワイニングス

フランス [編集]

- ・フォション

生産量 [編集]

1. インド
2. スリランカ

ブレンドティー [編集]

アールグレイティー
柑橘系の香りをつけた紅茶

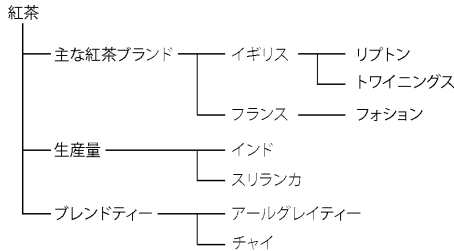
チャイ
インド式に甘く煮出したミルクティー

Category: 茶 | 喫茶文化

【 図 1 2 】

1	紅茶とは、摘み取った茶を乾燥させ、もみ込んで完全発酵させた茶葉。
2	= 主な紅茶ブランド =
3	== イギリス ==
4	* リプトン
5	* トワイニングス
6	== フランス ==
7	* フォション
8	== 生産量 ==
9	# インド
10	# スリランカ
11	= ブレンドティー =
12	;アールグレイティー:柑橘系の香りをつけた紅茶
13	;チャイ:インド式に甘く煮出したミルクティー
14	[[Category:茶]]
15	[[Category:喫茶文化]]

【 図 1 4 】



【 図 1 5 】

代表的な X, 代表 X, 主要な X, 主な X, 主要 X, 基本的な X, 基本 X, 著名な X, 大きな X, 他の X, 一部 X, 代表的 X, 基本的 X, 著名 X, 一部 X, X の一覧, X 一覧, X 詳細, X リスト, X の詳細

【 図 1 3 】

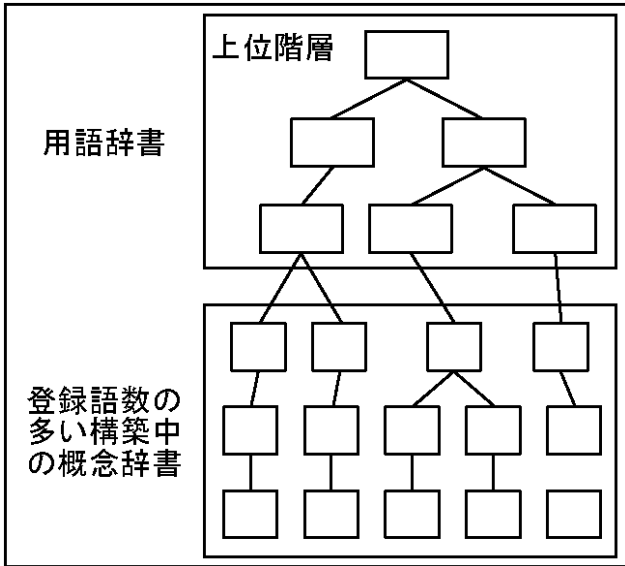
優先度	修飾記号の種類	記述方法	例
1	節見出し	=+ title =+	== イギリス ==
2	定義の箇条書き	;title:def.	;チャイ:ミルクティー
3	番号付き箇条書き	# + title	# インド
3	番号なし箇条書き	*+ title	*リプトン

注: title は、見出しを、+ は直前の記号が連続して出現しうることを示す。

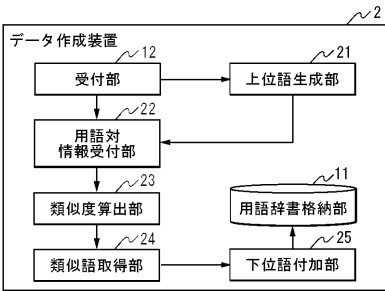
【 図 1 6 】

素性の種類	素性の発火条件
POS	上位語/下位語の末尾以外の形態素の品詞にXを含む 上位語/下位語の末尾の形態素の品詞がX
MORPH	上位語/下位語の末尾以外の形態素にXを含む 上位語/下位語の末尾の形態素がX
EXP	上位語/下位語がX
LCHAR	上位語と下位語の末尾の1文字が一致
ATTR	上位語/下位語が属性Xに一致
PAT	上位語が図15のパターンに一致
LAYER	階層構造で上位語/下位語に付与された修飾記号がX
DIST	階層構造で上位語と下位語の間の距離が2以上 階層構造で上位語と下位語の間の距離が1

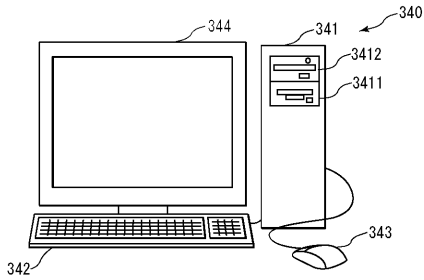
【図17】



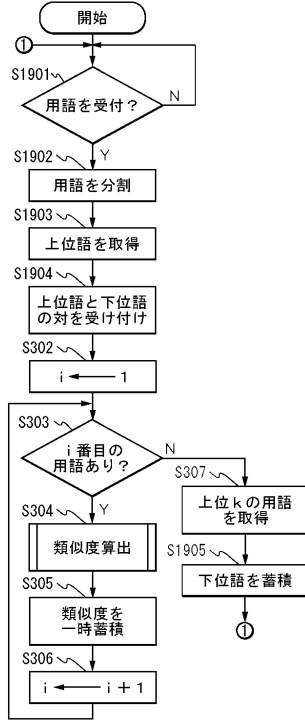
【図18】



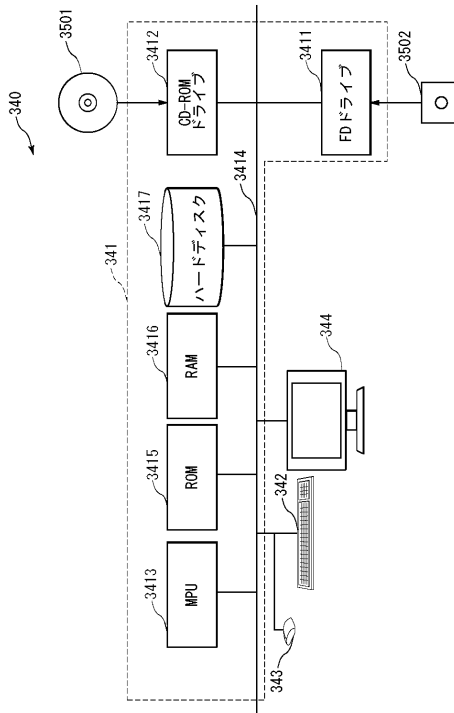
【図20】



【図19】



【図21】



フロントページの続き

- (72)発明者 黒田 航
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 村田 真樹
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 フランシス ボンド
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 ステイン デ サーガ
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 隅田 飛鳥
石川県鳳珠郡能登町字鶴川2 7 字 1 1 番地
- Fターム(参考) 5B075 ND03 ND35 NK43 QM08 UU01