(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5487781号
(P5487781)

(45)発行日　平成26年5月7日(2014.5.7)　(24)登録日　平成26年3月7日(2014.3.7)

(51)Int.Cl.　F I
*G06F 17/30 (2006.01)*　G06F 17/30 320D
G06F 17/30 210D
G06F 17/30 350C

請求項の数 10　(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)
(65)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)
審査請求日　平成24年6月18日(2012.6.18)

特許法第３０条第１項適用　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

(73)特許権者　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)【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または２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であって、
前記類似度算出部は、
用語と当該用語が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、
前記上位語の確率分布情報、および前記各用語の確率分布情報を用いて、前記上位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データ作成装置。

【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。
【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請求項１から請求項３いずれか記載のデータ作成装置。
【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。
【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、
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。
【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。
【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しており、かつ、
用語と当該用語が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または２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し、
前記類似度算出ステップは、
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上位語の確率分布情報、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ステップと、
前記上位語の確率分布情報、および前記各用語の確率分布情報を用いて、前記上位語と前記各用語の類似度を算出する類似度算出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、かつ、
用語と当該用語が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または２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って、
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上位語の確率分布情報、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、
前記上位語の確率分布情報、および前記各用語の確率分布情報を用いて、前記上位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、コンピュータを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＜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】

【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。

【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。

【００２５】

かかる構成により、大規模な用語辞書を構築できる。

【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
【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
【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】
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。
【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。
【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】

上位語取得装置１は、用語辞書格納部１１、受付部１２、類似度算出部１３、類似語取得部１４、上位語取得部１５、出力部１６、用語辞書蓄積装置１７を具備する。

【００３５】

類似度算出部１３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。

【００３６】

上位語取得部１５は、第一上位語取得手段１５１、階層差取得手段１５２、スコア算出手段１５３、第二上位語取得手段１５４を具備する。

【００３７】

用語辞書蓄積装置１７は、定義文パターン情報格納部１７１、定義文取得部１７２、第一用語対候補取得部１７３、第二用語対候補取得部１７４、階層関係定義情報格納部１７５、第三用語対候補取得部１７６、素性ベクトル構成部１７７、機械学習部１７８、用語対蓄積部１７９を具備する。

【００３８】

用語辞書格納部１１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１１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１１の用語辞書は、後述する用語辞書蓄積装置１７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１１で記憶されるようになってもよく、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１１で記憶されるようになってもよく、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１１で記憶されるようになってもよい。

【００３９】

受付部１２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１２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。

【００４０】

類似度算出部１３は、受付部１２が受け付けた対象語と、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。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１３は、通常、ＭＰＵやメモリ等から実現され得る。類似度算出部１３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。

【００４１】

確率分布情報格納手段１３１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１３１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１３１に確率分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。

【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。

【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。

【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、演算式の情報を予め格納している。

【数１】

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

【００４９】

数式１において、$n_{trg}$は、対象語である。Desc(nhyper)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($n_{hyper}$,$n_{hypo}$)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($n_{hyper}$ ,$n_{hypo}$)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。$d^{r(nhyper,nhypo)-1}$は、階層の深さに依存するペナルティである。sim($n_{trg}$,$n_{hypo}$)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。

【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。

【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ソフトまたは、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。

【００５３】

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。

【００５４】

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。

【００５５】

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。

【００５６】

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。

【００５７】

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ーンを予め保持している、とする。第七の素性「ＬＡＹＥＲ」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「ＤＩＳＴ」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「ＰＯＳ」「ＭＯＲＰＨ」「ＥＸＰ」「ＡＴＴＲ」「ＬＣＨＡＲ」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、上記８種類の素性を用いることは好適である。

【００６３】

素性ベクトル構成部１７７は、通常、ＭＰＵ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。

【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、ＭＰＵ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。

【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、ＭＰＵ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。

【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。

【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】

（ステップＳ３０５）類似度算出部１３は、ステップＳ３０４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Ｓ３０６）類似度算出部１３は、カウンタｉを１、インクリメントする。ステップＳ３０３に戻る。

【００７３】

（ステップＳ３０７）類似語取得部１４は、ステップＳ３０５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Ｓ３０８）上位語取得部１５は、カウンタｉに１を代入する。

【００７５】

（ステップＳ３０９）上位語取得部１５は、条件「ｉ＜＝ｋ」であるか否かを判断する。条件を満たせばステップＳ３１０に行き、条件を満たさなければステップＳ３１８に行く。

【００７６】

（ステップＳ３１０）上位語取得部１５は、カウンタｊに１を代入する。

【００７７】

（ステップＳ３１１）上位語取得部１５の第一上位語取得手段１５１は、用語辞書格納部１１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Ｓ３１２に行き、ｊ番目の上位語が存在しなければステップＳ３１７に行く。

【００７８】

（ステップＳ３１２）スコア算出手段１５３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（$\times \mathrm{sim}(n_{hyper},n_{hypo})$）を除いた部分である。なお、スコア算出手段１５３は、階層差取得手段１５２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Ｓ３１３）スコア算出手段１５３は、用語の類似度を読み出す。

【００８０】

（ステップＳ３１４）スコア算出手段１５３は、ステップＳ３１２で算出した階層差ペナルティと、ステップＳ３１３で良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１５３は、例えば、数式１により、スコアを算出する。

【００８１】

（ステップＳ３１５）第二上位語取得手段１５４は、ステップＳ３１４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Ｓ３１６）上位語取得部１５は、カウンタｊを１、インクリメントする。ステップＳ３１１に戻る。

【００８３】

（ステップＳ３１７）上位語取得部１５は、カウンタｉを１、インクリメントする。ステップＳ３０９に戻る。

【００８４】

（ステップＳ３１８）第二上位語取得手段１５４は、ステップＳ３１５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Ｓ３１９）出力部１６は、ステップＳ３１８で取得された上位語を出力する。ステップＳ３０１に戻る。

【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も良い。

【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。
【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。
【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。
【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。
【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。
【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】
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＜ｖ，ｒｅｌ＞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＜ｖ，ｒｅｌ＞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に、「名詞ｎ（用語）が＜ｖ，ｒｅｌ＞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１

３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。そのうちの一つが、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、図７に示す確率分布管理表から読み出す。

【０１２２】

次に、確率分布情報取得手段１３２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。

【０１２３】

次に、類似度算出手段１３３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。

【０１２４】

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１３３は、＜ｖ，ｒｅｌ，ｎ＞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１３３は、＜ｖ，ｒｅｌ，ｎ＞を２つの部分に分割する。第一は＜ｖ，ｒｅｌ＞である。第二は、＜ｎ＞である。次に、類似度算出手段１３３は、＜ｖ，ｒｅｌ＞の組の発生の条件付き確率「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）」を取得する。「Ｐ（＜ｖ，ｒｅｌ＞｜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１３３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る方法は種々ある。類似度算出手段１３３は、例えば、「Jensen-Shannon divergence」を用いる。「Jensen-Shannon divergence」は、２つの確率分布「Ｐ（・｜$n_1$）」「Ｐ（・｜$n_2$）」の距離「$D_{JS}$（Ｐ（・｜$n_1$）||Ｐ（・｜$n_2$）」を、以下の数式２のように算出する。

【数２】

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

【０１２５】

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】

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

【０１２６】

そして、類似度算出手段１３３は、例えば、「$D_{JS}$（Ｐ（・｜$n_1$）||Ｐ（・｜$n_2$）」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】

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

【０１２７】

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。

【０１２８】

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、１，０００，０００の名詞句と、１００，０００の動詞と助詞のセットを用いて、確率「Ｐ（＜ｖ，ｒｅｌ＞｜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（＜ｖ，ｒｅｌ＞｜ｎ）」は、以下の数式５により算出できた。

【数５】

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

【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（＜ｖ，ｒｅｌ＞｜ｎ）＝（ｆ（＜ｖ，ｒｅｌ，ｎ＞）＋１）／（Σｆ（＜ｖ，ｒｅｌ，ｎ＞）＋１）でも良い。

【０１３１】

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

【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、＜ｖ，ｒｅｌ，ｎ＞の組の出現確率は、以下の数式６で示される。

【数６】

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

【０１３３】

数式６において、「ａ」は＜ｖ，ｒｅｌ＞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（＜ｖ，ｒｅｌ＞｜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（＜ｖ，ｒｅｌ＞｜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（＜ｖ，ｒｅｌ＞｜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（＜ｖ，ｒｅｌ＞｜ａ）」、「Ｐ（ｎ｜ａ）」および「Ｐ（ａ）」が更新される。

【０１３５】

以上の処理により、各＜ｖ，ｒｅｌ＞、ｎ、およびａに対して、確率「Ｐ（＜ｖ，ｒｅ

ｌ＞｜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】

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

【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＜ｖ，ｒｅｌ＞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。

【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。

【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。

【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。

（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θ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$P_{child}(\cdot \mid n_{hyper})$が、以下の数式８により定義される。

【数８】

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

【０１５０】
数式８において、Ｃｈ（$n_{hyper}$）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】

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３００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。

【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。

【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。

【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ため、より多様な上位下位関係が獲得できる。

【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。

【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。

【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。

【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。

【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。

【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。

【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。

【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。

【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。

【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を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。

【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。

【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。

【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で取得さえた上位語と、ステップＳ１９０１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Ｓ１９０５）下位語付加部２５は、ステップＳ３０７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Ｓ１９０１に戻る。
【０２０１】
なお、図１９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１１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２３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１２は、上位語下位語対「ＰＣソフトーぷよぷよ通」を受け付けた、とする。そして、上位語生成部２１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２１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２２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２３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１１から取得する、とする。ここでは、類似度算出部２３は、用語辞書格納部１１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２３は、用語辞書格納部１１から、３つの用語「ＩＤ＝０６５６６０７７，ソフト」、「ＩＤ＝０３３２５９４１，ソフト」、および「ＩＤ＝０７６１４５００，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＝０６５６６０７７，ソフト」は、同義語として（ソフト，ソフトウェア，ソフトウエア，パッケージ）を有する。「ＩＤ＝０３３２５９４１，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＝０７６１４５００，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２３は、「ＩＤ＝０６５６６０７７，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２．４５５６７１９２と算出した、とする。なお、類似度算出部２３は、類似度算出部１３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２３は、「ＩＤ＝０３３２５９４１，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２３は、「ＩＤ＝０７６１４５００，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２４は、最も類似度が大きい、「ＩＤ＝０６５６６０７７，ソフト」を取得する。

【０２１１】

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ソフトーぷよぷよ通」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コードーマイクロ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。

【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。

【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コードーマイクロ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。

（具体例３）

【０２１７】

今、受付部１２は、「ＳＦ映画ーアルマゲドン」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（ドローイング，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，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、類似度が最大の用語「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）」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用語として蓄積した。つまり、親の用語は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】

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。

【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。

【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。

【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。

【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ープログラム等は、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。

【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。

【産業上の利用可能性】

【０２３９】

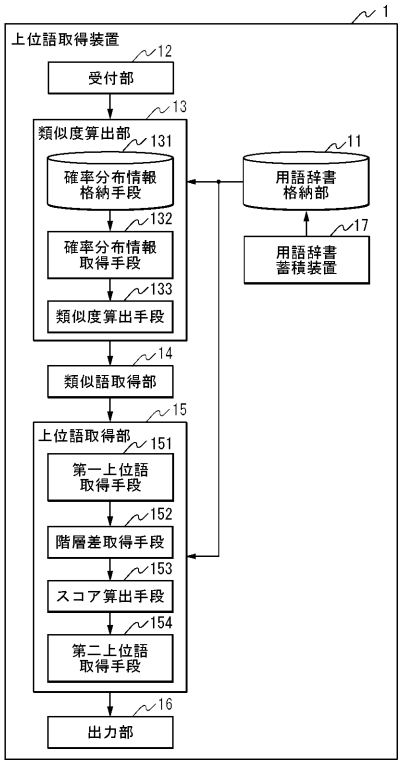
以上のように、本発明にかかるデータ作成装置は、大規模な上位語と下位語のデータベースを構築できるという効果を有し、データ作成装置等として有用である。

【符号の説明】

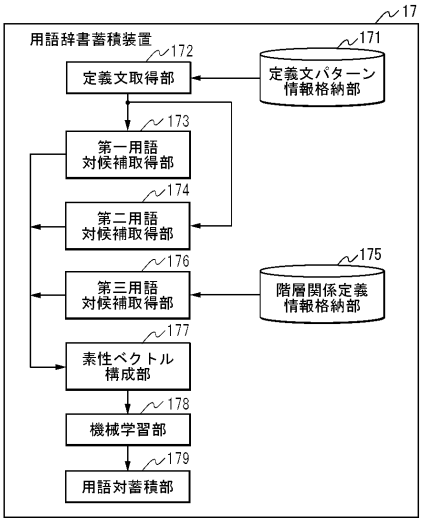
【０２４０】

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
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
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
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

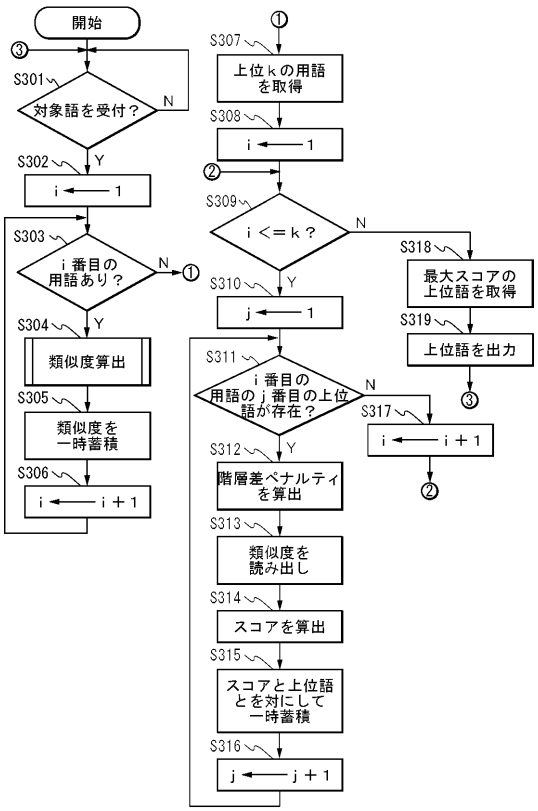
【図１】


1
上位語取得装置
12
受付部
13
類似度算出部
131
確率分布情報格納手段
132
確率分布情報取得手段
133
類似度算出手段
11
用語辞書格納部
17
用語辞書蓄積装置
14
類似語取得部
15
上位語取得部
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
16
出力部

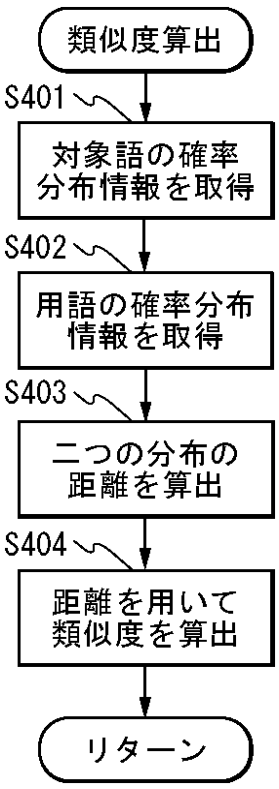
【図２】


17
用語辞書蓄積装置
172
定義文取得部
171
定義文パターン情報格納部
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
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

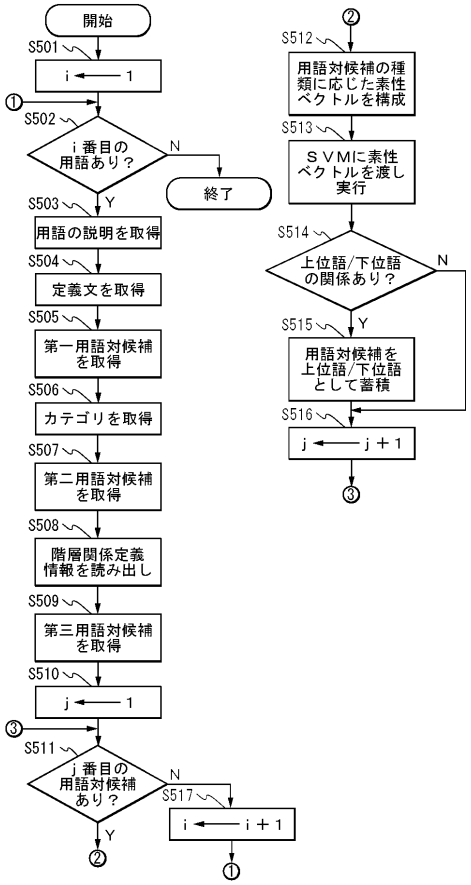
【図３】


開始
S301
対象語を受付？
N
Y
S302
i ← 1
S303
i 番目の用語あり？
N
①
Y
S304
類似度算出
S305
類似度を一時蓄積
S306
i ← i + 1
③
①
S307
上位 k の用語を取得
S308
i ← 1
②
S309
i <= k ？
N
Y
S310
j ← 1
S311
i 番目の用語の j 番目の上位語が存在？
N
Y
S312
階層差ペナルティを算出
S313
類似度を読み出し
S314
スコアを算出
S315
スコアと上位語とを対にして一時蓄積
S316
j ← j + 1
S317
i ← i + 1
②
S318
最大スコアの上位語を取得
S319
上位語を出力
③

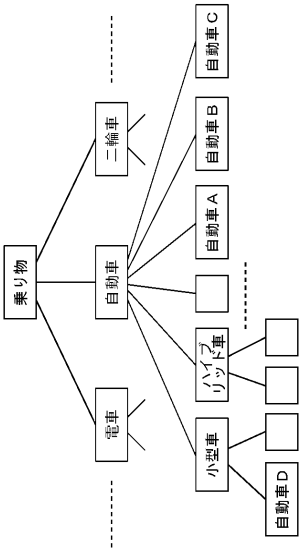
【図４】


類似度算出
S401
対象語の確率分布情報を取得
S402
用語の確率分布情報を取得
S403
二つの分布の距離を算出
S404
距離を用いて類似度を算出
リターン

【図５】



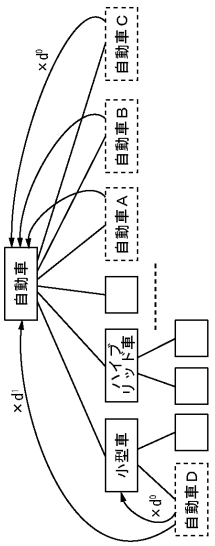

【図６】




【図７】

| 用語 | 確率分布情報 |
|---|---|
| 自動車 | (0.1, 0.05, 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１０】

| 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・バリー | 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 | グロメット | 商品 |
| 3.1 | スプリングバック | 現象 |

【図１１】

**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１２】

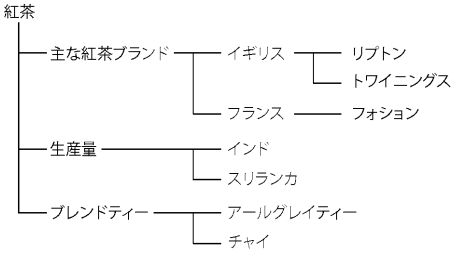
```
1  紅茶とは、摘み取った茶を乾燥させ、もみ込んで
   完全発酵させた茶葉。
2  = 主な紅茶ブランド =
3  == イギリス ==
4  * リプトン
5  * トワイニングス
6  == フランス ==
7  * フォション
8  == 生産量 ==
9  # インド
10 # スリランカ
11 = ブレンドティー =
12 ;アールグレイ ティー : 柑橘系の香りをつけた紅茶
13 ;チャイ : インド式に甘く煮出したミルクティー
14 [ [ Category: 茶 ] ]
15 [ [ Category: 喫茶文化 ] ]
```

【図１３】

| 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１４】



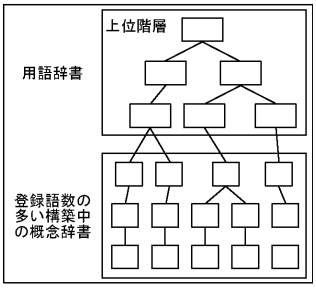

【図１５】

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

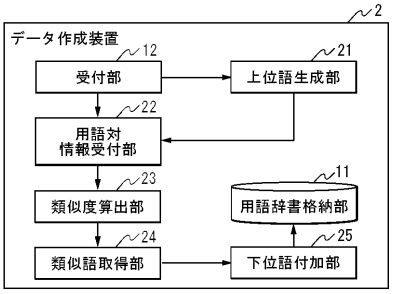
【図１６】

| 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<br>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<br>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<br>階層構造で上位語と下位語の間の距離が１ |

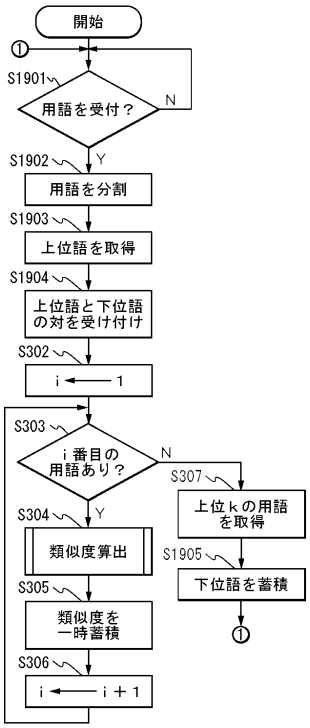
【図１７】



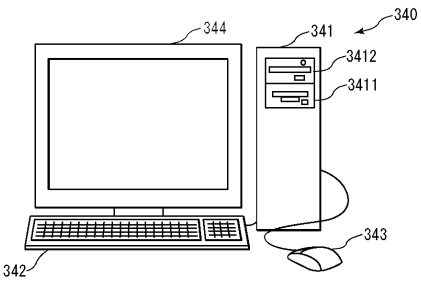

【図１８】



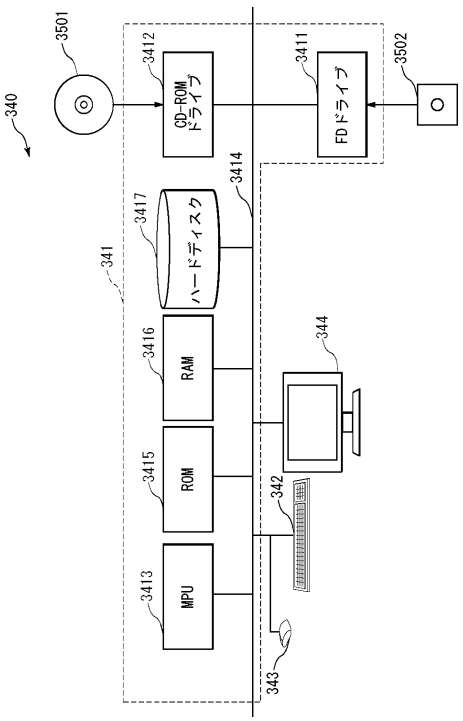

【図１９】




【図２０】




【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鵜川２７字１１番地

審査官　久々宇　篤志

(56)参考文献　特開２００７－０１１７７５（ＪＰ，Ａ）
正津　康弘　他，国語辞典とシソーラスの統合，情報処理学会研究報告，日本，社団法人情報処理学会，２００３年　1月２１日，第2003巻　第4号，pp.141-146
隅田　飛鳥　他，Ｗｉｋｉｐｅｄｉａの記事構造からの上位下位関係抽出，自然言語処理，日本，言語処理学会，２００９年　7月１０日，第16巻　第3号，pp.3-24

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　１７／３０