

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2014年6月26日(26.06.2014)



(10) 国際公開番号
WO 2014/097751 A1

- (51) 国際特許分類:
G06F 17/27 (2006.01) *G06F 17/21* (2006.01)
- (21) 国際出願番号: PCT/JP2013/079479
- (22) 国際出願日: 2013年10月24日(24.10.2013)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2012-274396 2012年12月17日(17.12.2012) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION)
[JP/JP]; 〒1088001 東京都港区芝五丁目7番1号
Tokyo (JP).
- (72) 発明者: 久野 綾子(KUNO, Ayako); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内
Tokyo (JP). 平尾 英司(HIRAO, Eiji); 〒1088001 東京都港区芝五丁目7番1号日本電気株式会社内
Tokyo (JP).
- (74) 代理人: 池田 憲保, 外(IKEDA, Noriyasu et al.); 〒1000011 東京都千代田区内幸町1丁目2番2号
日比谷ダイビル Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI

[続葉有]

(54) Title: DEFINED EXPRESSION EXTRACTION SYSTEM, METHOD, AND PROGRAM

(54) 発明の名称: 定義済み表現抽出システム、方法およびプログラム

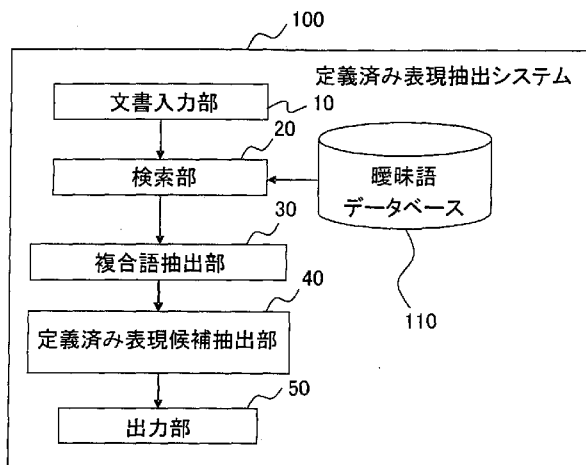


図 1

- 10 Document input unit
20 Search unit
30 Compound word extraction unit
40 Defined expression candidate extraction unit
50 Output unit
100 Defined expression extraction system
110 Fuzzy word database

(57) Abstract: A search unit searches for fuzzy words in an inputted document, querying a fuzzy word database wherein fuzzy words which are attached to dependent words and fuzzy word attributes are accumulated, and if fuzzy words are present, extracts the text containing the fuzzy words from the inputted document. A compound word extraction unit extracts, from the extracted text containing the fuzzy words, a group of nouns adjacent to the fuzzy words as compound words including the fuzzy words. A defined expression candidate extraction unit extracts expressions with the fuzzy words included from the compound words including the fuzzy words, and computes degrees of definition which denote the degrees of probability of said expressions being defined. A list of defined expression candidates is outputted from an output unit.

(57) 要約: 検索部は、付属語に属する曖昧語と曖昧語の属性を蓄積する曖昧語データベースを参照して入力文書中の曖昧語の有無を検索し、曖昧語が有る場合は曖昧語を含む文を入力文書から抽出する。複合語抽出部は、抽出された曖昧語を含む文から曖昧語に隣接する名詞群を曖昧語込み複合語として抽出する。定義済み表現候補抽出部は、曖昧語込み複合語から曖昧語付き表現を抽出し、定義済みの可能性の度合いを示す定義度を算出する。定義済み表現候補の一覧が出力部から出力される。

WO 2014/097751 A1



(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告 (条約第 21 条(3))

明 細 書

発明の名称

定義済み表現抽出システム、方法およびプログラム

5

技術分野

本発明は、定義済み表現抽出システム、方法およびプログラムに関する。

背景技術

- 10 近年、情報処理装置を用いて、自然言語で書かれた文書を分析して、その文書の曖昧性を分析する定義済み表現抽出システムが開発されている。その定義済み表現抽出システムに関する技術の一例が、特許文献1に「関連語自動抽出装置、複数重要語抽出プログラムおよび重要語の上下階層関係抽出プログラム」として開示されている。特許文献1に開示された関連語自動抽出装置は、文書群を格納
- 15 するデータベースを有するデータベース部と、重要語解析部と、カウント部と、関連後抽出部と、を備える。重要語解析部は、前記データベース中に含まれる文書の内容を特徴付けると考えられる名詞からなる重要語を抽出する。カウント部は、重要語解析部で抽出した重要語の前記データベース中の全出現回数と重要語が前記データベース中に含まれる文書数とを含む統計情報を取得し、前記重要語
- 20 解析部で抽出した特定の重要語が含まれる文書における特定の重要語から一定の範囲内に存在する別の重要語の出現数を取得する。関連後抽出部は、前記統計情報と前記特定の重要語が一定の範囲内に存在する別の重要語の出現数とから演算される特定の重要語と別の重要語とが一定範囲内に出現する出現確率に基づいて特定の重要語と別の重要語との関連度を判定する。このような構成により、一般
- 25 的な既存のシソーラス辞書には記載されていない、ユーザーが指定した特定分野に出現する専門用語や、新語および流行語を自動抽出し、さらにユーザーが指定した言葉に関連の深い重要語を抽出する機能を提供している。

さらに、定義済み表現抽出システムに関する技術の他の例が、特許文献2に「専門用語判別装置および専門用語判別方法」として開示されている。この専門用語

判別装置は、複合語分割部と、意味距離計算部と、専門性判定部と、を備える。複合語分割部は、構成語を組み合わせて構成される複合語を、各構成語に分割する。意味距離計算部は、複合語分割部により分割された構成語間の単語間の意味的な近さを表す意味距離を計算する。専門性判定部は、意味距離計算部により計算された意味距離の大きさに基づいて、複合語の専門性を判定する。このような構成により、あらかじめ構成語が専門用語であるか否かについて情報を保持していなくても、複合語を構成する構成語間の意味距離の大きさに基づいて、複合語の専門性を判定する機能を提供している。

10 先行技術文献

特許文献

特許文献 1：特許第 3 5 5 3 5 4 3 号公報

特許文献 2：特許第 4 9 6 1 8 6 9 号公報

15 発明の概要

発明が解決しようとする課題

特許文献 1、2 に開示された技術は、以下の課題を有している。例えば、「半」、「準」や「等」、「系」といった省略された概念を読み手に推定させる作用を持つ、接頭語や接尾語などの付属語に属する曖昧語を含む文書の曖昧性の評価に、上記技術による分析方法を適用しても、優先的な修正が必要な曖昧性の高い箇所の評価は行えない。その理由は以下の通りである。接頭語や接尾語などの付属語に属する曖昧語は、単語と組み合わせられた状態で語義が定義済みであるか未定義であるかでその曖昧性が大きく異なる。しかし、単に複合名詞を定義済みと解釈するだけでは、一つの名詞に接頭語や接尾語がついているだけというような単純な複合名詞も定義済みとしてしまい、曖昧性の高い箇所を見逃してしまう。

また、特許文献 2 の手法で用いられているような複合語の構成語間の意味距離で専門用語かどうか判定するような方法では、構成語が意味的に独立していることが求められ、接頭語・接尾語のように結合される単語によって意味が変わる複合語に適用しても、適切な専門用語の検出ができず、語義が定義済みで曖昧性が

低いと考えられるパターンを検出できない。なお、ここでの文書の曖昧性とは、文書の書き手と読み手の間の理解に齟齬が起きにくいことや、文書が読み手にとって理解しやすいかといった、少なくとも複数の解釈が起きる可能性を含む、文書の情報伝達の性能に関わる特徴を指す。

- 5 本発明は、上記課題に鑑み、文書の曖昧性を評価する定義済み表現抽出システム、方法およびプログラムを提供しようとするものである。

課題を解決するための手段

- 10 本発明に係る定義済み表現抽出システムは、「半」、「準」や「等」、「系」といった省略された概念を読み手に推定させる作用を持つ接頭語や接尾語などの付属語に属する曖昧語を含む文書について、独有用語や業界用語、法律用語、定型表現のように語義が定義済みで曖昧性が低いと考えられるパターンを検出して除外し、語義が未定義で曖昧性が高いと考えられるパターンを優先的な修正が必要な曖昧性の高い箇所として抽出することで文書の曖昧性を評価する。

- 15 本発明の第1の態様によれば、対象とする文書もしくは文書群を入力文書として受け付ける文書入力部と、曖昧語とその属性を蓄積した曖昧語データベースと、前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索し、曖昧語を含む文を抽出する検索部と、抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出する複合語抽出部と、抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、
20 抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出部と、前記定義済み表現候補を出力する出力部と、を備えた定義済み表現抽出システムが提供される。

- 25 本発明の第2の態様によれば、曖昧語とその属性を蓄積した曖昧語データベースを用意し、対象とする文書もしくは文書群の入力を入力文書として受け付け、前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索することにより、曖昧語を含む文を抽出し、抽出された曖昧語を含む文を形態素解析して、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出し、抽出され

た曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、かつ抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出し、前記定義済み表現候補を出力する、

5 定義済み表現抽出方法が提供される。

本発明の第3の態様によれば、コンピュータに、対象とする文書もしくは文書群を入力文書として受け付ける文書入力処理と、曖昧語とその属性を蓄積した曖昧語データベースを参照して、前記入力文書中の曖昧語の有無を検索し、曖昧語を含む文を抽出する検索処理と、抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出する複合語抽出処理と、抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出処理

10
15 と、前記定義済み表現候補を出力する出力処理と、を実行させるためのプログラムが提供される。

発明の効果

本発明によれば、「半」、「準」や「等」、「系」といった省略された概念を読み手に推定させる作用を持つ接頭語や接尾語などの付属語に属する曖昧語を含む文書について、語義が未定義で曖昧性が高いと考えられるパターンを優先的な修正が必要な曖昧性の高い箇所として抽出することで文書の曖昧性を評価することができる。

図面の簡単な説明

25 図1は、本発明の実施形態に係る定義済み表現抽出システムの構成を示すブロック図である。

図2は、図1に示した定義済み表現抽出システムの動作例を示すシーケンス図である。

図3は、文書入力部に入力する文の例を示す図である。

図４は、曖昧語データベースに蓄積するデータの例を示す図である。

図５は、出力部で出力する情報の例を示す図である。

図６は、出力部で出力する情報の例を示す図である。

5 発明を実施するための形態

[実施形態]

最初に、本発明の実施形態について、図面を参照して詳細に説明する。

図１は、本発明の実施形態に係る定義済み表現抽出システムの構成を示すブロック図である。

- 10 図１を参照すると、本発明の実施形態に係る定義済み表現抽出システム１００は、文書入力部１０と、検索部２０と、複合語抽出部３０と、定義済み表現候補抽出部４０と、出力部５０と、曖昧語データベース１１０と、を含み、基本的に、電子機器内、もしくはサーバと電子機器およびこれらを相互に接続するインターネット等の情報通信ネットワークからなるシステム内に備えられる。

- 15 図示の定義済み表現抽出システム１００は、「半」、「準」や、「等」、「系」といった省略された概念を読み手に推定させる作用を持つ接頭語や接尾語などの付属語に属する曖昧語を含む文書について、独自用語や業界用語、法律用語、定型表現のように語義が定義済みで曖昧性が低いと考えられるパターンを検出して除外し、語義が未定義で曖昧性が高いと考えられるパターンを優先的な修正が必
20 要な曖昧性の高い箇所として抽出することで文書の曖昧性を評価する。

- 電子機器で定義済み表現抽出システムを構成する場合、定義済み表現抽出システム１００は、プログラム制御により動作するコンピュータで実現可能である。図示はしないが、この種のコンピュータは、周知のように、データを入力する入力装置と、データ処理装置と、データ処理装置での処理結果を出力する出力装置
25 と、種々のデータベースとして働く補助記憶装置とを備えている。そして、データ処理装置は、プログラムを記憶するリードオンリメモリ（ROM）と、データを一時的に記憶するワークエリアとして使用されるランダムアクセスメモリ（RAM）と、ROMに記憶されたプログラムに従って、RAMに記憶されているデータを処理する中央処理装置（CPU）とから構成される。

この場合、データ処理装置が文書入力部 10、検索部 20、複合語抽出部 30、定義済み表現候補抽出部 40、出力部 50として機能し、補助記憶装置が曖昧語データベース 110として機能し、出力装置が出力部 50として機能する。

次に、定義済み表現抽出システム 100を構成する各構成要素の動作について
5 説明する。

文書入力部 10は、曖昧語を含む可能性が有り優先的な修正が必要な曖昧性の高い箇所を分析する文書もしくは文書群の入力を受け付ける。

曖昧語データベース 110は、省略された概念を読み手に推定させる作用を持ち、接頭語や接尾語などの付属語に属する曖昧語とその曖昧語が接頭語か接尾語
10 かを区別するための属性情報を蓄積したデータベースである。なお、曖昧語データベース 110として、インターネット上のデータベースを使用しても良い。

検索部 20は、曖昧語データベース 110を参照して、曖昧語を検索し、曖昧語と曖昧語を含む文を抽出する。ここで、文の区切りは句点、改行、記号、空白といった特定の文字列で判断すればよい。

15 複合語抽出部 30は、検索部 20で抽出した曖昧語を含む文に形態素解析を適用することで、曖昧語を含む複合語を取得する。曖昧語を含む複合語とは、曖昧語に隣接する名詞を結合した単語群であり、曖昧語の前後にそれぞれ名詞がある単語群である。つまり、曖昧語を先頭・末尾以外の箇所に含む複合語である。以後、これを曖昧語込み複合語と呼ぶ。なお、ここでは検索部 20で曖昧語を検索
20 してから、複合語抽出部 30で複合語を抽出するフローを想定した構成としているが、複合語抽出部 30で先に文章中の全複合語を抽出してから、検索部 20で曖昧語込み複合語を抽出してもよい。

定義済み表現候補抽出部 40は、複合語抽出部 30で抽出した曖昧語込み複合語から、曖昧語付き表現を抽出し、曖昧語込み複合語と曖昧語付き表現を定義済
25 みの可能性が高い定義済み表現候補と判定して取得する。ここで、曖昧語付き表現とは、曖昧語が接頭語の場合、曖昧語込み複合語の曖昧語から末尾までの文字列からなる単語部分のことであり、曖昧語が接尾語の場合は曖昧語込み複合語の先頭から曖昧語までの文字列からなる単語部分のことをいう。曖昧語込み複合語と曖昧語付き表現を定義済みの可能性が高い表現とみなせる理由は、文書内での

独自の複合語の生成パターンに起因する。すなわち、曖昧語付き表現が独自に意味を定義されていない状況で、曖昧語付き表現にさらに別の単語を組み合わせた曖昧語込み複合語が作成されることは、非常に稀であるためである。なお、曖昧語込み複合語および曖昧語付き表現を全て定義済みとみなしても、通常は高精度

5 に定義済みの単語の推定が行えるが、曖昧語込み複合語および曖昧語付き表現の形態素数か文字数、品詞の種類を利用して定義済みの可能性の強弱を示す定義度を推定し、定義度が所定の閾値を上回った場合に定義済みであると判定しても良い。ここで、定義度は曖昧語付き表現の形態素数、文字数、品詞の種類のいずれか、もしくはその組合せという語の構成条件を基に算出される。

- 10 出力部50は、定義済みと判定した曖昧語込み複合語や曖昧語付き表現を定義済み表現として出力する。定義度を算出した場合は定義済み表現と、定義度を対応づけて出力してもよい。

次に、図2のシーケンス図を参照して、本発明の実施形態に係る定義済み表現抽出システム100の全体の動作について詳細に説明する。なお、図2に示すシーケンス図および以下の説明は処理例であり、適宜求める処理に応じて処理順等

15 を入れ替えたり処理を戻したり繰り返したりすることを行っても良い。

文書入力部10は、曖昧語を含む可能性が有り優先的な修正が必要な曖昧性の高い箇所を分析する文書もしくは文書群の入力を受け付ける（図2のステップS100）。

- 20 検索部20は、曖昧語データベース110を参照して、文書中から曖昧語を検索し、曖昧語と曖昧語を含む文を取得する（ステップS101）。

複合語抽出部30は、曖昧語を含む文を形態素解析し、曖昧語と隣接して名詞が連続する部分を曖昧語込み複合語として取得する。その際、曖昧語が先頭、末尾にある場合は除く（ステップS102）。

- 25 定義済み表現候補抽出部40は、曖昧語込み複合語に含まれる曖昧語が接頭語の場合は曖昧語込み複合語中の曖昧語とその後ろに続く語を曖昧語付き表現として抽出し、曖昧語が接尾語の場合は、曖昧語込み複合語の先頭から曖昧語までの部分を曖昧語付き表現として抽出する（ステップS103）。そして、曖昧語込み複合語や曖昧語付き表現の文字数、形態素数、品詞の種類のいずれか、もしくは

はその組合せという語の構成条件により、定義度を算出する(ステップS104)。

出力部50は、曖昧語込み複合語や曖昧語付き表現と定義度を組にしたもの、あるいは、定義度が所定の閾値を上回ったものを定義済み表現一覧として出力する(ステップS105)。

- 5 次に、本発明の実施形態に係る定義済み表現抽出システム100の効果について説明する。

本実施形態に係る定義済み表現抽出システム100は、文書において曖昧語を含む曖昧語込み複合語と曖昧語付き表現を抽出し、曖昧語込み複合語や曖昧語付き表現の文字数、品詞の種類のいずれか、もしくはその組合せという語の構成条件により、曖昧語込み複合語や曖昧語付き表現が定義済みである可能性を定義度として算出し、算出した定義度に基づき曖昧語込み複合語と曖昧語付き表現の曖昧性を評価する。これにより、本実施形態に係る定義済み表現抽出システム100は、曖昧語を含む文書について、独自用語や業界用語、法律用語、定型表現のように語義が定義済みで曖昧性が低いと考えられるパターンを検出して除外し、
10 語義が未定義で曖昧性が高いと考えられるパターンを優先的な修正が必要な曖昧性の高い箇所として抽出することで文書の曖昧性を評価することができる。

なお、本実施形態に係る定義済み表現抽出システム100は、定義済み表現抽出方法として実現され得る。また、本実施形態に係る定義済み表現抽出システム100は、文書分析プログラムによりコンピュータによって実行させるようにしても良い。
20

[実施例]

次に、具体的な実施例を用いて、本発明の実施形態に係る定義済み表現抽出システム100の動作について説明するが、以下の動作例に限定されるものではない。

- 25 図3は、曖昧語を含む可能性が有り優先的な修正が必要な曖昧性の高い箇所を分析する文書Dの例を示す。このとき、曖昧語データベース110には図4に示す曖昧語とその属性がペアで登録されているとする。

まず、文書入力部10が文書Dの入力を受け付け、検索部20が曖昧語データベース110に含まれる曖昧語である「等」と「半」を文書Dから検索する。図

3で下線が引かれている部分が検索された曖昧語である。

次に、複合語抽出部30が、曖昧語を含む複合語を抽出する。ただし、複合語抽出部30は、曖昧語が複合語の先頭あるいは末尾にある場合は抽出しない。よって、複合語抽出部30は、図3の文書Dの中から「曖昧語半自動修正システム」と「仕様等読み取り部」を曖昧語込み複合語として抽出する。そして、定義済み表現候補抽出部40が、曖昧語込み複合語から曖昧語付き表現を抽出する。「曖昧語半自動修正システム」という曖昧語込み複合語には、「半」という曖昧語が含まれており、「半」という曖昧語の属性は曖昧語データベース110によると接頭語であるため、「半」とその後の文字列である「半自動修正システム」が曖昧語付き表現となる。また、「仕様等読み取り部」という曖昧語込み複合語には、「等」という曖昧語が含まれており、「等」という曖昧語の属性は曖昧語データベース110によると接尾語であるため、「等」とその前の文字列である「仕様等」が曖昧語付き表現となる。

また、定義済み表現候補抽出部40は、曖昧語込み複合語や曖昧語付き表現の定義度を、文字数、品詞の種類の内いずれか、もしくはその組合せという語の構成条件により算出する。算出方法の一例を挙げると、定義度をA、曖昧語を除いた文字数を N_a 、曖昧語を除いた形態素数を N_b 、特定の品詞に当てはまる形態素数を N_r とし、a、bを定数とすると、以下の式(1)で求めることができる。

$$A = \frac{N_a}{a} \times \frac{N_b}{(1+N_r)b} \quad \dots \text{式(1)}$$

このとき、 $a=2$ 、 $b=1.5$ とし、 N_r は品詞が「接尾」、「接頭」、「サ変接続」の内いずれかの名詞に当てはまる数とすると、「仕様等」の定義度Aは以下の式(2)に示すとおり0.67となる。

$$A = \frac{2}{2} \times \frac{1}{(1+0) \times 1.5} = 0.67 \quad \dots \text{式(2)}$$

また、「半自動修正システム」の定義度Aは以下の式(3)に示すとおり、4となる。「修正」がサ変接続の名詞であるため、 $N_r=1$ になっている。

$$A = \frac{8}{2} \times \frac{3}{(1+1) \times 1.5} = 4 \quad \dots \text{式(3)}$$

この数式の場合、文字数が多く、特定の品詞に当てはまらない形態素数が多いほど定義度が高く、定義済みの可能性が高くなる。よって、「仕様等」は文字数も形態素数も少ないため、定義度が低く、定義済みの可能性が低いのに対し、「半自動修正システム」は文字数、形態素数が多いため、定義度が高く、定義済みの可能性が高い。

出力部50は、図5に示すとおり、定義済み表現の候補となる曖昧語込み複合語と曖昧語付き表現の一覧とそれに対応する定義度をペアとして出力しても良いし、図6に示すとおり、定義度が所定の閾値（例えば1）を上回った表現のみを定義済み表現として一覧を出力しても良い。

10 このように定義済み表現抽出システム100により、文書Dから予め曖昧語を含む定義済みの表現を抽出することで、曖昧語を抽出する際、特に問題の大きい曖昧な表現のみを抽出することができる。図3の文書Dの例では、曖昧語「等」、「半」を含む表現として、「曖昧語半自動修正システム」、「仕様等」、「半自動修正システム」、「仕様等読み取り部」、「Office文書等」がある。定義済み表現抽出システム100によって予め定義済み表現である「曖昧語半自動修正システム」、「半自動修正システム」、「仕様等読み取り部」を抽出し、曖昧な表現の候補から除外することで、「仕様等」、「Office文書等」を曖昧な表現として抽出することができる。

20 なお、上記説明において、実施形態の動作（処理）は、プログラム、ソフトウェア、又はコンピュータによって実行されることが可能な命令でコード化された、コンピュータ読み取り可能な記憶媒体を用いて実行されてもよい。記憶媒体には、光ディスク、フロッピー（登録商標）ディスク、ハードディスク等の可搬型の記録媒体が含まれることはもとより、ネットワークのようにデータを一時的に記録保持するような伝送媒体も含まれる。

25 上記実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

（付記1）

対象とする文書もしくは文書群を入力文書として受け付ける文書入力部と、
曖昧語とその属性を蓄積した曖昧語データベースと、

前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索し、曖昧語を含む文を抽出する検索部と、

抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出する複合語抽出部と、

- 5 抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出部と、
前記定義済み表現候補を出力する出力部と、

- 10 を備えたことを特徴とする定義済み表現抽出システム。

(付記 2)

前記定義済み表現候補抽出部が、前記定義済み表現候補の文字数、形態素数、特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を算出する、ことを特徴とする付記 1 に記載の定義済み表現抽出システム。

- 15 (付記 3)

前記定義済み表現候補抽出部が、前記定義度の算出方法として、前記定義済み表現候補の曖昧語以外の文字数、特定の品詞に当てはまらない形態素数が多いほど定義度が高くなる計算式を用いる、ことを特徴とする付記 2 に記載の定義済み表現抽出システム。

- 20 (付記 4)

前記定義済み表現候補抽出部における前記特定の品詞とは、接頭と接尾とサ変接続の役割をもつ名詞である、ことを特徴とする付記 2 又は 3 に記載の定義済み表現抽出システム。

(付記 5)

- 25 前記出力部が、前記定義済み表現候補と前記定義度を組にして出力する、ことを特徴とする付記 2 乃至 4 のいずれかに記載の定義済み表現抽出システム。

(付記 6)

前記出力部が、前記定義度が所定の閾値を上回った定義済み表現候補のみを出力する、ことを特徴とする付記 2 乃至 5 のいずれかに記載の定義済み表現抽出シ

ステム。

(付記 7)

曖昧語とその属性を蓄積した曖昧語データベースを用意し、

対象とする文書もしくは文書群の入力を入力文書として受け付け、

- 5 前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索することにより、曖昧語を含む文を抽出し、

抽出された曖昧語を含む文を形態素解析して、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出し、

抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、

- 10 抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、かつ抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出し、

前記定義済み表現候補を出力する、

ことを特徴とする定義済み表現抽出方法。

- 15 (付記 8)

前記定義済み表現候補を抽出するステップが、定義済み表現候補の文字数、形態素数、特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を算出する、ことを特徴とする付記 7 に記載の定義済み表現抽出方法。

(付記 9)

- 20 前記定義済み表現候補を抽出するステップが、定義度の算出方法として、定義済み表現候補の曖昧語以外の文字数、特定の品詞に当てはまらない形態素数が多いほど定義度が高くなる計算式を用いる、ことを特徴とする付記 8 に記載の定義済み表現抽出方法。

(付記 10)

- 25 前記定義済み表現候補を抽出するステップにおける前記特定の品詞とは、接頭と接尾とサ変接続の役割をもつ名詞である、ことを特徴とする付記 8 又は 9 に記載の定義済み表現抽出方法。

(付記 11)

前記出力するステップが、前記定義済み表現候補と前記定義度を組にして出力

する、ことを特徴とする付記 8 乃至 10 のいずれかに記載の定義済み表現抽出方法。

(付記 12)

- 5 前記出力するステップが、前記定義度が所定の閾値を上回った定義済み表現候補のみを出力する、ことを特徴とする付記 8 乃至 10 のいずれかに記載の定義済み表現抽出方法。

(付記 13)

コンピュータに、

- 10 対象とする文書もしくは文書群を入力文書として受け付ける文書入力処理と、
曖昧語とその属性を蓄積した曖昧語データベースを参照して、前記入力文書中の曖昧語の有無を検索し、曖昧語を含む文を抽出する検索処理と、

抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出する複合語抽出処理と、

- 15 抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、
抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出処理と、

前記定義済み表現候補を出力する出力処理と、

- 20 を実行させるためのプログラム。

(付記 14)

前記定義済み表現候補抽出処理が、前記定義済み表現候補の文字数、形態素数、特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を算出する、ことを特徴とする付記 13 に記載のプログラム。

- 25 (付記 15)

前記定義済み表現候補抽出処理が、前記定義度の算出方法として、前記定義済み表現候補の曖昧語以外の文字数、特定の品詞に当てはまらない形態素数が多いほど定義度が高くなる計算式を用いる、ことを特徴とする付記 13 又は 14 に記載のプログラム。

(付記 1 6)

前記定義済み表現候補抽出処理における前記特定の品詞とは、接頭と接尾とサ変接続の役割をもつ名詞である、ことを特徴とする付記 1 4 又は 1 5 に記載のプログラム。

5 (付記 1 7)

前記出力処理が、前記定義済み表現候補と前記定義度を組にして出力する、ことを特徴とする付記 1 4 乃至 1 6 のいずれかに記載のプログラム。

(付記 1 8)

10 前記出力処理が、前記定義度が所定の閾値を上回った定義済み表現候補のみを出力する、ことを特徴とする付記 1 4 乃至 1 6 のいずれかに記載のプログラム。

(付記 1 9)

付記 1 3 乃至 1 8 の何れかに記載の定義済み表現抽出プログラムを記録したコンピュータ読み取り可能な記憶媒体。

産業上の利用可能性

15 本発明によれば、ソフトウェアやシステムの開発における要件定義などの作業においてやり取りされる各種文書に関して、文書の曖昧さを、特に問題の大きい場所から優先的に修正することが可能になり、文書作成や文書レビューの効率化に繋がる。また、複数の読み手の間に異なる解釈が起きる状況などが減少し、手戻りの減少や顧客満足の向上などシステム開発の効率化に寄与する。

20 この出願は、2012年12月17日に出願された日本出願特願第2012-274396号を基礎とする優先権を主張し、その開示のすべてをここに取り込む。

符号の説明

25 10 文書入力部
20 検索部
30 複合語抽出部
40 定義済み表現候補抽出部
50 出力部

1 0 0 定義済み表現抽出システム

1 1 0 曖昧語データベース

D 文書

請 求 の 範 囲

〔請求項 1〕

対象とする文書もしくは文書群を入力文書として受け付ける文書入力部と、

5 曖昧語とその属性を蓄積した曖昧語データベースと、

前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索し、
曖昧語を含む文を抽出する検索部と、

抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を
曖昧語込み複合語として抽出する複合語抽出部と、

10 抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、
抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそ
れぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された
曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出部と、

前記定義済み表現候補を出力する出力部と、

15 を備えたことを特徴とする定義済み表現抽出システム。

〔請求項 2〕

前記定義済み表現候補抽出部が、前記定義済み表現候補の文字数、形態素数、
特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を
算出する、ことを特徴とする請求項 1 に記載の定義済み表現抽出システム。

20 〔請求項 3〕

前記定義済み表現候補抽出部が、前記定義度の算出方法として、前記定義済み
表現候補の曖昧語以外の文字数、特定の品詞に当てはまらない形態素数が多いほ
ど定義度が高くなる計算式を用いる、ことを特徴とする請求項 2 に記載の定義済
み表現抽出システム。

25 〔請求項 4〕

前記定義済み表現候補抽出部における前記特定の品詞とは、接頭と接尾とサ変
接続の役割をもつ名詞である、ことを特徴とする請求項 2 又は 3 に記載の定義済
み表現抽出システム。

〔請求項 5〕

前記出力部が、前記定義済み表現候補と前記定義度を組にして出力する、ことを特徴とする請求項2乃至4のいずれか1項に記載の定義済み表現抽出システム。

〔請求項6〕

- 5 前記出力部が、前記定義度が所定の閾値を上回った定義済み表現候補のみを出力する、ことを特徴とする請求項2乃至5のいずれか1項に記載の定義済み表現抽出システム。

〔請求項7〕

- 曖昧語とその属性を蓄積した曖昧語データベースを用意し、
- 10 対象とする文書もしくは文書群の入力を入力文書として受け付け、
前記曖昧語データベースを参照して前記入力文書中の曖昧語の有無を検索することにより、曖昧語を含む文を抽出し、
抽出された曖昧語を含む文を形態素解析して、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出し、
- 15 抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、
抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、かつ抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出し、
前記定義済み表現候補を出力する、
- 20 ことを特徴とする定義済み表現抽出方法。

〔請求項8〕

前記定義済み表現候補を抽出するステップが、定義済み表現候補の文字数、形態素数、特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を算出する、ことを特徴とする請求項7に記載の定義済み表現抽出方法。

- 25 〔請求項9〕

コンピュータに、
対象とする文書もしくは文書群を入力文書として受け付ける文書入力処理と、
曖昧語とその属性を蓄積した曖昧語データベースを参照して、前記入力文書中の曖昧語の有無を検索し、曖昧語を含む文を抽出する検索処理と、

抽出された曖昧語を含む文を形態素解析し、曖昧語の前後に隣接する名詞群を曖昧語込み複合語として抽出する複合語抽出処理と、

抽出された曖昧語込み複合語の曖昧語が接尾語の場合は先頭から曖昧語まで、抽出された曖昧語込み複合語の曖昧語が接頭語の場合は曖昧語から末尾までをそれぞれ曖昧語付き表現として抽出し、抽出された曖昧語込み複合語と抽出された曖昧語付き表現を定義済み表現候補として抽出する定義済み表現候補抽出処理と、

前記定義済み表現候補を出力する出力処理と、
を実行させるためのプログラム。

10 〔請求項 10〕

前記定義済み表現候補抽出処理が、前記定義済み表現候補の文字数、形態素数、特定の品詞に当てはまる形態素数のいずれか、またはその組合せから、定義度を算出する、ことを特徴とする請求項 9 に記載のプログラム。

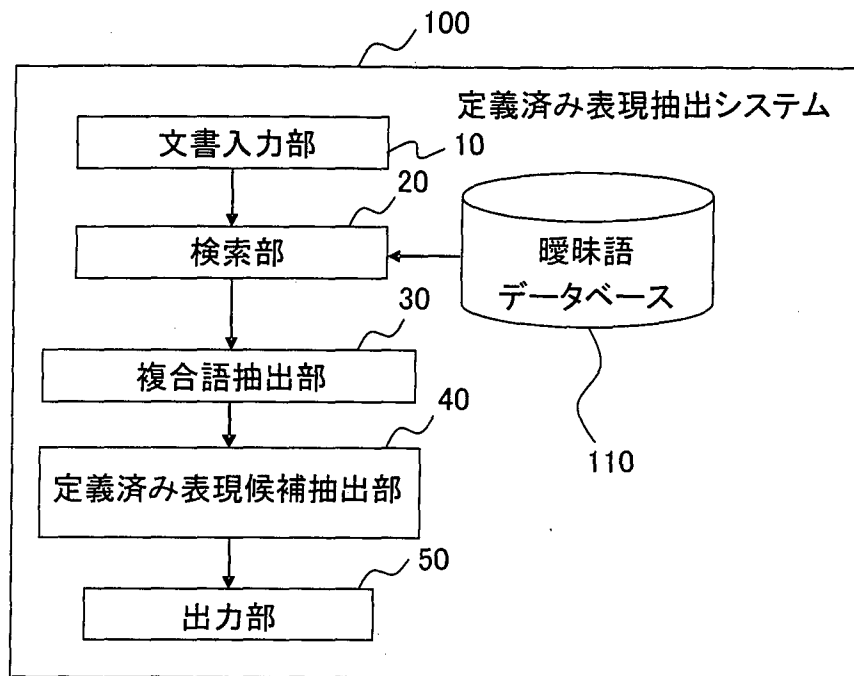


図 1

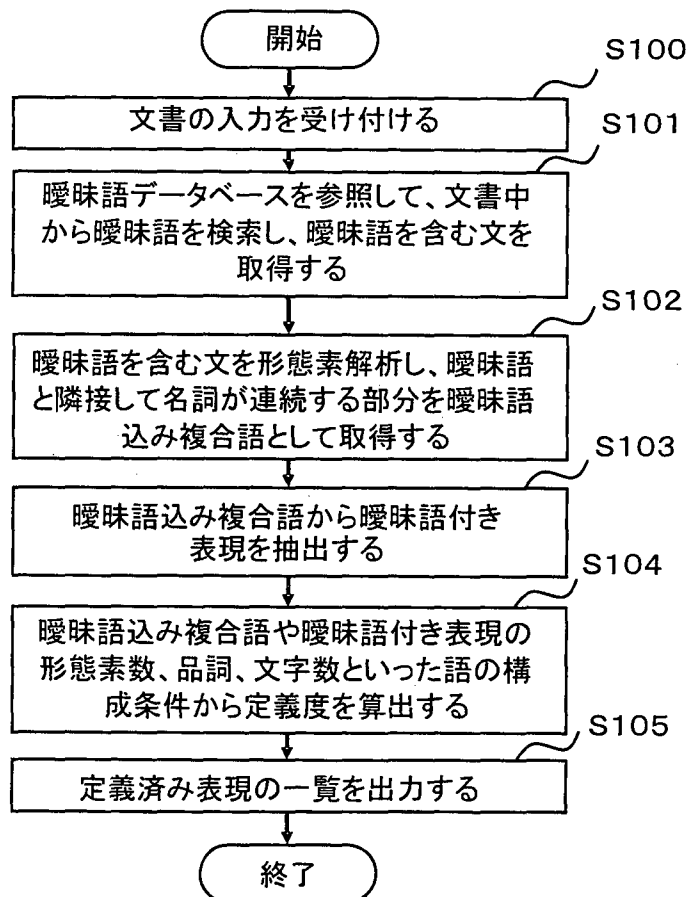


図 2

D

曖昧語半自動修正システムは、仕様等の文書中の曖昧語を自動的に検出し、修正候補を提示する。この半自動修正システムを用いることにより、レビューの工数を大幅に削減することができる。仕様等読み取り部では、テキストやOffice文書等を読み取り…

図 3

曖昧語	属性
半	接頭語
等	接尾語

図 4

定義済み表現候補	定義度
曖昧語半自動修正システム	6.11
仕様等	0.67
半自動修正システム	4.00
仕様等読み取り部	3.50

図 5

定義済み表現
曖昧語半自動修正システム
半自動修正システム
仕様等読み取り部

図 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2013/079479

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/27(2006.01) i, G06F17/21(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/20-17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2014
Kokai Jitsuyo Shinan Koho	1971-2014	Toroku Jitsuyo Shinan Koho	1994-2014

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 1-243116 A (Hitachi, Ltd.), 27 September 1989 (27.09.1989), entire text; all drawings (Family: none)	1-10
A	JP 2007-122660 A (Toshiba Corp.), 17 May 2007 (17.05.2007), paragraphs [0037] to [0047]; fig. 8 (Family: none)	1-10
A	Ayako KUNO et al., "Detection of Nocuous Ambiguities in Specifications", Proceedings of the 2012 IEICE General Conference, Joho System 1, 06 March 2012 (06.03.2012), page 27	1-10

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
07 January, 2014 (07.01.14)

Date of mailing of the international search report
21 January, 2014 (21.01.14)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/27(2006.01)i, G06F17/21(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/20-17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 4 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 4 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 4 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 1-243116 A（株式会社日立製作所）1989.09.27, 全文・全図（ファミリーなし）	1-10
A	JP 2007-122660 A（株式会社東芝）2007.05.17, 【0037】－【0047】段落, 【図8】（ファミリーなし）	1-10
A	久野綾子 他2名, 仕様書の曖昧性を検出するツールの試作と評価, 電子情報通信学会2012年総合大会講演論文集 情報・システム1, 2012.03.06, 27頁	1-10

☐ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

0 7 . 0 1 . 2 0 1 4

国際調査報告の発送日

2 1 . 0 1 . 2 0 1 4

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）
郵便番号100-8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

成瀬 博之

5 M

9 1 9 2

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 9 9