(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5936698号
(P5936698)
(45)発行日　平成28年6月22日(2016.6.22)　(24)登録日　平成28年5月20日(2016.5.20)

(51)Int.Cl.　F I
G06F 17/27 (2006.01)　G06F 17/27 695
G06F 17/27 615

請求項の数 5　(全 23 頁)

(21)出願番号　特願2014-532583(P2014-532583)
(86)(22)出願日　平成24年8月27日(2012.8.27)
(86)国際出願番号　PCT/JP2012/071535
(87)国際公開番号　WO2014/033799
(87)国際公開日　平成26年3月6日(2014.3.6)
審査請求日　平成26年11月11日(2014.11.11)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100100310
弁理士　井上　学
(74)代理人　100098660
弁理士　戸田　裕二
(74)代理人　100091720
弁理士　岩崎　重美
(72)発明者　森本　康嗣
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内

審査官　長　由紀子

最終頁に続く

(54)【発明の名称】単語意味関係抽出装置

(57)【特許請求の範囲】
【請求項１】
テキストから抽出した単語の組に対してそれぞれ異なる複数種類の方式で求められる特徴量を要素とする素性ベクトルを生成する手段と、
既知の辞書を参照し、前記素性ベクトルに対して単語意味関係を示すラベルを付与する手段と、
前記ラベルが付与された複数の素性ベクトルに基づいて単語意味関係を識別するために用いる単語意味関係識別用データを多カテゴリの識別問題として学習する手段と、
前記学習した単語意味関係識別用データに基づいて、任意の単語の組に対して単語意味関係を識別する手段とを備え、
前記素性ベクトルを生成する手段は、
注目する単語の前記テキスト中における出現箇所の近傍の単語を当該注目する単語の文脈情報として抽出する手段と、
前記単語の組の特徴量として、当該単語の組の２つの単語の文脈情報同士の類似度であって、単語の組の一方を基準に計算した類似度と他方を基準にして計算した類似度の２種類の特徴量を計算する手段と、を備えることを特徴とする単語意味関係抽出装置。
【請求項２】
テキストから抽出した単語の組に対してそれぞれ異なる複数種類の方式で求められる特徴量を要素とする素性ベクトルを生成する手段と、
既知の辞書を参照し、前記素性ベクトルに対して単語意味関係を示すラベルを付与する

手段と、
前記ラベルが付与された複数の素性ベクトルに基づいて単語意味関係を識別するために用いる単語意味関係識別用データを多カテゴリの識別問題として学習する手段と、
前記学習した単語意味関係識別用データに基づいて、任意の単語の組に対して単語意味関係を識別する手段とを備え、
前記素性ベクトルを生成する手段は、
前記単語の組の２つの単語に含まれる文字同士の対応関係を同一の文字であるか、文字の意味が類似しているかどうかに基づいて計算する手段と、
前記単語の組の特徴量として、前記文字同士の対応関係に基づいた類似度であって、単語の組の一方を基準に計算した類似度と他方を基準にして計算した類似度の２種類の特徴量を計算する手段と、を備えることを特徴とする単語意味関係抽出装置。
【請求項３】
テキストから抽出した単語の組に対してそれぞれ異なる複数種類の方式で求められる特徴量を要素とする素性ベクトルを生成する手段と、
既知の辞書を参照し、前記素性ベクトルに対して単語意味関係を示すラベルを付与する手段と、
前記ラベルが付与された複数の素性ベクトルに基づいて単語意味関係を識別するために用いる単語意味関係識別用データを多カテゴリの識別問題として学習する手段と、
前記学習した単語意味関係識別用データに基づいて、任意の単語の組に対して単語意味関係を識別する手段とを備え、
前記素性ベクトルを生成する手段は、
予め格納された、単語間の関係を示すパターンによって単語の組を抽出する手段と、
前記単語の組の特徴量として、前記単語の組の一方を基準に計算した前記単語の組の頻度に基づいた統計量と他方を基準に計算した前記単語の組の頻度に基づいた統計量の２種類の特徴量を計算する手段と、を備えることを特徴とする単語意味関係抽出装置。
【請求項４】
請求項１乃至３に記載の単語意味関係抽出装置であって、
前記単語意味関係は、前記単語の組を構成する２つの単語が、同義語であるか、上位・下位語であるか、対義語であるか、兄弟語であるか、あるいはそれらの何れでもないか、であることを特徴とする単語意味関係抽出装置。
【請求項５】
請求項１乃至３に記載の単語意味関係抽出装置であって、
前記単語の組を構成する２つの単語が固有名詞の場合であって、前記２つの単語が同じものを示さないときには、前記２つの単語を同義語でないと判定する手段を備えることを特徴とする単語意味関係抽出装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、テキスト中から、単語間の意味的な関係を抽出する技術に関する。
【背景技術】
【０００２】
パソコン及びインターネットの普及によって、ユーザがアクセス可能な電子化文書の量が増大している。このような大規模な文書情報の中から、所望の文書を効率的に発見するための技術が求められている。文書検索技術に代表される、自然言語を扱う技術では、言語の曖昧性、すなわち多義性と同義性を適切に扱うことが必要である。多義性は、同じ単語に対し複数の意味が存在することであり、ノイズの原因となる。一方、同義性は、同じ意味を持つ単語が複数存在することであり、漏れの原因となる。業務向けの応用では、特に漏れ、すなわち情報の見落としが問題となることが多いため、同義性の問題を解決することが重要である。
【０００３】

同義語辞書やシソーラスは、文書における言語表現の揺れを吸収し、同義性の問題を解決するための言語資源であり、各種の言語処理アプリケーションで使用される。価値が高いデータであることから、人手によって多くの辞書が古くから編纂されている。

【０００４】

同義語辞書やシソーラスの人手作成には大きなコストが必要であるため、同義語辞書やシソーラスをテキストデータから自動で作成することが従来から試みられている。同義語辞書やシソーラスを自動作成するための方法の一つとして、単語の出現文脈、すなわち着目している単語の近傍に現れる単語や文字列に着目する方法がある。非特許文献１に、出現文脈に基づく文脈ベース同義語抽出技術が開示されている。また、同義語の中で特に表記揺れを扱うための方法がある。非特許文献２に、発音に関する規則に基づいて、カタカナ表記の表記揺れを検出する表記ベース同義語抽出技術が開示されている。また、「ＡやＢなどのＣ」のような単語間の関係を明示的に示すパターンを用いる同義語抽出技術も存在する。非特許文献３には、パターンを用いることによるパターンベース同義語抽出技術が開示されている。

【０００５】

以上の同義語抽出技術は、教師なし学習、すなわち人手によって付与された正解を用いないタイプの学習技術によっている。教師なし学習では正解を作成する必要がないため、人手のコストが低いことが利点である。しかしながら、現在では人手で作成された大規模な辞書が広く利用可能となっており、これらを正解として用いることが可能となり、教師なし学習のメリットは少なくなっている。一方、教師あり学習では、人手による正解データを用いることで高い精度を得ることが可能である。

【０００６】

以上のような状況のもと、教師あり学習による同義語抽出方法が非特許文献５に開示されている。非特許文献５では、人手によって作成された同義語辞書を正解として、教師あり学習によって同義語抽出を行う。具体的には、後述する単語の文脈に基づいて単語の意味を表現し、正解である同義語辞書を用いることによって学習を行い、同義語を抽出する。

【０００７】

以上の先行技術は、同義語抽出技術に関するものであるが、シソーラスでは同義語以外の単語意味関係として、上位・下位語関係、対義語関係、兄弟語関係、部分・全体語関係などが定義されている。このような、同義語以外の関係を抽出する技術も存在する。特許文献１非特許文献６には、既存のシソーラスと文脈ベースの単語間類似度によって上位・下位語を抽出する技術が開示されている。また、非特許文献４には、単語の包含関係に基づいて単語の上位・下位語関係を抽出する技術が開示されている。

【０００８】

これらの単語意味関係は、部分・全体語を除き、同義語、上位・下位語、対義語、兄弟語のいずれも意味が類似しているという点が共通している。これらの単語意味関係を総称して類似語と呼ぶことにする。類似語中の特定種類の単語意味関係を抽出しようとすると、それ以外の種別の単語意味関係が誤って抽出され易い。例えば、同義語抽出をする際に、上位・下位語、対義語、兄弟語が誤って同義語として抽出される。そのため、このような類似度内のより詳細な単語意味関係の種別を詳細にする技術が提案されている。非特許文献７には、同義語抽出の際、パターンベースの方法で対義語を抽出する技術を用いることで、同義語を高精度に抽出する技術が開示されている。また、特許文献１には、教師ありのランキング学習によって、同義語とそれ以外の類似語、非類似語を区別する技術が開示されている。

【先行技術文献】

【特許文献】

【０００９】

【特許文献１】特開２０１１－１１８５２６号公報

【非特許文献】

【００１０】
【非特許文献１】相澤、「大規模テキストコーパスを用いた語の類似度計算に関する考察」、情報処理学会論文誌、vol．49-3、 pp．1426-1436 (2008).
【非特許文献２】久保田他、「カタカナ表記の統一方式　予備分類とグラフ比較によるカタカナ表記のゆらぎ検出法」、情報処理学会自然言語処理研究会報告、NL97-16、pp.111-117、1993.
【非特許文献３】M. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)、 pp. 539-545、 1992.
【非特許文献４】小山、竹内、 「日本語複合語用語の入れ子関係に基づく階層的体系化、 情報処理学会自然言語処理研究会報告」、 NL-180、 pp.49-54、 2007.
【非特許文献５】Masato Hagiwara: A Supervised Learning Approach to Automatic Synonym Identification based on Distributional Features、 Proc. of ACL 2008 Student Research Workshop、 pp. 1-6、 2008.
【非特許文献６】松本，須藤，中山，平尾: 複数の言語資源からのシソーラスの構築，情報処理学会情報学基礎研究会報告，FI42-4,pp.23-28，1996.
【非特許文献７】D. Lin、 S. Zhao、 L Qin、 and M. Zhou: “Identifying synonyms among distributionally similar words”、 IJCAI 2003、 pp. 1492-1493、 2003.
【発明の概要】
【発明が解決しようとする課題】
【００１１】
本発明の目的は、類似語内の詳細な単語意味関係の種別を、従来よりも高精度に区別して抽出することができる、単語意味関係抽出技術を実現することである。非特許文献７のような、教師なし学習のアプローチでは、人手作成のシソーラスを正解データとして用いることができないため、高い精度を達成することが困難である。一方で、教師あり学習を用いたアプローチでは、同義語、上位・下位語、対義語、兄弟語のような複数種類の単語意味関係の種別を任意の詳細さで判定する技術は存在しない。
【００１２】
例えば、非特許文献５に開示されている同義語抽出技術では、同義語か否かを判定する２値の識別問題として同義語抽出を解決するが、同義語以外の単語意味関係を抽出することはできない。同義語以外の類似語は、識別器が正しく動作して非類似語と認識されるか、誤って同義語として認識されるかのいずれかである。
【００１３】
また、特許文献１に開示されている単語意味関係抽出技術では、問題をランキング問題として扱うことで同義語とそれ以外の類似語を区別して扱おうとする。すなわち、同義語の場合は非常に類似性が高いということでランクとして１を付与し、上位・下位語や兄弟語の場合は同義語ほどではないが、ある程度類似性が高いということでランクとして２を付与し、そのいずれでもない場合には、類似性が低いということでランクとして３を付与する問題だと考える。しかしながら、特許文献１に開示されている方法でも、同義語以外の類似語をより詳細に、上位・下位語、兄弟語のように区別することはできない。
【００１４】
本発明は、以上の課題を解決するためになされたものであり、シソーラスを正解として活用して高精度な処理を実現すると同時に、複数種類の単語意味関係を詳細に抽出することができる単語意味関係抽出方式を提供することを目的とする。
【課題を解決するための手段】
【００１５】
本願において開示される発明のうち、代表的なものの概要を簡単に説明すれば、下記の通りである。
【００１６】
テキストから抽出した単語の組に対してそれぞれ異なる複数種類の類似度を要素とする

素性ベクトルを生成する手段と、既知の辞書を参照し、素性ベクトルに対して単語意味関係を示すラベルを付与する手段と、ラベルが付与された複数の素性ベクトルに基づいて単語意味関係を識別するために用いる単語意味関係識別用データを多カテゴリの識別問題として学習する手段と、学習した単語意味関係識別用データに基づいて、任意の単語の組に対して単語意味関係を識別する手段と、を備えることを特徴とする単語意味関係抽出装置である。

【発明の効果】

【００１７】

本発明によれば、高精度な単語意味関係抽出を行うことが可能となる。

【００１８】

上記した以外の課題、構成及び効果は、以下の実施形態の説明より明らかにされる。

【図面の簡単な説明】

【００１９】

【図１】計算機システムの構成例を示すブロック図である。

【図２】計算機システムにおける処理フローの説明図である。

【図３】類似度行列の説明図である。

【図４】教師なし学習による類似語抽出の概念的な説明図である。

【図５】２値の教師あり学習による類似語抽出の概念的な説明図である。

【図６】教師ありのランキング学習による類似語抽出の概念的な説明図である。

【図７】多クラスの教師あり学習による類似語抽出の概念的な説明図である。

【図８】単語意味関係抽出処理のフローチャートである。

【図９】シソーラスの説明図である。

【図１０】文脈行列の説明図である。

【図１１】文字重複度計算処理のフローチャートである。

【図１２】文字類似度計算処理のフローチャートである。

【図１３】文字類似度テーブルの説明図である。

【図１４】本発明の実施形態における、コンテンツクラウドシステムの一実現例を示す図である。

【発明を実施するための形態】

【００２０】

以下、図面を参照して本発明の実施の形態を説明する。

【実施例１】

【００２１】

まず、単語意味関係について説明する。単語意味関係としては、様々なものが存在するが、シソーラスを規定する規格として、ISO 2788 “Guidelines for the establishment and development of monolingual thesauri”やANSI/NISO Z39.19-2005 “Guidelines for the Construction、 Format、 and Management of Monolingual Controlled Vocabularies”が存在し、その中で下記のような種類が規定されている。

（１）同義語：同じ意味を持つ単語であり、テキスト中での置き換えが可能である単語のペア。「コンピュータ」と「電子計算機」など。

（２）上位／下位語：一方が他方の上位概念であるような単語ペア。「コンピュータ」と「サーバ」など。

（３）部分／全体語：一方が他方の一部であるような単語ペア。「帽子」と「つば」など。

（４）対義語：対となる概念を示す単語ペア。「男」と「女」など。

（５）兄弟語：同義ではないが、共通の上位概念を持つ単語ペア。「ルータ」と「サーバ」など。

（６）関連語：類似しておらず、階層的でもないが、概念的に連想される単語ペア。「細胞」と「細胞学」など。

【００２２】

また、同義語、上位・下位語、対義語、兄弟語のいずれも意味が類似しているという点が共通しているため、本明細書においては、これらの単語意味関係を総称して類似語と呼ぶ。

【００２３】

第１の実施の形態として、複数種類の単語意味関係を同時に抽出する単語意味関係抽出装置について説明する。図１は、本実施形態を実現する計算機システムの構成例を示すブロック図である。図１に示した計算機システムは、本発明の第１の実施の形態に用いられる。なお、実施の形態によっては使用されない機能も含んでいる。

【００２４】

単語意味関係抽出装置１００は、ＣＰＵ１０１、主メモリ１０２、入出力装置１０３及びディスク装置１１０を備える。ＣＰＵ１０１は、主メモリ１０２に記憶されるプログラムを実行することによって各種処理を行う。具体的には、ＣＰＵ１０１は、ディスク装置１１０に記憶されるプログラムを、主メモリ１０２上に呼び出して実行する。主メモリ１０２は、ＣＰＵ１０１によって実行されるプログラム及びＣＰＵ１０１によって必要とされる情報などを記憶する。入出力装置１０３には、ユーザから情報が入力される。また、入出力装置１０３は、ＣＰＵ１０１の指示に応じて、情報を出力する。例えば、入出力装置１０３は、キーボード、マウス及びディスプレイのうち少なくとも一つを含む。

【００２５】

ディスク装置１１０は、各種情報を記憶する。具体的には、ディスク装置１１０は、ＯＳ１１１、単語意味関係抽出プログラム１１２、テキスト１１３、シソーラス１１４、類似度行列１１５、文脈行列１１６、品詞パターン１１７、共起類似度テーブル１１８、識別モデル１１８、文字類似度テーブル１２０を記憶する。

【００２６】

ＯＳ１１１は、単語意味関係抽出装置１００の処理の全体を制御する。

【００２７】

単語意味関係抽出プログラム１１２は、テキスト１１３及びシソーラス１１４から単語意味関係を抽出するプログラムであり、素性ベクトル抽出サブプログラム１１２１、正解ラベル設定サブプログラム１１２２、識別モデル学習サブプログラム１１２３、識別モデル適用サブプログラム１１２４からなる。

【００２８】

テキスト１１３は、単語意味関係抽出プログラム１１２への入力となるテキストであり、特別な形式である必要はない。ＨＴＭＬ文書、ＸＭＬ文書などのタグを含む文書の場合は、タグを除去する前処理を施すことが望ましいが、タグが含まれた状態でも処理は可能である。

【００２９】

シソーラス１１４は、人手によって作成された同義語、上位・下位語、兄弟語が格納された辞書である。

【００３０】

類似度行列１１５は、テキスト及び同義語辞書から抽出された単語ペアに関する素性ベクトル、同義語かどうかを示すラベルなどを格納した行列である。文脈行列１１６は、文脈ベース類似度を計算するために必要な単語の文脈情報を格納した行列である。識別モデル１１８は、類似度行列から学習された、単語ペアが同義語であるかどうかを識別するためのモデルである。識別モデル１１８は、類似度行列から学習された、単語ペアが何れの単語意味関係に属するかを識別するためのモデルである。文字類似度テーブル１１９は、意味が類似した文字間の関係を格納するテーブルである。

【００３１】

処理の流れは、図２に示すようになる。素性ベクトル抽出サブプログラム１１２１は、テキスト１１３を読み込んでテキスト中の全ての単語を抽出し、任意の単語の組に対して各種の類似度を計算し、類似度行列１１５として出力する。その際に必要な情報である文脈行列１１６を事前に作成しておく。品詞パターン１１７は、文脈行列１１６の作成に用

いられる。なお、第１の実施の形態では、正解ラベル設定サブプログラム１１２２は、シソーラス１１４を正解データとして読み込み、類似度行列１１５中の各単語ペアに正解、各種の単語意味関係の種別を示すラベルを設定する。識別モデル学習サブプログラム１１２３は、類似度行列１１５を読み込み、単語ペアの単語意味関係種別を識別するための識別モデル１１８を学習する。識別モデル適用サブプログラム１１２４は、識別モデル１１８を読み込み、類似度行列１１５中の単語ペアに対し、単語意味関係種別の判定結果を付与する。

【００３２】

以下では、図３に示す類似度行列の例を用いて本実施形態の基本的な考え方を説明する。

【００３３】

テキストデータ中に含まれる、任意の単語のペアを考える。例えば、単語のペアを＜コンピュータ、計算機＞とする。このとき、単語ペアがどのような単語意味関係を持つかを判定するための様々な尺度を想定することができる。

【００３４】

例えば、単語の出現文脈間の類似度（以下、文脈ベース類似度と呼ぶ）を用いる方法がある。また、重複する文字数に着目するなど表記に基づいた類似度（以下、表記ベース類似度と呼ぶ）が考えられる。さらに、語彙統語パターンと呼ばれるパターン（以下、パターンベース類似度と呼ぶ）を用いることも可能である。

【００３５】

さらに、各手法において、様々なバリエーションが存在する。例えば、文脈ベース類似度において、単語の出現文脈をどのように定義するか、あるいは距離の計算方法をどのように定義するかによってバリエーションが存在する。本実施形態では、このような様々な尺度を、単語ペアの素性であると考え、単語ペアを素性毎の値からなる素性ベクトルで表現する。各単語関係種別に適した素性の構成方法については後述する。図３の例では、例えば、＜コンピュータ、コンピューター＞という単語ペアは、素性１の次元の値が０．３、素性２の次元の値が０．２、素性Ｎの次元の値が０．８であるベクトルで表現されている。ここで、素性１は、例えば、文脈類似度によるスコアであり、素性２は、表記ベース類似度によるスコアである。

【００３６】

以上のように、単語ペアを様々な尺度によるスコアでベクトル表現した後、各単語ペアがどのような単語意味関係であるかを、シソーラスを用いて判断し、ラベル付けを行う。すなわち、＜コンピュータ、計算機＞がシソーラス中で同義語であれば類似度行列に同義語に相当するラベルを付与し、＜コンピュータ、パソコン＞が上位・下位語であれば、上位・下位語に相当するラベルを付与する。類似語でない場合は、非類似語であるというラベルを付与する。なお、類似語内の単語意味関係の内、上位・下位語のみが方向を持ち、それ以外は方向を持たない。方向を持たない関係については、順序の異なる単語ペア、例えば、＜コンピュータ、計算機＞と＜計算機、コンピュータ＞を区別する必要がない。そのため、以下では、単語ペアは文字の昇順に単語を並べて、両者を同一のものとして扱うこととし、上位・下位語については関係の方向を考慮して、左側の単語が上位語の場合は上位・下位語、左側の単語が下位語の場合は下位・上位語と呼ぶことにする。図３の例では、同義語の場合のラベルが１、下位・上位語のラベルが２、上位・下位語のラベルが３、対義語のラベルが４、兄弟語のラベルが５、非類似語のラベルが－１、未知の単語ペアのラベルが０となっている。

【００３７】

以上のように、単語ペアを素性の値のベクトルで表現し、さらに正解データを付与することにより、多クラス（カテゴリ）の識別問題として解くことが可能となる。多クラスの識別問題とは、未知の事例が３個以上のクラスのいずれに属するかを識別するタスクであり、教師あり学習によって識別のモデルを学習する方法が知られている。同義語、上位・下位語、対義語、兄弟語等の単語意味関係種別は排他的であり、同時に複数のカテゴリに

属することは、単語が多義語である場合以外には原則的にはない。そのため、単語意味関係種別を多クラスの識別問題として解くことにより、類似語内の詳細な単語意味関係の種別が区別できるだけではなく、各単語意味関係、例えば同義語の抽出精度を向上することが可能となる。以上が本実施形態の基本的な考え方である。

【００３８】

以下では、単語意味関係毎にどのような尺度が有効であるかを説明する。

（１）上位・下位語

（ａ）文脈ベース方式

単純な文脈ベース方式では、ある単語ペアに関する類似度がスカラ値で与えられ、数値が大きい場合には（狭義の）同義語、中程度以下の場合には、同義語以外の類似語のいずれかだと考える。よって、上位・下位語、対義語、兄弟語の区別を行うことが困難である。

【００３９】

本実施形態では、非対称なスコアそれぞれを素性として用いることで教師あり学習を行う。非対称な２種類のスコアを素性として用いると、両方のスコアが高い場合は同義語、一方が他方よりも高い場合は上位・下位語、両方が中程度に高い場合には兄弟語といったように、境界を設定することが可能となる。

【００４０】

非対称な類似度とは、単語ペア<A、B>があるときに、単語Aを基準とした場合の単語Bに対する値と、Bを基準とした場合のAに対する値が異なるような類似度を言う。例えば、単純な例として、単語ペア<A、B>に対して、共通する文脈語の個数を類似度とする場合を考える。この場合、AとBのどちらを基準にしても、値は変わらないため、この類似度は対称である。一方、この値に基づいて、以下のように非対称な類似度を構成することができる。Aを基準として類似した単語のランキングを生成し、そのランキング中でBが何位にランクされるかを考える。このランクの逆数を類似度と考えると、Aを基準とした場合と、Bを基準とした場合では、値が異なる。例えば、「メーカ」と「電機メーカ」のような上位・下位語を考えた場合、「メーカ」を基準にすると、「商社」のような語が類似する語として抽出されるが、「電機メーカ」に対してはそうではない。一般に、上位語の方が類似する語の種類が多いため、上位語「メーカ」に関する「電機メーカ」のランクの方が、下位語「電機メーカ」に関する「メーカ」のランクよりも下位にあることが多い。このような、文脈語の分布の違いを反映する非対称な類似度を用いることによって、上位・下位語を判定することが可能となる。

【００４１】

（ｂ）表記ベース方式

本実施例では、「回路」と「電子回路」のような単語レベルでの包含関係にある上位・下位語を抽出する技術を用いる。このような複合語とその主辞となる単語の単語ペアに対してスコアが高くなるようなスコアを特徴量として用いる。この特徴量は、「犬」と「動物」のような種類の上位・下位語は抽出できず汎用的ではないが、専門用語では包含関係を持つ上位・下位語が多く存在し、実用上は強力な手掛かりとなる。

【００４２】

（ｃ）パターンベース方式

パターンベース方式は、単語ペア種別の識別に最も多用される方式であり、抽出するパターンを工夫することによって、様々な単語ペア種別を抽出することができる。上位・下位語については、「Ａ等のＢ」、「ＡのようなＢ」等のパターンを用いる。

【００４３】

（２）対義語

（ａ）文脈ベース方式

文脈ベースの特徴量では、対義語の抽出は難しい。対義語は、ある１つの属性を除き、他の属性が全て一致している単語ペアであり、文脈上は非常に類似していることが理由である。本実施形態では、一部の対義語を抽出するための特徴量として以下に述べる特徴量

を使用する。対義語の中には、「天国」と「地獄」、「善」と「悪」のように一方がポジティブ、他方がネガティブな意味を持つものが多く存在する。そこで、単語がポジティブな意味を持つか、ネガティブな意味を持つかを文脈によって判定し、単語ペアがポジティブ／ネガティブな単語の組である場合にスコアが大きくなる量を考え、対義語であるかどうかを示す特徴量として使用する。単語のポジティブさ、ネガティブさを判定する技術は、公知の技術を採用可能である。一例としては、「を被る」のようなネガティブな表現、「を達成する」のようなポジティブな表現をポジティブ用語、ネガティブ用語の辞書を用いて抽出し、これらの語が文脈に含まれている割合に基づいて、単語のポジティブ／ネガティブさ(マイナスのポジティブ度)を判定する。対義語特徴量としては、単語ペアのポジティブ度の積がマイナスで大きいほど対義語度が高いと考えることとする。この特徴量だけでは、ポジティブな単語とネガティブな単語のペア、例えば＜天国、悪＞、が抽出されるが、他の類似度と組み合わせることで、対義語の識別が可能となる。

【００４４】

（ｂ）表記ベース方式

漢字は表意文字であり、対義語の多くは、対義である漢字を含むことが多い。漢字はそれほど種類が多くないことから、正解の対義語データから、対義である漢字ペアを抽出し、これを手掛かりとすることで、対義語を抽出することが可能であると考えられる。ただし、対義である漢字ペアを含むかどうかだけでは、対義語であるとは言えないため、補助的な条件を加える。対義語の多くは、「連勝」と「連敗」のように対義である漢字ペア以外の文字が一致している場合が多い。また、完全には一致していなくても、「極寒」と「酷暑」のように、「極」と「酷」のように似た意味の漢字を含むことが多い。よって、対義である漢字ペアを含み、かつ同じあるいは類似した意味を持つ漢字を共通に含むかどうかによって特徴量を構成する。また、英語のような表音文字からなる言語に対しても、同様な処理が可能となる。すなわち、単語を意味のある形態素単位で考えることで、”for”と”back”や、”pre”と”post”のような対義関係にある形態素を抽出することが可能であり、漢字のみに限定するものではない。

【００４５】

（ｃ）パターンベース方式

「や」、「と」などの並列助詞は、類似語抽出において最も基本的に用いられるパターンである。通常、同義語が抽出できると考えられがちだが、実際には、「男と女」、「日本や中国」のように、対義語や兄弟語を導く場合が多く、逆に厳密な意味での同義語には使われない。例えば、表記揺れは最も厳密な意味での同義語だが、「コンピュータやコンピューター」のような言い方は、通常用いられない。そこで、並列表現のパターンを対義語、兄弟語抽出のための特徴量として導入する。

【００４６】

ただし、抽出結果を分析すると、同義語が並列で現れる場合も存在する。これは、表記揺れ以外の同義語ペアでは、単語が意味する範囲が完全に一致することはまれであり、意味に差があることが理由である。そのため、並列表現だけでは、単語種別の区別は困難である。そのため、以下のようなパターンを併用するものとした。対義語、兄弟語を含むパターンを分析すると、「地獄から天国」のような表現が多く出現する。これらは、パターンの前後の単語ペアが同義ではないことを示す表現である。このような非同義語パターンと並列表現を組み合わせて使用する。

【００４７】

（３）兄弟語

（ａ）文脈ベース

非対称な類似度の両方が中程度に高い場合が、兄弟語になると考えられる。

（ｂ）表記ベース

兄弟語のみを抽出するための特徴量は特に追加しなかった。

（ｃ）パターンベース

対義語と同じパターンを用いた。兄弟語に固有のパターンは使用していない。

【００４８】
（４）その他
単語ペアに関する特徴量ではないが、単語が固有名詞であるかどうかは重要な情報である。「イラク」と「アフガニスタン」のような単語ペアは、文脈ベース類似度では非常に類似している。しかしながら、固有名詞の場合には、指しているものが同じでなければ、同義語とは言えない。よって、単語ペアが両方とも固有名詞の場合であって、同じものを示さないときには、２つの単語を同義語でないと判定する。
【００４９】
以上のような素性によって単語ペアを表現した後に、多クラスの識別問題として問題を解く。本実施例と従来技術との違いを説明する。図４に、教師なし学習による類似語抽出の概念図を示す。各単語ペアの素性ベクトルは、素性１～Ｎで表現されるＮ次元空間上のある点に相当し、図４では黒丸で表現されている。各単語関係に属する単語ペアを示す黒丸は空間中の近い領域に分布していることが期待される。このとき、教師なし学習では、類似度を計算する関数によってスコアを計算しているが、これは各単語ペアを一次元の直線に射影することに相当する。一次元の直線上に射影されることでランキングが定義され、かつ閾値を設けることによって、類似語かどうかの区別を行う。教師なし方式の問題点は、射影関数(類似度関数)が人手によって決定され、正解等による修正が行い難い点と、閾値が自動的に決定できない点にある。
【００５０】
次に、図５に、２値の教師あり学習による類似語抽出の概念図を示す。２値の教師あり学習では、正解データにしたがって、２つのクラスを区別するのに最も適切な境界が自動的に決定される。このように、教師なしのアプローチによる課題が解決されているが、２つの種類を区別できるだけであり、多くの種類の単語関係を区別する目的には適さない。
【００５１】
次に、図６に、教師ありのランキング学習による類似語抽出の概念図を示す。ランキング学習では、２値の教師あり学習とは異なり、３種類以上のクラスへの分類を扱うことが可能である。正解データに基づいて事例の順序、類似語抽出の場合は単語ペアが似ている程度を学習するので、非常に良く類似している同義語、少し類似している上位・下位語、似ていない非類似語を区別することが可能である。しかしながら、類似している程度という一次元の値を学習するだけなので、上位・下位語、兄弟語、対義語のような類似の仕方が異なる単語ペアを区別することができない。
【００５２】
次に、図７に、本実施形態の多クラスの教師あり学習による類似語抽出の概念図を示す。多クラスの教師あり学習による類似語抽出では、各単語意味関係に対してクラスを割り当てる、各単語意味関係の単語ペアが属する領域を定める境界を自動決定する。これにより、複数の視点による単語ペアの区別が行えるため、類似語内の詳細な単語ペア種別の区別が可能となる。
【００５３】
未知の点、すなわち単語意味関係種別が不明である単語ペアが与えられたとき、いずれの領域に所属するかによって単語意味関係を判定することが多クラスの識別モデルの適用である。
【００５４】
図８は、本発明の第１の実施の形態の単語意味関係抽出装置によって実行される単語意味関係抽出処理のフローチャートである。
【００５５】
ステップ１１において、全ての単語ペアの処理を終了したかどうか判定する。終了していたら、ステップ１７に進む。処理していない単語ペアが存在すれば、ステップ１２に進む。
【００５６】
ステップ１２では、全ての種類の素性について処理を終了したかどうかを判定する。終

了していたらステップ１６に進む。処理していない素性が存在すれば、ステップ１３に進む。

【００５７】

ステップ１３では、ｉ番目の単語ペアを取得する。単語ペアの取得は、例えば、テキストを形態素解析して全単語リストを予め作成しておき、その中から任意の２個の単語の組み合わせを取得すれば良い。

【００５８】

ステップ１４では、取得したｉ番目の単語ペアについて、ｊ番目の素性の計算を行う。ステップ１４の処理の詳細は後述する。

【００５９】

次に、ステップ１５に進み、素性の計算結果を類似度行列に格納する。類似度行列の例は、図３で説明した通りである。

【００６０】

ステップ１６では、類似度行列にラベルを設定する。ラベルはシソーラスを参照することによって設定する。

【００６１】

シソーラスの例を図９に示す。シソーラスは、単語ペアとその単語関係種別を記述したデータである。図９の例では、ある単語ペアに対し、一方の単語を見出し語欄、他方を関連語欄に格納し、タイプ欄に見出し語に対する関連語のタイプを格納している。例えば、図９の例の場合、＜コンピュータ、パソコン＞のような上位・下位語関係にある単語ペアに対し、「コンピュータ」が見出し、「パソコン」が関連語であり、「パソコン」が「コンピュータ」の「下位語」（より具体的な語）であることが格納されている。図９のシソーラスは、辞書引きの都合上、冗長にデータを保持しているものとする。すなわち、＜コンピュータ、パソコン＞という単語ペアに対し、「コンピュータ」を見出し語とした行と、「パソコン」を見出し語とした行の両方を保持しているものとする。ここで、特に単語ペアが上位・下位語関係にある場合には、順序を逆にしたペアのタイプは同様に逆になることに注意が必要である。例えば、「コンピュータ」は「パソコン」の上位語となる。

【００６２】

類似度行列へのラベルの設定では、まず、単語ペアの一方の単語を用いてシソーラスの見出し欄を検索し、見出しが一致した行に対して更に関連語を探索することによって、単語ペアが一致する行を特定する。次に、シソーラスのタイプ欄を取得し、ラベルを設定する。ただし、タイプが上位語、下位語の場合には、関係を考慮して、上位・下位語、下位・上位語のいずれかのラベルを設定する必要がある。図３の例では、同義語の場合のラベルが１、下位・上位語のラベルが２、上位・下位語のラベルが３、対義語のラベルが４、兄弟語のラベルが５、である。単語ペアがシソーラス中に存在しない場合は、以下のように処理する。シソーラス中に、単語ペアを含む行はないが、単語それぞれはシソーラスの別の行に含まれている場合には、非同義語のラベルとして「－１」を付与する。単語の組の少なくとも一方の単語がシソーラスに含まれていない場合には、不明のラベルとして「０」を付与する。

【００６３】

図８に戻り、ステップ１７では識別モデルを学習する。類似度行列中から、ラベルが０ではない行のみを対象に、多クラスの識別モデルを学習する。多クラスの識別モデルの学習方法としては、任意方法を使用することができる。例えば、J. Weston and C. Watkins. Multi-class support vector machines. Royal Holloway Technical Report CSD-TR-98-04、 1998.に開示されている、One versus Rest (One-against-the-Rest)法を用いる。

【００６４】

ステップ１８では、識別モデルに従って、類似度行列の値から単語意味関係抽出を行う。行列中の全ての単語ペアについて、素性ベクトルを学習済みの識別器に入力し、単語意味関係を識別する。識別器の判定結果は、類似度行列の判定結果欄に格納する。これにより、ラベルが「不明」すなわち「０」であった単語ペアに対し、単語意味関係に相当する

ラベルが格納される。また、人手によるシソーラスの誤りチェックに使用することもできる。既に「不明」以外のラベルが付与されている単語ペアに対し、ラベルと判定結果が異なるもののみを抽出し、人手によって確認することによりシソーラスを効率的にチェックすることができる。

【００６５】

以下では、図８のステップ１４の処理を詳細に説明する。ステップ１４では、単語ペアを表現するための素性として、各種の類似度を計算する。以下、類似度のタイプ毎に説明を行う。

【００６６】

（１）文脈ベース類似度

文脈ベース類似度は、単語の文脈の類似性によって単語ペアの類似度を計算する方法である。ある単語の文脈とは、その単語がテキスト中に出現している箇所の「近傍」の単語、あるいは単語列等のことである。何をもって「近傍」と定義するかによって、様々な文脈が定義できる。代表的な方法として、以下では、文脈として、後続する動詞及び直前に出現する形容詞・形容動詞を出現文脈として用いる例を説明するが、これ以外の出現文脈を代替して使用する、あるいは追加・組み合わせて使用することも可能である。また、文脈同士の類似度計算式にも様々な方法が存在する。

【００６７】

以下の例では、文脈ベース類似度を文脈行列１１６に基づいて計算する。文脈行列は、見出し欄と文脈情報欄からなり、見出し欄中の単語に対し、文脈単語列とその頻度の組の繰り返しからなる文脈情報が格納されている。

【００６８】

図１０に文脈行列の例を示す。図１０の例は、着目した単語に後続する助詞＋述語を文脈とした場合を示す。例えば、「コンピュータ」には、「が起動する」が１５回、「を接続する」が４回出現していることを示している。このような文脈行列に対し、任意の２個の単語に相当する行の文脈情報を取得し、文脈単語列の頻度ベクトルに基づいて類似度を計算する。文脈ベース類似度としては、タームベクトルモデルによる文書検索に用いられている方法を用いることができ、例えば、北、津田、獅々掘「情報検索アルゴリズム」共立出版（２００２年）に開示されている方法を用いることができる。本実施の形態では、一例として下式の類似度計算方法によって類似度sを計算する。

【００６９】

【数１】

数１

$$s(b|d)=\frac{1}{L+\kappa*[dlen(b)-L]}*\frac{1}{n}\sum_{i}w(t_i|d)*\nu(t_i|b)$$

ただし、

$$w(t_i|d)=\log\left(1+\frac{\#D}{df(t_i)}\right)*\nu(t_i|d)$$

$$\nu(t_i|d)=\frac{1+\log(tf(t_i|d))}{1+\log(tf(\bullet|d))}$$

d：入力単語

t_i：入力単語のi番目の文脈単語列

b：類似度を計算するターゲット単語

$\#D$：総単語数

$df(t)$：文脈単語列tを文脈として持つ単語数

$tf(t|d)$：入力単語dにおける文脈単語列tの出現回数

$tf(\cdot|d)$：入力単語dに現れる単語の出現回数の平均

$dlen(b)$：ターゲット単語bの持つ文脈単語列の種類数

L：単語毎の文脈単語列種類数の平均値

κ：文脈単語列種類数正規化のための定数

【００７０】

ここで、s(b|d)とs(d|b)の値は一般的には異なる、すなわち非対称であるため、単語ペア(b、d)に対し、s(b|d)とs(d|b)の両方を計算し、それぞれを異なる素性として用いる。このように、本実施例では、単語の組の類似度として、単語の組の２つの単語の文脈情報同士の類似度であって、非対称な単語の組の一方を基準に計算した類似度と他方を基準にして計算した類似度の２種類の類似度を計算する。つまり、非対称な２種類のスコアを素性として用いることにより、両方のスコアが高い場合は同義語、一方が他方よりも高い場合は上位・下位語、両方が中程度に高い場合には兄弟語といったように、境界を設定することが可能となる。

【００７１】

文脈行列の作成方法については、テキストを形態素解析した後、形態素解析結果に対して品詞パターンを適用する、構文解析を行うなどの方法で作成するなど、公知の手法を適用可能である。

【００７２】

（２）表記ベース類似度

以下では、表記ベース類似度を計算する方法について説明する。表記ベース類似度は、単語の組に対し、文字の情報に基づいて類似度を計算する。同義語が特に、「コンピュータ」と「コンピューター」のような異表記語の場合、非特許文献２に開示されているように、多くの文字が重複していることから文字の重複している割合は類似度として用いることができる。異表記語は原則的にカタカナ語であるが、漢字からなる単語ペアでも、意味

が類似している場合に、「分析」と「解析」、「信頼」と「信用」のように同じ文字が含まれることは多い。以下では、文字の重複割合に基づく類似度を文字重複度と呼ぶ。漢字からなる単語の場合、特に２文字単語のような文字数が短い単語の場合は、「分析」と「透析」のように同じ文字を含んでいても意味が異なる単語が多く存在する。本実施例では、文脈ベース類似度のような異なる種類の類似度と組み合わせることによって、文字重複度が有効に作用する。

【００７３】

さらに、漢字の場合には、異なる文字であっても意味が類似している文字が存在する。例えば、「慕（う）」、「憧（れる）」のような文字は類似した意味を持っている。このような文字の類似性を教師データから学習することができれば、文字が完全に一致していない場合でも、単語間の表記ベース類似度を計算することが出来る。文字の類似性に基づく単語の類似度を類似文字重複度と呼ぶ。

【００７４】

（ａ）文字重複度

文字の重複度は、様々な方法で計算することができるが、ここでは一例として２個の単語間で共通に含まれている文字をカウントし、２個の単語のうち短い方の単語の文字列長で正規化することで計算する方法を説明する。同じ文字が複数含まれている場合には、一方にｍ個、他方の単語にｎ個含まれている場合には、ｍ対ｎの対応関係となる。このような場合は、ｍ又はｎの小さい方の個数の文字が重複したものとする。

【００７５】

以下では、２個の単語ｉと単語ｊの表記ベース類似度の計算方法について図１１を用いて説明する。

【００７６】

ステップ１４１１において、単語ｉの全ての文字を処理したかどうか調べる。処理していれば、ステップ１４１５に進む。未処理の文字があれば、ステップ１４１２に進む。ステップ１４１２では、単語ｊの全ての文字を処理したかどうか調べる。処理していれば、ステップ１４１１に進む。未処理の文字があれば、ステップ１４１３に進む。

【００７７】

ステップ１４１３では、単語ｉのｍ番目の文字と単語ｊのｎ番目の文字を比較し、一致するかどうか調べる。一致していれば、ステップ１４１４に進む。一致していなければ、ステップ１４１２に進む。ステップ１４１４では、単語ｉのｍ番目の文字と単語ｊのｎ番目の文字にそれぞれフラグを立てる。その後、ステップ１４１２に進む。

【００７８】

ステップ１４１５では、単語ｉ、単語ｊのフラグが立った文字数をそれぞれカウントし、小さい方を一致文字数とする。例えば、「ウインドウ」と「ウィンドー」が処理対象であると仮定すると、「ウ」、「ン」、「ド」の３文字が一致する。「ウ」については、「ウインドウ」に２文字含まれているため、「ウインドウ」中でフラグが立った文字は４文字、「ウィンドー」中でフラグが立った文字は３文字となる。よって、３文字が一致したものとする。

【００７９】

以上の方法以外にも、２個の単語の語頭からの共通部分文字列長を重複度とする、２個の単語の語末からの共通部分文字列長を重複度とする、正規化する文字列長を両者の平均とする、長い方とするなどのバリエーションが考えられる。また、より精緻な方法として、例えば、ＤＰマッチングなどによって２個の単語を照合し、マッチした文字数に基づいて表記ベース類似度を計算することも可能であり、利用可能な計算リソースに応じて、より多数の表記ベース類似度を計算することもできる。また、文字の頻度に基づいて、文字が一致した際の重みを変更することもできる。文書の検索において、単語の重みを計算する方法としてＩＤＦ（Inversed Document Frequency）が知られているが、同様の考え方で多くの単語に共通して含まれている文字の重要性は小さいと考えることで文字の重みを計算することができる。

【００８０】
（ｂ）類似文字重複度
同義語辞書から文字の類似度を学習し、類似文字も含めて文字の重複度を計算する。文字の類似度の計算方法について、図１２に示すフローチャートを用いて説明する。
【００８１】
ステップ１４２１において、同義語辞書から同義語である単語ペアを取得する。次に、ステップ１４２２において、単語ペアの一方の単語から取り出した文字と他方の単語から取り出した文字からなる文字ペアを全ての組み合わせについて取得する。例えば、「敬慕」、「憧憬」が同義語である単語ペアの場合、「敬」／「憧」、「敬」／「憬」、「慕」／「憧」、「慕」／「憬」という４種類の文字ペアを取得する。
【００８２】
次に、ステップ１４２３に進み、同義語辞書中の全ての単語に含まれる文字の頻度を計算する。次に、ステップ１４２４に進み、全ての文字ペアについて文字類似度を計算する。文字類似度は、文字ペアの頻度を、文字ペアを構成する２個の文字の頻度で割ったもの（Dice係数）を用いる。自己相互情報量等を類似度として用いても良い。
【００８３】
ステップ１４２５では、ステップ１４２４で計算した類似度について、同じ文字についての類似度と異なる文字についての類似度を正規化する。具体的には、同じ文字についての類似度の平均ＡＳと異なる文字についての類似度の平均ＡＤをそれぞれ計算する。同じ文字については、計算した類似度に関わらず、１．０を設定する。異なる文字については、ステップ１４２４で計算した値にＡＤ／ＡＳを掛け算した値を最終的な類似度とする。
【００８４】
文字類似度テーブルの例を図１３に示す。文字類似度テーブルを利用して類似文字重複度を計算することが可能である。類似文字重複度の計算は、文字重複度の計算と同様に行えば良い。異なる部分は、文字重複度では文字が一文字一致した場合に、文字数１を加算していたのに対し、類似文字重複度の場合は、類似文字テーブルを参照し、類似文字である場合には、文字類似度を加算する点である。文字が一致する場合には、類似文字テーブルには１．０が格納されているため、文字重複度と同じである。
【００８５】
なお、意味が類似した形態素(単語の部分文字列)の類似性を用いる方法、非特許文献４に開示されているような単語の包含関係を用いる方法によって得られる類似度を用いることができる。
【００８６】
以下では、詳細な単語意味関係を抽出するために必要な類似度を構成する方法について述べる。表記ベース類似度においても、文脈ベース類似度の場合と同様に、単語の組の一方を基準に計算した類似度と他方を基準にして計算した類似度の２種類の類似度、すなわち非対称な類似度の組を構成することが可能である。例えば、Jaccard係数を例に考える。Jaccard係数は、２種類の集合の類似度を、積集合の要素数の和集合の要素数の割合で示すものである。例えば、「銀行」と「投資銀行」のような単語ペアがあったときに、これを「銀」、「行」という文字から集合と「投」、「資」、「銀」、「行」という４文字からなる集合だと考えると、積集合（一致した文字）の要素数は2、和集合の要素数は4であり、Jaccard係数は0.5となる。Jaccard係数は対称である。ここで、和集合ではなく、単語ペアの一方の単語に着目し、その単語に含まれる文字を用いることを考える。すると、「銀行」に着目した場合は、スコアは2/2=1.0であり、「投資銀行」に着目したときには、2/4=0.5となり、非対称となり、「銀行」が「投資銀行」の上位語であることを表現している。このように非対称な特徴量の組を構成し、両方を特徴量として用いることで詳細な単語意味関係を精度良く抽出することが可能となる。
【００８７】
（３）パターンベース類似度
パターンベース類似度は、「AのようなB」、「AやBなどのC」のような単語意味関係を

明示的に示すパターンを使用する。予め定められたパターンと文字列、あるいは形態素解析結果と照合することによって、パターンと合致する単語ペアを取得する。抽出した単語ペアの数を集計し、正規化などの統計処理を行い素性の次元の値とする。パターンベース類似度の計算方法は、非特許文献３に開示されているので、説明は省略する。

【００８８】

以下では、詳細な単語意味関係を抽出するために必要な類似度の構成方法について述べる。単語の組の一方を基準にして算出された素性の値と他方を基準にして算出された素性の値の２種類を算出する。例えば、「AのようなB」、「AなどのB」のような上位・下位語を抽出するためのパターンについては、パターン自体に方向性がある。すなわち、「AのようなB」が自然な表現の場合、「BのようなA」が使用されることはない。類似度行列では、単語ペア＜A、B＞と＜B、A＞を区別せず、ラベルとして上位・下位語と下位・上位語を用いて表現することとしたため、このような上位・下位語を示すパターンから得られる特徴量は、「AのようなB」が出現したことを示す素性と「BのようなA」が出現したことを示す素性の両方を準備する。「顧客関係管理（ＣＲＭ）」のような括弧表現は、同義語を示すことが多い表現であり、有効である。しかしながら、必ずしも同義語のみで使われる訳ではない。例えば、「Ａ社（東京都）」のような、名詞とその属性のような場合に使用されることもある。このような場合にも、同義語の場合には、括弧外と括弧内の表現が交換可能であり、方向性がなく、属性表現の場合には、括弧外と括弧内の表現は交換できない。よって、「A（B）」が出現したことを示す特徴量と「B（A）」が出現したことを示す特徴量を両方用いることで、同義語のケースと属性のケースを区別できる。「AやB」、「AとB」のような並列表現については、本質的には方向性はないが、文の構造の解析が正しく行えないと正確な処理ができない。例えば、「A社と契約を締結」のような表現では、「と」は並列を示す助詞ではないが、誤って並列助詞として処理されてしまう可能性がある。このようなケースについても、「契約とA社」のような表現があるかどうかを考慮して特徴量を構成することで、真に同義である単語ペアのみを抽出可能である。

【００８９】

こうして本発明の第１の実施の形態の単語意味関係抽出装置によると、人手作成によるシソーラスなどの付加的な情報源を正解として用いると同時に、文脈ベース、表記ベース、パターンベースなどの異なるタイプの類似度を統合することにより、従来と比較して高精度な単語意味関係抽出を行うことが可能となる。特に、類似語内の同義語、上位・下位語、対義語、兄弟語などのより詳細な種別を判定することが可能となる。また、各種別の詳細な区別が可能になることにより、種別毎の抽出精度が向上する。

【実施例２】

【００９０】

図１４はコンテンツクラウドシステムの概略図である。コンテンツクラウドシステムは、Ｅｘｔｒａｃｔ　Ｔｒａｎｓｆｏｒｍ　Ｌｏａｄ（ＥＴＬ）２７０３モジュール、ストレージ２７０４、検索エンジン２７０５モジュール、メタデータサーバ２７０６モジュール、マルチメディアサーバ２７０７モジュールから構成される。コンテンツクラウドシステムは１つ以上のＣＰＵ、メモリ、記憶装置を備えた一般的な計算機上で動作し、システム自体は様々なモジュールで構成されている。また、それぞれのモジュールが独立した計算機で実行されることもあり、その場合、各ストレージとモジュール間はネットワーク等で接続されおり、それらを介してデータ通信を行う分散処理で実現される。コンテンツクラウドシステムには、アプリケーションプログラム２７０１がネットワーク等を経由してリクエストを送り、コンテンツクラウドシステムはリクエストに応じた情報をアプリケーション２７０１に送信する。

【００９１】

コンテンツクラウドシステムは入力として音声データ２７０１－１、医療データ２７０１－２、メールデータ２７０１－３などの任意の形式のデータを対象とする。各種データは、例えば、コールセンター通話音声、メールデータ、文書データなどであり、構造化されていたり、されていなかったりしてもよい。コンテンツクラウドシステムへ入力される

データは各種ストレージ２７０２に一時的に蓄えられる。

【００９２】

コンテンツクラウドシステムにおけるＥＴＬ２７０３は、ストレージを監視しており、ストレージへの各種データ２７０１の蓄積が完了すると、そのデータに合わせた情報抽出処理モジュールを動作させ、抽出された情報（メタデータ）をコンテンツストレージ２７０４にアーカイブ化して保存する。ＥＴＬ２７０３は、例えば、テキストのインデックスモジュール、画像認識モジュールなどで構成されており、メタデータの例としては、時刻、Ｎ－ｇｒａｍインデックスや画像認識結果（物体名）、画像特徴量とその関連語、音声認識結果、などが該当する。これらの情報抽出モジュールには、何らかの情報（メタデータ）抽出を行うプログラムすべてを用いることができ、公知の技術を採用することができるので、ここでは各種情報抽出モジュールの説明を省略する。必要ならば、メタデータはデータ圧縮アルゴリズムによって、データサイズの圧縮が行われてもよい。また、各種モジュールで情報を抽出したのち、データのファイル名、データ登録年月日、元データの種類、メタデータテキスト情報などをＲｅｌａｔｉｏｎａｌ　Ｄａｔａ　Ｂａｓｅ（ＲＤＢ）へ登録する処理が行われても良い。

【００９３】

コンテンツストレージ２７０４には、ＥＴＬ２７０３で抽出された情報及びストレージ２７０２に一時的に蓄えられている処理前のデータ２７０１が保存される。検索エンジン２７０５は、アプリケーションプログラム２７０１からのリクエストがあると、例えば、テキスト検索であれば、ＥＴＬ２７０３で作成されたインデックスを元に、テキストの検索を行い、検索結果をアプリケーションプログラム２７０１に送信する。ここで、検索エンジンやそのアルゴリズムに関しては、公知の技術を適用することができる。検索エンジンはテキストだけでなく、画像、音声などのデータを検索するモジュールが含まれうる。

【００９４】

メタデータサーバ２７０６は、ＲＤＢに蓄えられたメタデータの管理を行う。例えば、ＥＴＬ２７０２において、データのファイル名、データ登録年月日、元データの種類、メタデータテキスト情報、などがＲＤＢに登録されているとすると、アプリケーション２７０１からリクエストの要求があると、リクエストに従って、データベース内の情報をアプリケーション２７０１に送信する。

【００９５】

マルチメディアサーバ２７０７では、ＥＴＬ２７０３で抽出されたメタデータ同士の情報を互いに関連付けを行い、グラフ形式で構造化してメタ情報の保存を行う。関連図付けの一例としては、コンテンツストレージ２７０４に蓄えられた「リンゴ」という音声認識結果に対して、元の音声ファイルや画像データ、関連語などがネットワーク形式で表現される。マルチメディアサーバ２７０７もアプリケーション２７０１からのリクエストがあると、それに応じたメタ情報をアプリケーション２７０１に送信する。例えば、「リンゴ」というリクエストがあると、構築されたグラフ構造に基づき、リンゴの画像や平均相場、アーティストの曲名、などの関連メタ情報を提供する。

【００９６】

コンテンツクラウドシステムにおいて、シソーラスは以下のように使用される。

【００９７】

まず、メタデータの検索において活用するというのが第１のパターンである。音声認識結果が「リンゴ」のようなメタデータで表現されている場合に、「林檎」のようなクエリが入力された場合、クエリを、シソーラスを用いて同義語に変換することによって検索可能とすることができる。また、付与されたメタデータが一貫しておらず、あるデータには、「リンゴ」、別のデータには「林檎」が付与されている場合に、同一のメタデータが付与されているものとして扱うことが可能となる。

【００９８】

次に、メタデータの付与の際、特にテキスト情報を利用したメタデータの付与の際に活用するというのが第２のパターンである。例えば、HTML文書のようなテキスト中に画像が

含まれるようなテキストを用いて画像にメタデータを付与するタスクを考える。画像のメタデータは、テキスト中に含まれる単語を統計処理することによって得られるが、スパースネスと呼ばれる、データ量が不足しており正確に統計処理を行えない問題によって、精度が低下することが知られている。シソーラスを用いることで、このような問題を回避することが可能となり、高い精度でメタデータを抽出することが可能となる。

【００９９】

以上、本発明の実施形態について説明したが、本発明は上記実施形態に限定されるものではなく、種々変形実施可能であり、上述した各実施形態を適宜組み合わせることが可能であることは、当業者に理解されよう。

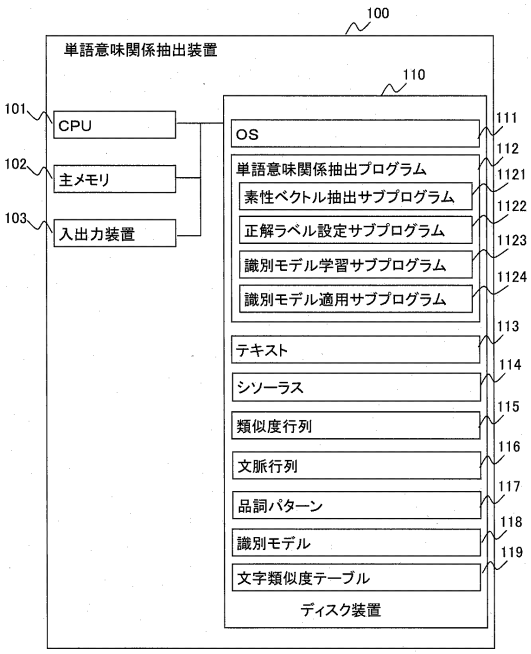
【符号の説明】

【０１００】

１００　単語意味関係抽出装置
１０１　ＣＰＵ
１０２　主メモリ
１０３　入出力装置
１１０　ディスク装置
１１１　ＯＳ
１１２　単語意味関係抽出プログラム
１１２１　素性ベクトル抽出サブプログラム
１１２２　正解ラベル設定サブプログラム
１１２３　識別モデル学習サブプログラム
１１２４　識別モデル適用サブプログラム
１１３　テキスト
１１４　シソーラス
１１５　類似度行列
１１６　文脈行列
１１７　品詞パターン
１１８　識別モデル
１１９　文字類似度テーブル

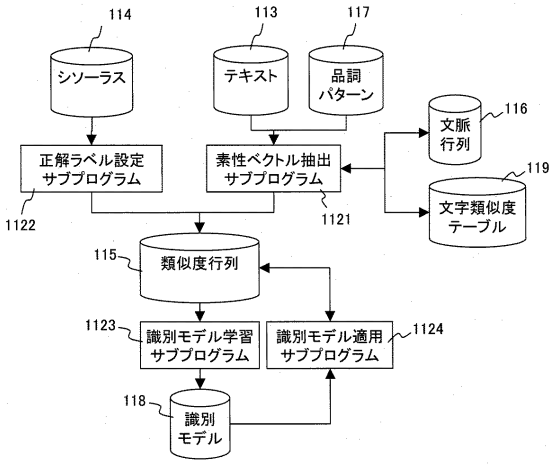
【図１】

図1



【図２】

図2



【図３】

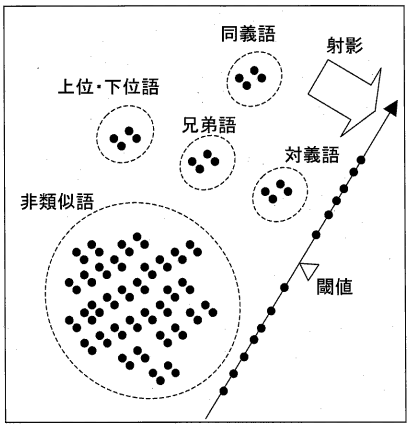
図3

単語ペア	素性1	素性2	素性3	⋮	素性N	ラベル	判定結果
…	…	…	…	…	…	…	…
<コンピュータ, コンピューター>	0.3	0.2	0.4	…	0.8	1	1
…	…	…	…	…	…	…	…
<コンピュータ, パソコン>	0.3	0.2	0.4	…	0.8	3	1
…	…	…	…	…	…	…	…
<コンピュータ, プログラム>	0.0	0.1	0.0	…	0.2	−1	−1
…	…	…	…	…	…	…	…
<コンピュータ, メインフレーム>	0.0	0.1	0.0	…	0.2	2	−1
…	…	…	…	…	…	…	…
<コンピュータ, 計算機>	0.1	0.2	0.0	…	0.3	1	1
…	…	…	…	…	…	…	…
<サーバ, ルータ>	0.0	0.1	0.0	…	0.2	5	−1
…	…	…	…	…	…	…	…
<起動, 停止>	0.1	0.2	0.0	…	0.3	4	1
…	…	…	…	…	…	…	…
<仮想化技術, 計算機>	0.2	0.1	0.0	…	0.2	0	−1
…	…	…	…	…	…	…	…

115

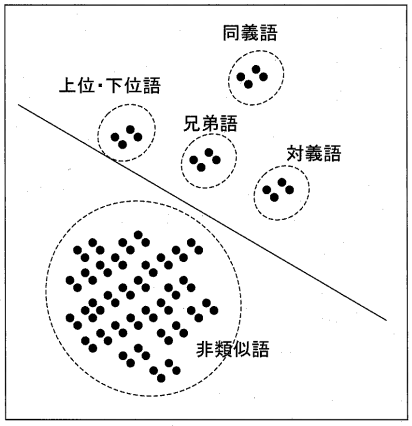
【図４】

図4



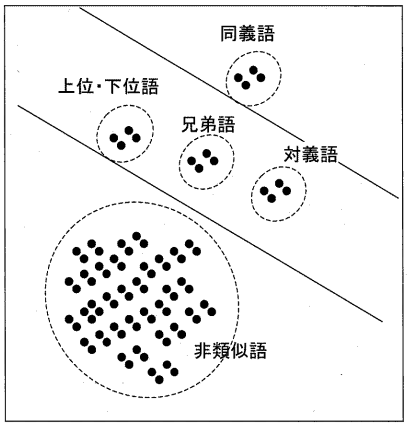
【図５】

図5



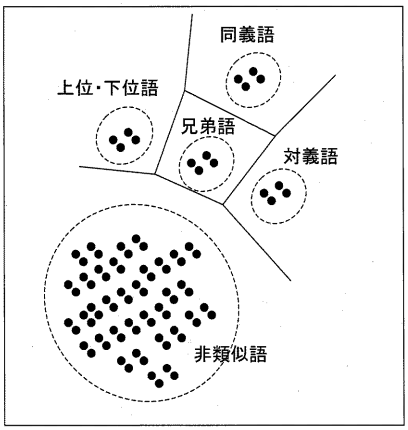
【図６】

図6



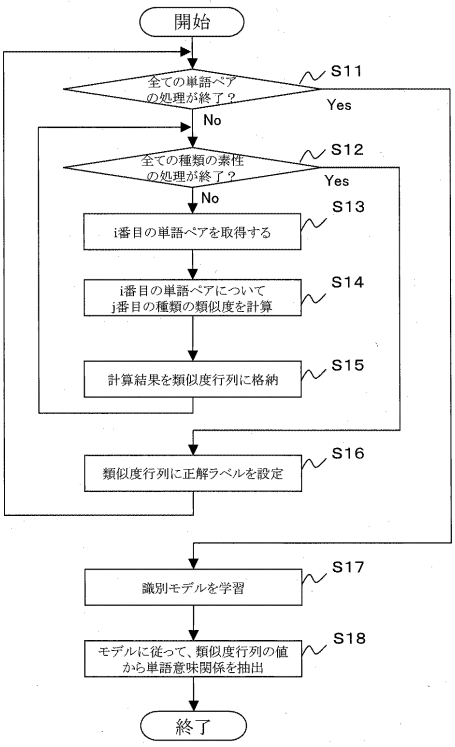
【図７】

図7



【図８】

図8



【図９】

図9

見出し	関連語	タイプ
…	…	…
コンピュータ	コンピューター	同義語
…	…	…
コンピュータ	パソコン	下位語
…	…	…
コンピュータ	メインフレーム	上位語
…	…	…
コンピューター	コンピュータ	同義語
…	…	…
パソコン	コンピュータ	上位語
…	…	…
メインフレーム	コンピュータ	上位語
…	…	…
起動	停止	対義語
…	…	…
停止	起動	対義語
…	…	…

【図１０】

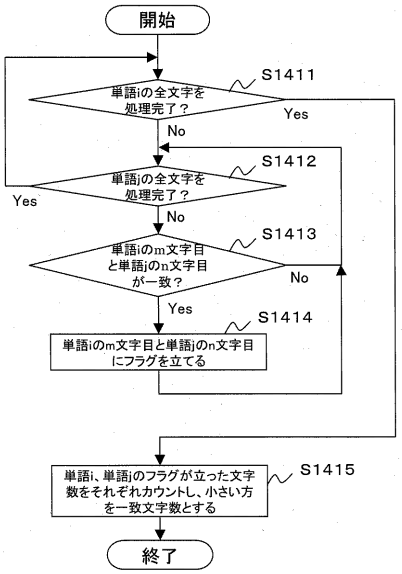
図10

見出し	文脈情報					
	文脈単語列	頻度	文脈単語列	頻度	文脈単語列	頻度
…	…	…	…	…	…	…
コンピュータ	が起動する	15	を接続する	4	を停止する	8
コンピューター	が起動する	10	を接続する	2	を停止する	4
…	…	…	…	…	…	…
プログラム	が終了する	4	が停止する	3	を開発する	6
…	…	…	…	…	…	…
計算機	が起動する	12	を接続する	2	を停止する	5
…	…	…	…	…	…	…

116

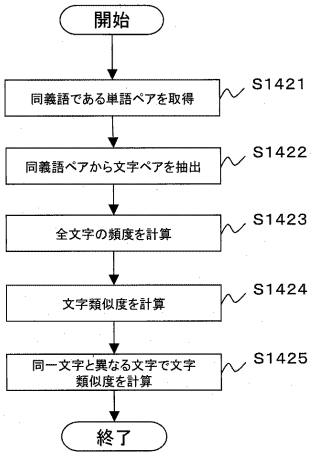
【図１１】

図11



【図１２】

図12



【図１３】

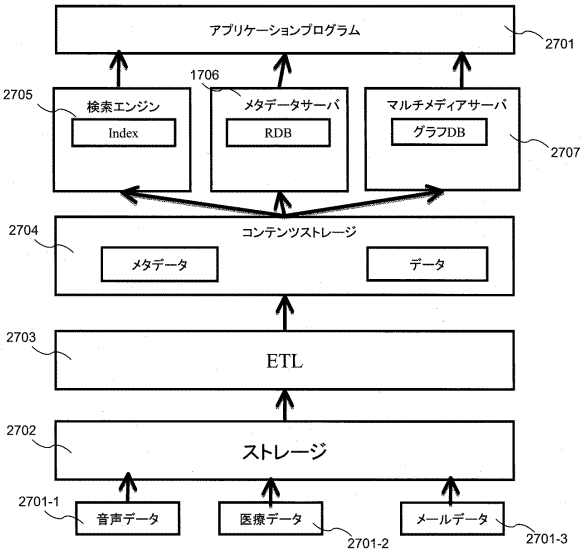
図13

見出し	共起語	文字類似度
…	…	…
安	穏	0.50
…	…	…
叡	恵	0.40
…	…	…
禍	災	0.35
…	…	…

119

【図１４】

図14



フロントページの続き

(56)参考文献 特開２０１１－１１８５２６（ＪＰ，Ａ）
特開２０１１－１７５４９７（ＪＰ，Ａ）
特開２０１２－１０８５７０（ＪＰ，Ａ）
特開２００７－０１１７７５（ＪＰ，Ａ）
特開２００５－３０９７０６（ＪＰ，Ａ）
特開２００６－２２８０４２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０－２８