

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2017-201478

(P2017-201478A)

(43) 公開日 平成29年11月9日(2017.11.9)

(51) Int.Cl. F I テーマコード (参考)
G 0 6 F 17/30 (2006.01)
 G 0 6 F 17/30 3 5 0 A
 G 0 6 F 17/30 1 7 0 A

審査請求 未請求 請求項の数 12 O L (全 24 頁)

(21) 出願番号	特願2016-93227 (P2016-93227)	(71) 出願人	000004226
(22) 出願日	平成28年5月6日 (2016.5.6)		日本電信電話株式会社
			東京都千代田区大手町一丁目5番1号
		(74) 代理人	110001519
			特許業務法人太陽国際特許事務所
		(72) 発明者	大塚 淳史
			東京都千代田区大手町一丁目5番1号 日
			本電信電話株式会社内
		(72) 発明者	別所 克人
			東京都千代田区大手町一丁目5番1号 日
			本電信電話株式会社内
		(72) 発明者	平野 徹
			東京都千代田区大手町一丁目5番1号 日
			本電信電話株式会社内
		最終頁に続く	

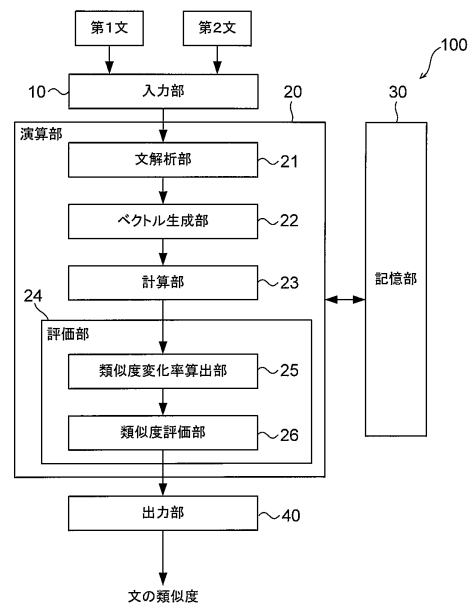
(54) 【発明の名称】 キーワード評価装置、類似度評価装置、検索装置、評価方法、検索方法、及びプログラム

(57) 【要約】

【課題】 自然言語で記述された文に類似する文を精度よく検索するようにする。

【解決手段】 類似度評価装置100は、第1文に含まれる単語と第2文に含まれる単語との組み合わせの中で類似度が最も高い組み合わせにおける第1文に含まれる単語を第1のキーワード、第2文に含まれる単語を第2のキーワードとし、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算し、少なくとも2つの類似度の変化値の絶対値が小さいほど、第1文と第2文とが類似していると評価する。

【選択図】 図1



【特許請求の範囲】

【請求項 1】

入力された第 1 文から抽出された第 1 のキーワードと、入力された第 2 文の中で前記第 1 のキーワードと類似するキーワードである第 2 のキーワードと、に基づき、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも 2 つの類似度を計算する計算部と、

前記計算部で計算された前記少なくとも 2 つの類似度の変化値の絶対値が小さいほど、前記第 1 のキーワードの重要度を高く評価する評価部と、
を含むキーワード評価装置。

【請求項 2】

前記評価部は、前記計算部で計算された前記キーワード同士の類似度と前記文節同士の類似度との変化値の絶対値が小さいほど、前記第 1 のキーワードの重要度を高く評価し、又は、前記計算部で計算された前記文節同士の類似度と前記文節の係り受け同士の類似度との変化値の絶対値が小さいほど、前記第 1 のキーワードの重要度を高く評価する

請求項 1 記載のキーワード評価装置。

【請求項 3】

入力された第 1 文に含まれる単語と入力された第 2 文に含まれる単語との組み合わせの中で類似度が最も高い組み合わせにおける前記第 1 文に含まれる単語を第 1 のキーワード、前記第 2 文に含まれる単語を第 2 のキーワードとし、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち

少なくとも 2 つの類似度を計算する計算部と、
前記計算部で計算された前記少なくとも 2 つの類似度の変化値の絶対値が小さいほど、前記第 1 文と前記第 2 文とが類似していると評価する評価部と、

を含む類似度評価装置。

【請求項 4】

前記評価部は、前記計算部で計算された前記キーワード同士の類似度と前記文節同士の類似度との変化値の絶対値が小さいほど、前記第 1 文と前記第 2 文とが類似していると評価し、又は、前記計算部で計算された前記文節同士の類似度と前記文節の係り受け同士の類似度との変化値の絶対値が小さいほど、前記第 1 文と前記第 2 文とが類似していると評価する

請求項 3 記載の類似度評価装置。

【請求項 5】

前記計算部は、入力された前記第 1 文に含まれる単語の各々について、前記第 1 文に含まれる単語を第 1 のキーワード、前記第 1 文に含まれる単語との類似度が最も高い、前記第 2 文に含まれる単語を第 2 のキーワードとし、前記キーワード同士の類似度、前記キーワードを含む文節同士の類似度、及び前記キーワードを含む文節の係り受け同士の類似度を計算し、

前記評価部は、前記第 1 文に含まれる単語の各々について、前記第 1 のキーワードと前記第 2 のキーワードとの類似度と、前記キーワード同士の類似度と前記文節同士の類似度との変化値の絶対値と、前記文節同士の類似度と前記文節の係り受け同士の類似度との変化値の絶対値とに基づいて算出されるスコアの平均値に基づいて、前記第 1 文と前記第 2 文とが類似しているかを評価する

請求項 4 記載の類似度評価装置。

【請求項 6】

予め用意された複数の検索対象文毎に、前記検索対象文に含まれる各キーワードを表すキーワードベクトル、前記各キーワードについてのキーワードを含む文節を表す文節ベクトル、及び前記各キーワードについてのキーワードを含む前記文節の係り先を含む係り受け関係を表す係り受けベクトルを記憶する記憶部と、

前記複数の検索対象文毎に、入力されたクエリ文に含まれるキーワードと前記検索対象文に含まれるキーワードとの組み合わせの中で類似度が最も高い組み合わせにおける前記

10

20

30

40

50

クエリ文に含まれるキーワードを第1のキーワード、前記検索対象文に含まれるキーワードを第2のキーワードとし、キーワードベクトルに基づくキーワード同士の類似度、文節ベクトルに基づくキーワードを含む文節同士の類似度、及び係り受けベクトルに基づくキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を前記複数の検索対象文毎に計算する計算部と、

前記複数の検索対象文毎に、前記計算部で計算された前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記クエリ文と前記検索対象文とが類似していると評価する評価部と、

前記評価部による評価結果に基づいて、前記クエリ文に類似する検索対象文を検索する検索部と、

を含む検索装置。

【請求項7】

入力された第1文から抽出された第1のキーワードと、入力された第2文の中で前記第1のキーワードと類似するキーワードである第2のキーワードと、に基づき、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算し、

前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1のキーワードの重要度を高く評価する

キーワードの評価方法。

【請求項8】

入力された第1文に含まれる単語と入力された第2文に含まれる単語との組み合わせの中で類似度が最も高い組み合わせにおける前記第1文に含まれる単語を第1のキーワード、前記第2文に含まれる単語を第2のキーワードとし、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算し、

前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1文と前記第2文とが類似していると評価する

文の類似度の評価方法。

【請求項9】

予め用意された複数の検索対象文毎に、前記検索対象文に含まれる各キーワードを表すキーワードベクトル、前記各キーワードについてのキーワードを含む文節を表す文節ベクトル、及び前記各キーワードについてのキーワードを含む前記文節の係り先を含む係り受け関係を表す係り受けベクトルを記憶し、

前記複数の検索対象文毎に、入力されたクエリ文に含まれるキーワードと前記検索対象文に含まれるキーワードとの組み合わせの中で類似度が最も高い組み合わせにおける前記クエリ文に含まれるキーワードを第1のキーワード、前記検索対象文に含まれるキーワードを第2のキーワードとし、キーワードベクトルに基づくキーワード同士の類似度、文節ベクトルに基づくキーワードを含む文節同士の類似度、及び係り受けベクトルに基づくキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を前記複数の検索対象文毎に計算し、

前記複数の検索対象文毎に、計算した前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記クエリ文と前記検索対象文とが類似していると評価し、

前記評価に基づいて、前記クエリ文に類似する検索対象文を検索する

文の検索方法。

【請求項10】

請求項1又は請求項2記載のキーワード評価装置の各部としてコンピュータを機能させるためのプログラム。

【請求項11】

請求項3～請求項5の何れか1項に記載の類似度評価装置の各部としてコンピュータを機能させるためのプログラム。

10

20

30

40

50

【請求項 1 2】

請求項 6 記載の検索装置の各部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、音声又はテキストを入力とするキーワード評価装置、類似度評価装置、検索装置、評価方法、検索方法、及びプログラムに関する。

【背景技術】

【0002】

情報検索システムとして、ユーザが入力したクエリに対して、キーワードマッチ等の検索処理を行い、クエリに適合する文書を検索するシステムが知られている。しかし、キーワードマッチを用いた検索処理の場合には、クエリのキーワードと文書内のキーワードが完全一致していなくてはならず、検索の再現率（Recall）が低下してしまうという問題が発生することがある。そこで、クエリ中に含まれるキーワードを拡張して増加させ、ユーザが入力したクエリを、より幅広い文書にマッチさせるクエリ拡張が行われる場合がある（特許文献 1）。

【0003】

また、キーワードマッチ以外の検索手法として、概念検索が知られている。概念検索は、キーワードを連続値の N 次元ベクトルで表現し、N 次元ベクトルの重心をクエリベクトルと見なす。同様に、文書ベクトルも文書内のキーワードベクトルの重心で表現し、クエリベクトルと文書ベクトルの類似度を計算し、類似度が高い順に検索結果を出力することで、クエリに適合する文書の検索を実行する。概念検索ではキーワードマッチと異なり、キーワードが完全一致しなくてもクエリに近い話題に関する文書が検索可能になるという利点がある（特許文献 2）。

【先行技術文献】

【特許文献】

【0004】

【特許文献 1】特開 2010-123036 号公報

【特許文献 2】特開 2010-182041 号公報

【発明の概要】

【発明が解決しようとする課題】

【0005】

しかしながら、キーワードマッチ及び概念検索では、クエリをキーワードの集合（bag-of-words）とみなすことを前提としている。つまり、自然言語の形式でクエリが入力された場合、形態素解析やキーワード抽出により、自然言語からキーワードを抽出することで、クエリに近い話題を含む文書等の検索を行っている。

【0006】

したがって、自然言語からキーワードを抽出してキーワード集合を生成する過程で、自然言語が本来持っていた語順や構文情報が欠落してしまう状況が発生する場合がある。

【0007】

例えば、FAQ 検索において、「メールが送信できない」と「送信できないメールがある」という、自然言語で表現された 2 つのクエリが入力されたとする。この 2 つのクエリが表す意味は異なるが、2 つのクエリをキーワード集合に変換すると、共に「メール、送信、できない」という同じ要素を含むキーワード集合になってしまい、2 つのクエリの違いが区別できなくなってしまう。このように、自然言語による情報検索を行う際には、文に含まれるキーワードのみでなく、キーワード周辺の文脈を考慮しなければならないことがある。

【0008】

本発明は、上記の事情を鑑みて成されたものであり、自然言語で表される入力文に含まれるキーワードの重要度を精度よく評価することができるキーワード評価装置、キーワー

10

20

30

40

50

ドの評価方法、及びプログラムを提供することを目的とする。また、自然言語で表される入力文と、比較対象となる文と、の類似度を評価し、入力文に類似する文を精度よく検索することができる類似度評価装置、検索装置、類似度の評価方法、検索方法、及びプログラム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】

【0009】

上記の目的を達成するために本発明に係るキーワード評価装置は、入力された第1文から抽出された第1のキーワードと、入力された第2文の中で前記第1のキーワードと類似するキーワードである第2のキーワードと、に基づき、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算する計算部と、前記計算部で計算された前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1のキーワードの重要度を高く評価する評価部と、を含んで構成されている。

10

【0010】

本発明に係る類似度評価装置は、入力された第1文に含まれる単語と入力された第2文に含まれる単語との組み合わせの中で類似度が最も高い組み合わせにおける前記第1文に含まれる単語を第1のキーワード、前記第2文に含まれる単語を第2のキーワードとし、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算する計算部と、前記計算部で計算された前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1文と前記第2文とが類似していると評価する評価部と、を含んで構成されている。

20

【0011】

本発明に係る検索装置は、予め用意された複数の検索対象文毎に、前記検索対象文に含まれる各キーワードを表すキーワードベクトル、前記各キーワードについてのキーワードを含む文節を表す文節ベクトル、及び前記各キーワードについてのキーワードを含む前記文節の係り先を含む係り受け関係を表す係り受けベクトルを記憶する記憶部と、前記複数の検索対象文毎に、入力されたクエリ文に含まれるキーワードと前記検索対象文に含まれるキーワードとの組み合わせの中で類似度が最も高い組み合わせにおける前記クエリ文に含まれるキーワードを第1のキーワード、前記検索対象文に含まれるキーワードを第2のキーワードとし、キーワードベクトルに基づくキーワード同士の類似度、文節ベクトルに基づくキーワードを含む文節同士の類似度、及び係り受けベクトルに基づくキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を前記複数の検索対象文毎に計算する計算部と、前記複数の検索対象文毎に、前記計算部で計算された前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記クエリ文と前記検索対象文とが類似していると評価する評価部と、前記評価部による評価結果に基づいて、前記クエリ文に類似する検索対象文を検索する検索部と、を含んで構成されている。

30

【0012】

本発明に係るキーワードの評価方法は、入力された第1文から抽出された第1のキーワードと、入力された第2文の中で前記第1のキーワードと類似するキーワードである第2のキーワードと、に基づき、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算し、前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1のキーワードの重要度を高く評価する。

40

【0013】

本発明に係る文の類似度の評価方法は、入力された第1文に含まれる単語と入力された第2文に含まれる単語との組み合わせの中で類似度が最も高い組み合わせにおける前記第1文に含まれる単語を第1のキーワード、前記第2文に含まれる単語を第2のキーワードとし、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度のうち少なくとも2つの類似度を計算し、前記少なくとも2つの類似度の変化値の絶対値が小さいほど、前記第1文と前記第2文とが類似してい

50

ると評価する。

【0014】

本発明に係る文の検索方法は、予め用意された複数の検索対象文毎に、前記検索対象文に含まれる各キーワードを表すキーワードベクトル、前記各キーワードについてのキーワードを含む文節を表す文節ベクトル、及び前記各キーワードについてのキーワードを含む前記文節の係り先を含む係り受け関係を表す係り受けベクトルを記憶し、前記複数の検索対象文毎に、入力されたクエリ文に含まれるキーワードと前記検索対象文に含まれるキーワードとの組み合わせの中で類似度が最も高い組み合わせにおける前記クエリ文に含まれるキーワードを第1のキーワード、前記検索対象文に含まれるキーワードを第2のキーワードとし、キーワードベクトルに基づくキーワード同士の類似度、文節ベクトルに基づく

10

【0015】

本発明に係るキーワード評価装置のプログラムは、キーワード評価装置の各部としてコンピュータを機能させるためのプログラムである。

【0016】

本発明に係る類似度評価装置のプログラムは、類似度評価装置の各部としてコンピュータを機能させるためのプログラムである。

20

【0017】

本発明に係る検索装置のプログラムは、検索装置の各部としてコンピュータを機能させるためのプログラムである。

【発明の効果】

【0018】

以上説明したように、本発明のキーワード評価装置、キーワードの評価方法、及びプログラムによれば、自然言語で表される入力文に含まれるキーワードの重要度を精度よく評価することができる、という効果が得られる。

また、本発明の類似度評価装置、検索装置、類似度の評価方法、検索方法、及びプログラムによれば、自然言語で表される入力文と、比較対象となる文と、の類似度を評価し、入力文に類似する文を精度よく検索することができる、という効果が得られる。

30

【図面の簡単な説明】

【0019】

【図1】第1実施形態に係る類似度評価装置の構成例を示す概略図である。

【図2】係り受け解析の実行結果の一例について説明する図である。

【図3】類似度評価装置における類似度評価処理ルーチンの処理の一例を示すフローチャートである。

【図4】類似度評価処理ルーチンを実行した場合におけるスコアの算出過程の一例を示す図である。

40

【図5】第2実施形態に係るキーワード評価装置の構成例を示す概略図である。

【図6】キーワード評価装置におけるキーワード評価処理ルーチンの処理の一例を示すフローチャートである。

【図7】第3実施形態に係る検索装置の構成例を示す概略図である。

【図8】検索装置における検索処理ルーチンの処理の一例を示すフローチャートである。

【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詳細に説明する。なお、以下では、同じ働きを担う構成要素又は処理には全図面を通して同じ符号を付与し、重複する説明を適宜省略する。

50

【0021】

<第1実施形態>

第1実施形態では、自然言語で記述された第1文及び第2文の2つの文を入力とし、2つの文の類似度を数値化してスコアとして出力する類似度評価装置100について説明する。

【0022】

<システム構成>

図1は、類似度評価装置100のシステム構成例を示す図である。図1に示すように、類似度評価装置100は、CPUと、RAMと、後述する類似度評価処理ルーチンを実行するためのプログラムを記憶したROMと、を備えたコンピュータで構成され、機能的には次に示すように構成されている。

10

【0023】

類似度評価装置100は、入力部10、演算部20、記憶部30、及び出力部40を備える。また、演算部20は、文解析部21、ベクトル生成部22、計算部23、及び評価部24を含むと共に、評価部24は、類似度変化率算出部25及び類似度評価部26を含む。

【0024】

入力部10に入力される第1文及び第2文は、文解析部21に出力される。文解析部21では、第1文及び第2文の各々に対して、キーワード抽出及び文節の係り受け解析といった言語処理を行い、第1文と第2文の言語構造を解析する。

20

【0025】

ベクトル生成部22は、文解析部21で行われた第1文と第2文の言語構造の解析結果を入力として、各々の文について、キーワード、文節、及び文節の係り受けといった異なる表現単位毎に概念ベクトルを生成する。

【0026】

そして、計算部23は、ベクトル生成部22で生成された各表現単位毎の概念ベクトルを入力として、キーワード同士、文節同士、及び文節の係り受け同士の類似度を計算する。

【0027】

評価部24は、計算部23で計算されたキーワード同士、文節同士、及び文節の係り受け同士の類似度を入力として、類似度変化率算出部25で類似度変化率を算出し、算出した類似度変化率に基づいて、類似度評価部26で第1文と第2文の最終的な類似度を数値化してスコアとして評価する。

30

【0028】

出力部40は、評価部24で評価されたスコアを出力して、第1文と第2文の類似度を通知する。

【0029】

なお、以降では、キーワード、文節、及び文節の係り受けといった各々の表現単位の粒度について述べる場合がある。キーワードに比べて文節の方が表現単位の粒度が粗く、文節に比べて文節の係り受けまで含めた範囲の方が表現単位の粒度が粗いため、キーワードの粒度が最も細かく、文節の係り受けの粒度が最も粗くなる。

40

【0030】

次に、演算部20の処理内容について詳細に説明する。

【0031】

<文解析部>

文解析部21は、第1文及び第2文の各文に対して言語処理を実行して、係り受け解析及びキーワード抽出を行う。

【0032】

文解析部21は、第1文及び第2文の各々の文に対して、例えば係り受け解析器を用いて、文中の形態素、各形態素の品詞、文節情報（文節数、文節に含まれる形態素）、文節

50

間の係り受け関係等の言語構造に関する情報を取得する。なお、文解析部 2 1 で用いる係り受け解析器に制限はなく、一例として上記に示した言語構造に関する情報を取得することができるものであれば、どのような係り受け解析器を用いてもよい。

【0033】

図 2 は、文解析部 2 1 での係り受け解析の実行結果の一例を示す図であり、例えば第 2 文として「メールが送信できない」を受け付けた場合、「メールが」が文節 1、「送信できない」が文節 2 であり、文節 1 の係り先が文節 2 であることが示される。また、各々の文節は形態素毎に分類されて各形態素の品詞が示される。

【0034】

更に、文解析部 2 1 は、係り受け解析の結果に基づいて、キーワード抽出ルールに従って、各々の文からキーワードを抽出する。なお、文解析部 2 1 で使用するキーワード抽出ルールに特に制限はなく、文同士の類似性の評価目的に応じて自由に規定することができる。一般的には、名詞及び動詞等の自立語となる単語をキーワードとして文から抽出することが好ましい。

【0035】

文解析部 2 1 は、第 1 文及び第 2 文から取得した言語構造に関する情報、並びに抽出したキーワードを、各々の文と対応付けて管理する。

【0036】

<ベクトル生成部>

ベクトル生成部 2 2 は、文解析部 2 1 での第 1 文及び第 2 文の解析結果に基づいて、文同士の類似度を評価するための概念ベクトルを作成する。

【0037】

ベクトル生成部 2 2 は、まず各々の文に含まれるキーワードの概念ベクトルであるキーワードベクトルを生成する。また、ベクトル生成部 2 2 は、キーワードベクトルを合成することにより、文節及び係り受けのそれぞれの概念ベクトルである文節ベクトル及び係り受けベクトルを生成する。

【0038】

キーワードベクトルを生成するためには、予め概念ベクトルモデルを生成しておく必要がある。概念ベクトルモデルは、概念ベクトル生成用の文書集合を用意し、概念ベクトル生成手法を用いて生成することができる。

【0039】

なお、ベクトル生成部 2 2 で用いる概念ベクトル生成用の文書集合、及び概念ベクトル生成手法に特に制限はない。概念ベクトル生成用の文書集合には、例えば第 1 文として入力される可能性のある文の集合（入力文集合）と内容が重複する Wikipedia（登録商標）のページ集合を用いてもよく、また、例えば入力文集合から抽出したキーワードを用いて Web 検索を行った場合の検索結果に含まれる Web のページ集合等、任意の文書集合を用いることができる。

【0040】

また、概念ベクトル生成手法についても、例えば特異値分解を用いた潜在意味インデックス（Latent Semantic Indexing: LSI）、トピックモデル、及びニューラルネットワークを用いたモデル等、任意の概念ベクトル生成モデルを用いることができる。この際、ベクトル生成部 2 2 で用いる概念ベクトル生成モデルでは、文に含まれるキーワードのみならず、助詞等全ての形態素のベクトルも生成するようにする。

【0041】

そして、ベクトル生成部 2 2 は、文節に含まれるキーワードのキーワードベクトル、及び当該文節に含まれるキーワード以外の形態素のベクトルを合成して、キーワードを含む文節の文節ベクトルを第 1 文及び第 2 文の各文について生成する。

【0042】

なお、ベクトルの合成方法に制限はなく、重心ベクトルによる合成方法の他、Recursive AutoEncoder (RAE) 等のニューラルネットワークを用いたベクトルの合成方法を用いても

10

20

30

40

50

よい。

【0043】

具体的には、重心ベクトルによる合成方法を用いる場合、ベクトル生成部22は、文節内に含まれるキーワードを含む全ての形態素のベクトルの重心を計算し、当該重心を表すベクトルを、キーワードを含む文節の文節ベクトルとする。また、RAEによる合成方法を用いる場合、ベクトル生成部22は、まず、キーワードを含む文節の先頭から数えて1番目及び2番目に現われる2つの形態素のベクトルを合成し、当該合成した形態素のベクトルを、今度は3番目に現われる形態素のベクトルと合成する。以降、ベクトル生成部22は、形態素の語順に従って、合成した形態素のベクトルと、次に現われる形態素のベクトルと、を順次合成する処理を、文節に含まれる全ての形態素のベクトルを合成するまで繰り返し、最終的に合成された形態素のベクトルを、キーワードを含む文節の文節ベクトルとする。

10

【0044】

次に、ベクトル生成部22は、文節の係り受け関係を表す係り受けベクトルを生成する。ここで文節の係り受け関係とは、2つの文節間の依存関係を表す。したがって、ベクトル生成部22は、文節の係り元と係り先における2つの文節の文節ベクトルを合成することによって係り受けベクトルを生成する。当該文節ベクトルの合成方法については、文節ベクトルの生成時と同様に制限はなく、重心ベクトルによる合成方法の他、RAE等のニューラルネットワークを用いたベクトルの合成方法といった任意のベクトルの合成方法を用いることができる。

20

【0045】

日本語における文節の係り受けの場合、単方向の係り受け関係となるため、述部に相当する文節以外の文節には、係り先となる文節が1つ存在する。

【0046】

例えば、「横浜で赤い帽子を買った」という文では、文節「赤い」の係り先の文節は「帽子を」となるため、ベクトル生成部22は、文節「赤い」に対応する文節ベクトルと、文節「帽子を」に対応する文節ベクトルとを合成して、「赤い帽子を」に対応する係り受けベクトルを生成する。

【0047】

また、文節「横浜で」及び文節「帽子を」の係り先の文節は、共に文節「買った」となるため、ベクトル生成部22は、それぞれ「横浜で買った」に対応する係り受けベクトルと、「帽子を買った」に対応する係り受けベクトルを生成する。

30

【0048】

なお、述部に相当する文節「買った」については、係り先となる文節が存在しないため、ベクトル生成部22は、文節「買った」に対応する文節ベクトルをそのまま係り受けベクトルとして用いる。

【0049】

以上の処理により、ベクトル生成部22は、第1文及び第2文の各文について、キーワードベクトル群、キーワードを含む文節の文節ベクトル群、及びキーワードを含む文節の係り受け関係を表す係り受けベクトル群を生成し、各々の文と対応づける。

40

【0050】

<計算部>

計算部23は、ベクトル生成部22で生成したキーワードベクトル、文節ベクトル、及び係り受けベクトルに基づいて、キーワード同士の類似度、文節同士の類似度、及び文節の係り受け同士の類似度を計算する。この際、第1文と第2文のどちらの文を、類似度の評価を行いたい基準の文にするかによって類似度の計算結果が異なるが、前述したように、ここでは第1文を基準の文にして類似度を計算する。

【0051】

まず、計算部23は、基準となる第1文に含まれるキーワードを1つ選択する。なお、第1文に含まれるキーワードは、文解析部21によって既に第1文から抽出されている。

50

そして、計算部 23 は、第 1 文から選択したキーワード（注目キーワード）と、第 2 文に含まれる全てのキーワードとの類似度を、各々のキーワードに対応するキーワードベクトルを用いて計算し、第 2 文に含まれるキーワードのうち、注目キーワードと最も類似度が高くなるキーワード（対応キーワード）を選択する。

【0052】

なお、計算部 23 で用いるキーワード同士の類似度の計算方法は、計算した類似度を示す値が 0 以上 1 以下の範囲の値を取るように正規化されるものであれば、どのような計算方法を用いてもよく、例えばコサイン距離等を用いることができる。類似度を示す値が“0”の場合は、キーワード同士が類似していないことを示し、類似度を示す値が大きくなるほど、キーワード同士の類似度が高いことを示す。そして、類似度を示す値が“1”の場合は、キーワード同士の類似度が最大であることを示す。

10

【0053】

次に、計算部 23 は、第 1 文から選択した注目キーワードを含む文節と、注目キーワードと最も類似度が高い対応キーワードを含む第 2 文の文節と、の類似度を、各々の文節に対応する文節ベクトルを用いて、例えばキーワード同士の類似度を計算する際に用いた計算方法と同じ計算方法で計算する。

【0054】

なお、第 2 文において、対応キーワードを含む文節が複数存在する場合には、計算部 23 は、注目キーワードを含む文節と、対応キーワードを含む全ての文節と、の類似度を各々計算し、類似度を示す値が最も大きくなる文節同士の組み合わせを選択する。

20

【0055】

そして、計算部 23 は、注目キーワードを含む文節の係り受け関係と、対応キーワードを含む文節の係り受け関係と、の類似度を、各々の係り受け関係に対応する係り受けベクトルを用いて、例えばキーワード同士の類似度を計算する際に用いた計算方法と同じ計算方法で計算する。

【0056】

すなわち、計算部 23 は、第 1 文から選択した注目キーワードに対して、第 2 文に含まれる対応キーワードとの間のキーワード同士の類似度、注目キーワードを含む文節と対応キーワードを含む文節との間の文節同士の類似度、及び注目キーワードを含む文節の係り受け関係と対応キーワードを含む文節の係り受け関係との間の文節の係り受け同士の類似度の 3 種類の類似度を計算する。

30

【0057】

そして、計算部 23 は、第 1 文に含まれる全てのキーワードの各々を注目キーワードとして順次選択し、第 1 文に含まれる各々のキーワードに対して、上記に示した 3 種類の類似度を計算する。

【0058】

<評価部>

評価部 24 は、計算部 23 で第 1 文に含まれる各々のキーワードに対して計算した 3 種類の類似度に基づいて、第 1 文と第 2 文との類似度を評価する。

【0059】

具体的には、まず、類似度変化率算出部 25 において、3 種類の類似度における類似度変化率を算出する。

40

【0060】

ここで類似度変化率とは、キーワードから文節、文節から文節の係り受け関係へと、第 1 文と第 2 文との類似判定単位の粒度を粗くした場合における、類似度の変化を示す尺度である。

【0061】

例えば、キーワード同士の類似度に対して文節同士の類似度が低下するほど、第 1 文と第 2 文とはキーワードレベルでは類似しているが、キーワードの周辺を含めた文節の単位では類似していないことを示すことになる。すなわち、キーワードレベルより粗い単位で

50

見た場合、キーワード同士の類似度で表されるほど、お互いの文節は類似していないことを示している。

【0062】

換言すれば、キーワード同士の類似度に対する、キーワードを含む文節まで拡張した文節同士の類似度の低下の度合いが少ないほど、第1文と第2文とは、キーワードレベルでの類似性をキーワードの周辺を含めた文節の単位でも維持していることを示すことになる。

【0063】

同様に、文節同士の類似度に対して、文節の係り受け同士の類似度が低下するほど、第1文と第2文とは文節レベルでは類似しているが、文節の係り受け関係まで含めた単位では類似していないことを示すことになる。すなわち、文節より粗い単位で見た場合、文節同士の類似度で表されるほど、お互いの文節の係り受け関係は類似していないことを示している。

【0064】

換言すれば、文節同士の類似度に対する、文節の係り受け同士の類似度の低下の度合いが少ないほど、第1文と第2文とは、文節レベルでの類似性を文節の係り受け関係まで含めた単位でも維持していることを示すことになる。

【0065】

以上をまとめれば、より粗い粒度で類似度を比較しても類似度の低下の度合いが少ない文ほど、各々の文の類似性が高いといえることができる。

【0066】

したがって、類似度変化率は、上記に示した類似度の変化の状況と適合するように、例えば類似判定単位を粗くすることによって第1文と第2文との類似度が低下するほど、類似変化率の絶対値が大きくなるように設定される。具体的には、類似度変化率算出部25は、第1文に含まれる全てのキーワードに対する類似度変化率を、(1)式及び(2)式を用いて計算する。

【0067】

【数1】

$$d_{ws}(\text{word}_1) = \text{sim}(\text{seg}_1, \text{seg}_2) - \text{sim}(\text{word}_1, \text{word}_2) \quad \cdots (1)$$

$$d_{sd}(\text{word}_1) = \text{sim}(\text{dep}_1, \text{dep}_2) - \text{sim}(\text{seg}_1, \text{seg}_2) \quad \cdots (2)$$

【0068】

ここで、 $d_{ws}(\text{word}_1)$ は第1文に含まれる任意のキーワード word_1 について、キーワードから文節へ類似判定単位を粗くした場合の類似度変化率を示す。 word_2 はキーワード word_1 に対応する第2文の対応キーワードであり、 $\text{sim}(\text{word}_1, \text{word}_2)$ は、キーワード同士の類似度を示す。また、 seg_1 はキーワード word_1 を含む文節を示し、 seg_2 はキーワード word_2 を含む文節を示す。したがって、 $\text{sim}(\text{seg}_1, \text{seg}_2)$ は、対応する文節同士の類似度を示す。

【0069】

また、 $d_{sd}(\text{word}_1)$ は第1文に含まれる任意のキーワード word_1 について、キーワード word_1 を含む文節から、当該文節の係り受け関係へ類似判定単位を粗くした場合の類似度変化率を示す。 dep_1 はキーワード word_1 を含む文節の係り受け関係を示し、 dep_2 はキーワード word_2 を含む文節の係り受け関係を示す。したがって、 $\text{sim}(\text{dep}_1, \text{dep}_2)$ は、対応する文節の係り受け同士の類似度を示す。

【0070】

なお、(1)式及び(2)式では、より類似判定単位が粗い場合の類似度から類似判定単位が細かい場合の類似度を減算しているが、より類似判定単位が細かい場合の類似度から類似判定単位が粗い場合の類似度を減算して、 $d_{ws}(\text{word}_1)$ 及び $d_{sd}(\text{word}_1)$ を算出するようにしてもよい。

【0071】

類似度評価部26は、計算部23で計算したキーワード同士の類似度と、類似度変化率

10

20

30

40

50

算出部 25 で算出した類似度変化率とを用いて、第 1 文と第 2 文との類似度の程度をスコアとして表す。

【0072】

具体的には、類似度評価部 26 は、第 1 文と第 2 文との類似度の程度を示すスコア $SIM(S_1, S_2)$ を (3) 式を用いて計算する。

【0073】

【数 2】

$$SIM(S_1, S_2) = \frac{1}{N} \sum_{w \in S_1} \frac{\text{sim}(w, w'_{s_2})}{3} * [1.0 + (1.0 - |d_{ws}(w)|) * \{1.0 + (1.0 - |d_{sd}(w)|)\}] \cdots (3)$$

【0074】

ここで、 S_1 は第 1 文、 S_2 は第 2 文を表す。 w は第 1 文 S_1 に含まれるキーワードを表し、 N は第 1 文 S_1 に含まれるキーワードの個数を表す。また、 w'_{s_2} はキーワード w と最も類似度が高くなる第 2 文 S_2 の対応キーワードを表している。(3) 式からわかるように、スコア $SIM(S_1, S_2)$ は、0 以上 1 以下の範囲の値を取るように正規化され、スコア $SIM(S_1, S_2)$ が “1” に近づくほど、第 1 文 S_1 と第 2 文 S_2 が類似していることを示す。

【0075】

以上により、演算部 20 で第 1 文と第 2 文との類似度が算出される。

【0076】

<類似度評価装置の作用>

次に、第 1 実施形態に係る類似度評価装置 100 の作用について説明する。

【0077】

類似度評価装置 100 は、自然言語で記述された第 1 文及び第 2 文をそれぞれ入力部 10 で受け付けると、受け付けた第 1 文及び第 2 文を例えば記憶部 30 に格納する。そして、類似度評価装置 100 は、CPU で図 3 に示す類似度評価処理ルーチンを実行する。

【0078】

まず、ステップ S100 において、係り受け解析器を用いて、第 1 文及び第 2 文に対して係り受け解析を行い、各々の文から取得した言語構造に関する情報を取得する。そして、当該言語構造に関する情報に基づいて、第 1 文及び第 2 文の各々の文からキーワードを抽出し、取得した言語構造に関する情報及びキーワードを、各々の文と対応付けて記憶部 30 に格納する。

【0079】

ステップ S102 において、ステップ S100 で取得した、第 1 文及び第 2 文の各々の文に対応する言語構造に関する情報及びキーワードを参照して、予め定めた概念ベクトルモデルに基づいて、各々の文に含まれるキーワードの各々についてキーワードベクトルを生成する。

【0080】

そして、言語構造に関する情報に含まれる文節情報を参照し、文節に含まれるキーワードのキーワードベクトル、及び当該文節に含まれるキーワード以外の形態素のベクトルを合成することによって、第 1 文及び第 2 文の各々の文に対してキーワードを含む文節の各々の文節ベクトルを生成する。

【0081】

更に、言語構造に関する情報に含まれる文節の係り受け関係を参照し、係り受け関係を有する文節ベクトルを合成することによって、第 1 文及び第 2 文の各々の文に対して、文節の係り受け関係の各々の係り受けベクトルを生成する。

【0082】

なお、生成したキーワードベクトル、文節ベクトル、及び係り受けベクトルは、第 1 文及び第 2 文の各々の文と対応付けて記憶部 30 に格納する

【0083】

ステップ S104 において、ステップ S100 で第 1 文に対応付けられた未選択のキーワードを注目キーワードとして、記憶部 30 から 1 つ選択する。

10

20

30

40

50

【0084】

ステップS106において、ステップS104で選択した注目キーワードに対応するキーワードベクトルを記憶部30から取得する。そして、注目キーワードに対応するキーワードベクトルと、第2文に対応付けられた全てのキーワードベクトルと、のコサイン距離を各々計算して、注目キーワードに対応するキーワードベクトルと最もコサイン距離が短い第2文に対応付けられたキーワードベクトルで表されるキーワードを、対応キーワードとして取得する。この際、注目ベクトルと対応ベクトルとのコサイン距離を、キーワード同士の類似度として記憶部30に格納する。

【0085】

ステップS108において、ステップS100で取得した言語構造に関する情報を参照して、ステップS104で選択した注目キーワードを含む第1文の文節と、ステップS106で取得した対応キーワードを含む第2文の文節と、を取得する。

10

【0086】

そして、ステップS102で生成した、注目キーワードを含む第1文の文節に対応した文節ベクトルと、対応キーワードを含む第2文の文節に対応した文節ベクトルと、のコサイン距離を文節同士の類似度として計算し、計算した文節同士の類似度を記憶部30に格納する。

【0087】

なお、文節同士の類似度は、コサイン距離以外の指標で表してもよいことは言うまでもない。

20

【0088】

ステップS110において、ステップS100で取得した言語構造に関する情報を参照して、ステップS104で選択した注目キーワードを含む第1文の文節の係り受け関係と、ステップS106で取得した対応キーワードを含む第2文の文節の係り受け関係と、を取得する。

【0089】

そして、ステップS102で生成した、注目キーワードを含む第1文の文節の係り受け関係に対応した係り受けベクトルと、対応キーワードを含む第2文の文節の係り受け関係に対応した係り受けベクトルと、のコサイン距離を文節の係り受け同士の類似度として計算し、計算した文節の係り受け同士の類似度を記憶部30に格納する。

30

【0090】

なお、文節の係り受け同士の類似度は、コサイン距離以外の指標で表してもよいことは言うまでもない。

【0091】

ステップS104～S110の処理によって、注目キーワードと最も類似する第2文の対応キーワードとの間のキーワード同士の類似度、注目キーワードを含む文節と対応キーワードを含む文節との間の文節同士の類似度、及び、注目キーワードを含む文節の係り受け関係と対応キーワードを含む文節の係り受け関係との間の文節の係り受け同士の類似度が各々算出される。

【0092】

ステップS112において、ステップS104で第1文に含まれる全てのキーワードを選択したか否か判定し、まだステップS104で選択されていないキーワードが存在する場合には、ステップS104に移行する。そして、第1文に含まれるキーワードの中から未選択のキーワードがなくなるまで、第1文に含まれる未選択のキーワードを注目キーワードとして選択することを繰り返すことで、第1文に含まれる各キーワードに対して、キーワード同士の類似度、文節同士の類似度、及び文節の係り受け同士の類似度が算出される。

40

【0093】

一方、ステップS112の判定処理が肯定判定の場合、すなわち、ステップS104で第1文に含まれる全てのキーワードを選択した場合には、ステップS114に移行する。

50

【0094】

ステップS114において、ステップS106で算出した、第1文に含まれるキーワード $word_1$ と、当該キーワードに対応する第2文の対応キーワード $word_2$ とのキーワード同士の類似度 $sim(word_1, word_2)$ と、ステップS108で算出した、当該各々のキーワードに対応する文節同士の類似度 $sim(seg_1, seg_2)$ と、に基づいて、上記(1)式に従って、類似度変化率 $d_{ws}(word_1)$ を第1文に含まれるキーワード $word_1$ 毎に算出する。

【0095】

また、ステップS108で算出した文節同士の類似度 $sim(seg_1, seg_2)$ と、ステップS110で算出した、当該各々の文節 seg_1 及び文節 seg_2 に対応する文節の係り受け同士の類似度 $sim(dep_1, dep_2)$ と、に基づいて、上記(2)式に従って、類似度変化率 $d_{sd}(word_1)$ を第1文に含まれるキーワード $word_1$ 毎に算出する。

10

【0096】

そして、ステップS116において、ステップS106で算出したキーワード同士の類似度 $sim(word_1, word_2)$ と、ステップS114で算出した類似度変化率 $d_{ws}(word_1)$ 及び $d_{sd}(word_1)$ と、に基づいて、上記(3)式に従って、第1文 S_1 と第2文 S_2 との類似度の程度を示すスコア $SIM(S_1, S_2)$ を算出する。算出したスコア $SIM(S_1, S_2)$ は、記憶部30に格納され、出力部40によって、例えばディスプレイ等の表示装置に、第1文 S_1 と第2文 S_2 とのスコア $SIM(S_1, S_2)$ が出力される。

【0097】

なお、(3)式では、スコア $SIM(S_1, S_2)$ の算出に類似度変化率 $d_{ws}(word_1)$ 及び $d_{sd}(word_1)$ を用いているが、例えば類似度変化率 $d_{ws}(word_1)$ 及び $d_{sd}(word_1)$ の少なくとも一方を用いてスコア $SIM(S_1, S_2)$ を算出するようにしてもよい。

20

【0098】

<類似度評価装置の実行結果>

図4は、「PWの変更をしたい」を第1文、「パスワードを変えたらログインできない」を第2文とした場合の、第1実施形態に係る類似度評価装置100でのスコア $SIM(S_1, S_2)$ の算出過程の一例を示した図である。

【0099】

この場合、第1文のキーワードとして、例えば「PW」及び「変更」が抽出され、第2文のキーワードとして、例えば「パスワード」、「変える」、「ログイン」が抽出される。なお、「変える」は「変えたら」の標準表記である。

30

【0100】

キーワード同士の類似度を算出した場合、「PW」と最も類似度が高い第2文のキーワードは「パスワード」であり、類似度は0.90であった。また、「変更」と最も類似度が高い第2文のキーワードは「変える」であり、類似度は0.95であった。

【0101】

文節同士の類似度を算出した場合、第1文のキーワード「PW」を含む文節「PWの」と、第2文のキーワード「パスワード」を含む文節「パスワードを」との文節同士の類似度は0.75であった。また、第1文のキーワード「変更」を含む文節「変更を」と、第2文のキーワード「変える」を含む文節「変えたら」との文節同士の類似度は0.32であった。

40

【0102】

更に、文節の係り受け同士の類似度を算出した場合、第1文の文節「PWの」の係り先を含めた係り受け関係「PWの変更を」と、第2文の文節「パスワードを」の係り先を含めた係り受け関係「パスワードを変えたら」との類似度は0.15であった。また、第1文の文節「変更を」の係り先を含めた係り受け関係「変更をしたい」と、第2文の文節「変えたら」の係り先を含めた係り受け関係「変えたらログインできない」との類似度は0.04であった。

【0103】

更に、キーワード「PW」に対する類似度変化率 $d_{ws}(PW)$ は、(1)式から“-0.15”となり、類似度変化率 $d_{sd}(PW)$ は、(2)式から“-0.60”となる。また、キーワード「

50

変更」に対する類似度変化率 d_{ws} (変更)は、(1)式から“-0.63”となり、類似度変化率 d_{sd} (変更)は、(2)式から“-0.28”となる。

【0104】

したがって、第1文「PWの変更をしたい」と第2文「パスワードを変えたらログインできない」との類似度の程度を示すスコアSIM(PWの変更をしたい, パスワードを変えたらログインできない)は、(3)式から“0.59”となる。

【0105】

このように第1実施形態に係る類似度評価装置100は、自然言語で記述された2つの文の類似度を評価する場合、各々の文に含まれるキーワード同士の類似性の比較だけではなく、キーワードが含まれる文節及び文節の係り受け表現の意味といった、語順や構文の類似性まで考慮して、2つの文の類似性を評価する。

10

【0106】

したがって、類似度評価装置100は、文に含まれるキーワード同士の類似性のみによって各々の文の類似性を評価する従来の類似度評価装置と比較して、比較対象となる文に対する類似度を精度よく評価することができる。

【0107】

例えば、第1文「メールが送信できなくなった」と、第2文A「メールが送信できない」及び第2文B「送信できないメールがある」と、を各々比較した場合、キーワード「メール」だけに着目すると、第2文A及び第2文B共にキーワード「メール」を含むため、第1文と第2文Aの類似度、及び第1文と第2文Bの類似度の間に違いはない。

20

【0108】

また、類似性の判定単位を文節まで広げた場合も、第1文の文節「メールが」は、第2文Aにも第2文Bにも含まれるため、第1文と第2文Aの類似度、及び第1文と第2文Bの類似度の間に違いはない。

【0109】

しかしながら、類似性の判定単位を文節の係り受け関係まで広げた場合、第1文の「メールが送信できなくなった」に対して、第2文Aは「メールが送信できない」、第2文Bは「メールがある」となる。したがって、第1文は第2文Bよりも第2文Aに類似していることがわかり、その評価値がスコアとして出力される。

【0110】

なお、第1実施形態に係る類似度評価装置100では、一例として、入力部10で第1文及び第2文をテキストとして受け付けるように説明した。しかし、例えば、入力部10で第1文及び第2文に対応する音声を受け付け、受け付けた音声に対して、音声をテキストに変換する公知の音声認識を行うことで、テキスト化された第1文及び第2文を取得するようにしてもよい。

30

【0111】

この場合、類似度評価装置100の入力インターフェースとして音声を用いることができるため、類似度を評価する内容を予めテキストにする必要がない。したがって、入力としてテキストを受け付ける場合に比べて、類似度評価装置100の操作性を向上することができる。

40

【0112】

<第2実施形態>

第1実施形態では、キーワード同士の類似度、キーワードを含む文節同士の類似度、及びキーワードを含む文節の係り受け同士の類似度に基づいて算出した類似度変化率の低下の度合いが少ない文同士ほど、各々の文の類似性が高いことを説明したが、換言すれば、これは、文に含まれるキーワードの中で、より粗い類似判定単位の粒度で類似度を比較しても類似度の低下の度合いが少ないキーワードほど、文の類似性の判定に与える影響が大きい重要なキーワードであることを示している。

【0113】

したがって、第2実施形態では、類似する第1文及び第2文の2つの文を入力して、第

50

1文に含まれる各々のキーワードの重要度を評価するキーワード評価装置200について説明する。

【0114】

<システム構成例>

図5は、キーワード評価装置200のシステム構成例を示す図である。図5のキーワード評価装置200のシステム構成が第1実施形態に係る図1の類似度評価装置100のシステム構成例と異なる点は、類似度評価部26がキーワード重要度評価部26Aに置き換えられ、それに伴い評価部24が評価部24Aに置き換えられた点である。

【0115】

なお、その他のキーワード評価装置200の構成は、類似度評価装置100のシステム構成例と同様である。

【0116】

キーワード重要度評価部26Aは、計算部23で計算したキーワード同士の類似度と、類似度変化率算出部25で算出した類似度変化率と、に基づいて、第1文に含まれるキーワード毎にキーワードの重要度を計算する。

【0117】

第2文 S_2 との類似性を判定する場合において、第1文 S_1 に含まれるキーワード w の重要度 $SIM(S_{1,w}, S_2)$ は、例えば(4)式で計算される。

【0118】

【数3】

$$SIM(S_{1,w}, S_2) = \frac{sim(w, w'_{s_2})}{3} * [1.0 + (1.0 - |d_{ws}(w)|)] * \{1.0 + (1.0 - |d_{sd}(w)|)\} \cdots (4)$$

【0119】

ここで、 w'_{s_2} は、(3)式で説明したように、第1文 S_1 に含まれるキーワード w と最も類似度が高い第2文 S_2 の対応キーワードを表している。また、(4)式からわかるように、重要度 $SIM(S_{1,w}, S_2)$ は、0以上1以下の範囲の値を取るように正規化され、重要度 $SIM(S_{1,w}, S_2)$ が“1”に近づくほど、キーワード w の重要度が高いことを示す。

【0120】

<キーワード評価装置の作用>

キーワード評価装置200は、類似する第1文及び第2文をそれぞれ入力部10で受け付けると、受け付けた第1文及び第2文を例えば記憶部30に格納する。そして、キーワード評価装置200は、CPUで図6に示すキーワード評価処理ルーチンを実行する。

【0121】

図6に示すキーワード評価処理ルーチンが、図3に示した第1実施形態に係る類似度評価装置100の類似度評価処理ルーチンと異なる点は、ステップS116がステップS118に置き換えられた点であり、その他の処理は、類似度評価装置100の類似度評価処理ルーチンと同じである。したがって、以下ではステップS118の処理について説明する。

【0122】

ステップS118において、ステップS106で算出したキーワード同士の類似度 $sim(word_1, word_2)$ と、ステップS114で算出した類似度変化率 $d_{ws}(word_1)$ 及び $d_{sd}(word_1)$ と、に基づいて、上記(4)式に従って、第2文に対する第1文の類似度の判定において、判定結果に影響を与える度合いを示すキーワード w の重要度 $SIM(S_{1,w}, S_2)$ をキーワード w 毎に算出する。

【0123】

算出した重要度 $SIM(S_{1,w}, S_2)$ は記憶部30に格納され、出力部40によって、例えばディスプレイ等の表示装置にキーワード w 毎の重要度 $SIM(S_{1,w}, S_2)$ が出力される。

【0124】

なお、キーワード w の重要度 $SIM(S_{1,w}, S_2)$ を算出する(4)式は一例であり、キーワード同士の類似度と、文節同士の類似度と、の変化値の絶対値が小さいほど、重要度 $SIM(S_1$

10

20

30

40

50

$,w,S_2)$ を高く評価し、又は、文節同士の類似度と、文節の係り受け同士の類似度と、の変化値の絶対値が小さいほど、重要度 $SIM(S_1,w,S_2)$ を高く評価することができれば、(4)式の代わりに他の評価式を用いて重要度 $SIM(S_1,w,S_2)$ を算出してもよい。

【0125】

このように第2実施形態に係るキーワード評価装置200によれば、少なくとも2つの類似度変化率 $d_{ws}(w)$ 及び $d_{sd}(w)$ の絶対値が小さいほど、キーワード w の重要度 $SIM(S_1,w,S_2)$ を高く評価する。

【0126】

したがって、例えばキーワードを入力して文を検索する検索システム等において、キーワード評価装置200で得られた、より重要度の高いキーワードを優先的に入力すれば、
目的とする文を精度よく検索することができる。

10

【0127】

なお、キーワード評価装置200では、第1実施形態に係る類似度評価装置100と同様に、第1文及び第2文に対応する音声を受け付け、音声をテキストに変換する公知の音声認識を行うことで、テキスト化された第1文及び第2文を取得するようにしてもよい。

【0128】

<第3実施形態>

第1実施形態に係る類似度評価装置100、及び第2実施形態に係るキーワード評価装置200では、文同士の異なる粒度における類似度変化率 $d_{ws}(w)$ 及び $d_{sd}(w)$ を算出し、算出した類似度変化率 $d_{ws}(w)$ 及び $d_{sd}(w)$ に基づいて、文同士の類似度、又は、文同士の類似度の判定に用いられるキーワードの重要度を評価した。

20

【0129】

第3実施形態では、第1実施形態に係る類似度評価装置100、及び第2実施形態に係るキーワード評価装置200と同様に、類似度変化率 $d_{ws}(w)$ 及び $d_{sd}(w)$ を算出し、算出した類似度変化率 $d_{ws}(w)$ 及び $d_{sd}(w)$ に基づいて、複数の文の中から、第1文に最も類似した文を検索する検索装置300について説明する。

【0130】

<システム構成例>

図7は、検索装置300のシステム構成例を示す図である。図7の検索装置300のシステム構成例が第1実施形態に係る図1の類似度評価装置100のシステム構成例と異なる点は、類似度評価部26がクエリ文類似度評価部26Bに置き換えられ、それに伴い評価部24が評価部24Bに置き換えられた点である。更に、検索装置300には検索部27が追加され、記憶部30に検索対象文DB30Aが予め構築される。

30

【0131】

その他の検索装置300の構成は、類似度評価装置100のシステム構成例と同様である。

【0132】

検索対象文DB30Aには、複数の検索対象文が、上記第1実施形態と同様に生成されたキーワードベクトル、文節ベクトル、及び係り受けベクトルと対応付けられて予め格納されているものとする。

40

【0133】

ここで、検索対象文の各々に対応付けられるキーワードベクトル、文節ベクトル、及び係り受けベクトルは、例えば検索装置300に検索対象文を入力した場合におけるベクトル生成部22の出力結果を用いることができる。

【0134】

検索装置300では、検索対象文DB30Aに予め格納される複数の検索対象文の中から、自然言語で記述されたクエリ文の内容に最も類似する検索対象文を検索する。

【0135】

このように、検索装置300では、複数の検索対象文が予め検索対象文DB30Aに格納されているため、第1実施形態に係る類似度評価装置100、及び第2実施形態に係る

50

キーワード評価装置 200 とは異なり、クエリ文のみが入力部 10 に入力される。

【0136】

したがって、文解析部 21 は、類似度評価装置 100 と同様の手法によって、入力部 10 から受け付けたクエリ文に対して係り受け解析を実行し、係り受け解析の結果に基づいて、クエリ文からキーワードを抽出する。

【0137】

そして、ベクトル生成部 22 は、文解析部 21 でクエリ文から抽出したキーワードに基づいて、類似度評価装置 100 と同様の手法によってキーワードベクトル、文節ベクトル、及び係り受けベクトルを生成する。

【0138】

また、計算部 23 は、ベクトル生成部 22 で生成したクエリ文に対応するキーワードベクトル、文節ベクトル、及び係り受けベクトルと、検索対象文 DB 30A に格納される検索対象文に対応するキーワードベクトル、文節ベクトル、及び係り受けベクトルと、に基づいて、類似度評価装置 100 と同様の手法によってキーワード同士の類似度、文節同士の類似度、及び文節の係り受け同士の類似度を計算する。

【0139】

クエリ文類似度評価部 26B は、第 1 実施形態に係る類似度評価装置 100 の類似度評価部 26 と同様に、計算部 23 で計算したキーワード同士の類似度と、類似度変化率算出部 25 で算出した類似度変化率と、に基づいて、クエリ文 S_1 に対して、(3) 式のスコア $SIM(S_1, S_2)$ を検索対象文 S_2 毎に算出する。そして、クエリ文類似度評価部 26B は、例えばスコア $SIM(S_1, S_2)$ が最も “1” に近くなる検索対象文 S_2 を、クエリ文 S_1 の内容に類似する検索対象文 S_2 として評価する。

【0140】

しかしながら、検索装置 300 のように、クエリ文の内容に対応する検索対象文を検索する装置の場合、キーワード同士の類似度、文節同士の類似度、及び文節の係り受け同士の類似度が高いからといって、必ずしもクエリ文の内容に対応した適切な検索対象文が検索されるとは限られない。

【0141】

例えば、文中に頻繁に出現するキーワードは文の主題を表しやすい一方、複数の文に頻繁に出現するキーワードは重要なキーワードではないといった傾向が見られる。

【0142】

したがって、キーワード同士の類似度、文節同士の類似度、及び文節の係り受け同士の類似度に加えて、更に、文中におけるキーワードの重みを考慮することが好ましい。

【0143】

文中におけるキーワードの重み算出手法には、例えば Term Frequency-Inverse Document Frequency (TF-IDF) 法、又は BM25 法などの公知の手法が存在するが、クエリ文類似度評価部 26B には、こうした公知のキーワードの重み算出手法を適用することができる。

【0144】

したがって、クエリ文類似度評価部 26B は、公知のキーワードの重み算出手法を用いて算出したキーワードの重みを考慮した、クエリ文 S_1 と検索対象文 S_2 との類似度合いを示すスコア $Score(S_1, S_2)$ を、例えば (5) 式を用いて計算する。

【0145】

【数 4】

$$Score(S_1, S_2) = \frac{1}{N} \sum_{w \in S_1} SIM(S_{1,w}, S_2) * weight(\arg \max_{w \in S_2} (sim(w, w_{s_2}))) \cdots (5)$$

【0146】

ここで、 $SIM(S_{1,w}, S_2)$ は、(4) 式で表されるクエリ文 S_1 に含まれるキーワード w の重要度、すなわち、クエリ文 S_1 に含まれるキーワード w のみに着目した場合の、クエリ文 S_1 と検索対象文 S_2 との類似度である。また、 $weight$ は重み値を表す。したがって、 $weight(\arg \max_{w \in S_2} (sim(w, w_{s_2})))$ は、クエリ文 S_1 に含まれるキーワード w と最も類似度が高くなる検

10

20

30

40

50

索対象文 S_2 中のキーワード w_{s_2} の重み値である。

【0147】

検索部27は、クエリ文類似度評価部26Bで算出されたスコア $\text{Score}(S_1, S_2)$ を用いて、例えばスコア $\text{Score}(S_1, S_2)$ に関して予め定めた条件を満たす検索対象文 S_2 を検索対象文DB30Aから検索して、出力部40に出力する。

【0148】

なお、検索装置300が、例えばインターネット等のネットワークに接続される場合、検索対象文DB30Aをネットワークに接続される記憶装置等の外部装置に格納し、検索装置300が外部装置に格納された検索対象文DB30Aを参照するようにしてもよい。

【0149】

<検索装置の作用>

検索装置300は、自然言語で記述されたクエリ文を入力部10で受け付けると、受け付けたクエリ文を例えば記憶部30に格納する。そして、検索装置300は、CPUで図8に示す検索処理ルーチンを実行する。

【0150】

図8に示す検索処理ルーチンが、図3に示した第1実施形態に係る類似度評価装置100の類似度評価処理ルーチンと異なる点は、ステップS106の代わりにステップS107が追加された点である。また、検索処理ルーチンでは、ステップS103、ステップS120、及びS122が新たに追加される。なお、その他の処理は、類似度評価装置100の類似度評価処理ルーチンと同じである。したがって、以下では類似度評価処理ルーチンと異なる処理を中心にして、検索処理ルーチンを説明する。

【0151】

ステップS103において、検索対象文DB30Aに予め記憶されている複数の検索対象文のうち、未選択の検索対象文を1つ選択する。

【0152】

そして、ステップS107において、ステップS103で選択した検索対象文に対応付けられたキーワードの中から、ステップS104で選択した注目キーワードに最も類似する対応キーワードを抽出する。なお、キーワード同士の類似度は、図3におけるステップS106と同様に、例えば注目キーワードと対応キーワードとに対応する各々のキーワードベクトルのコサイン距離によって計算すればよく、計算したコサイン距離をキーワード同士の類似度として記憶部30に格納する。

【0153】

以降、ステップS108及びS110で、注目キーワードを含む文節と対応キーワードを含む文節同士の類似度、及び注目キーワードを含む文節の係り受けと対応キーワードを含む文節の係り受け同士の類似度を計算し、ステップS112でクエリ文に含まれる全てのキーワードについてステップS104～S112の処理を実行したか判定する。

【0154】

そして、ステップS114において、ステップS107で算出した、クエリ文に含まれるキーワード $word_1$ と、当該キーワードに対応する検索対象文の対応キーワード $word_2$ とのキーワード同士の類似度 $\text{sim}(word_1, word_2)$ と、ステップS108で算出した、当該各々のキーワードに対応する文節同士の類似度 $\text{sim}(\text{seg}_1, \text{seg}_2)$ と、に基づいて、上記(1)式に従って、類似度変化率 $d_{ws}(word_1)$ をクエリ文に含まれるキーワード毎に算出する。

【0155】

また、ステップS108で算出した文節同士の類似度 $\text{sim}(\text{seg}_1, \text{seg}_2)$ と、ステップS110で算出した、当該各々の文節 seg_1 及び文節 seg_2 に対応する文節の係り受け同士の類似度 $\text{sim}(\text{dep}_1, \text{dep}_2)$ と、に基づいて、上記(2)式に従って、類似度変化率 $d_{sd}(word_1)$ をクエリ文に含まれるキーワード毎に算出する。

【0156】

次に、ステップS116において、ステップS107で算出したキーワード同士の類似度 $\text{sim}(word_1, word_2)$ と、ステップS114で算出した類似度変化率 $d_{ws}(word_1)$ 及び $d_{sd}(wo$

10

20

30

40

50

rd_1)と、に基づいて、上記(4)式に従って、クエリ文 S_1 に含まれるキーワード w のみに着目した場合のクエリ文 S_1 と検索対象文 S_2 との類似度 $SIM(S_1, w, S_2)$ を、クエリ文 S_1 のキーワード毎に算出する。

【0157】

更に、本ステップで算出したキーワード w 毎の類似度 $SIM(S_1, w, S_2)$ と、クエリ文 S_1 に含まれるキーワード w と最も類似度が高くなる検索対象文 S_2 中のキーワード w_{s_2} の重み値と、に基づいて、上記(5)式に従って、スコア $Score(S_1, S_2)$ を算出する。算出したスコア $Score(S_1, S_2)$ は検索対象文 S_2 と対応付けて、例えば検索対象文DB30Aに格納する。

【0158】

なお、キーワード w_{s_2} の重み値は、前述したようにTF-IDF法等の公知の重み算出手法を用いて算出すればよい。

10

【0159】

ここでは一例として、(5)式に従ってクエリ文と検索対象文との類似度を算出したが、クエリ文と検索対象文との類似度の算出方法はこれに限られない。例えば、(3)式に従って類似度を算出してもよく、また、(4)式に従って算出した、クエリ文におけるキーワード毎の重要度 $SIM(S_1, w, S_2)$ の和を、クエリ文と検索対象文との類似度を示すスコアとしてもよい。

【0160】

ステップS120において、ステップS103で検索対象文DB30Aに含まれる全ての検索対象文を選択したか否か判定し、まだステップS103で選択されていない検索対象文が存在する場合には、ステップS103に移行する。

20

【0161】

そして、検索対象文DB30Aに含まれる検索対象文の中から未選択の検索対象文がなくなり、ステップS120の判定処理が肯定判定になるまでステップS103で検索対象文を繰り返し選択することで、ステップS116でクエリ文 S_1 に対する各検索対象文 S_2 のスコア $Score(S_1, S_2)$ が算出される。

【0162】

一方、ステップS120の判定処理が肯定判定となる場合には、ステップS122に移行する。

【0163】

ステップS122において、各々の検索対象文に対応付けられたスコア $Score(S_1, S_2)$ を参照し、予め定めた閾値以上のスコア $Score(S_1, S_2)$ が対応付けられた検索対象文を、検索対象文DB30Aから検索して取得する。そして、出力部40で、ステップS122で取得した検索対象文を、例えばディスプレイ等の表示装置に出力する。

30

【0164】

なお、ステップS122で取得する検索対象文は、予め定めた閾値以上のスコア $Score(S_1, S_2)$ が対応付けられた検索対象文に限られない。例えば、スコア $Score(S_1, S_2)$ の大きい方から順に予め定めた数の検索対象文を検索対象文DB30Aから検索して取得するようにしてもよい。

【0165】

なお、検索装置300では、入力部10でクエリ文に対応する音声を受け付け、受け付けた音声に対して、音声をテキストに変換する公知の音声認識を行うことで、テキスト化されたクエリ文を取得するようにしてもよい。

【0166】

この場合、検索装置300で受け付けた音声をそのまま検索に用いることができるため、入力としてテキストを受け付ける場合に比べて、検索装置300の操作性を向上することができる。

【0167】

このように第3実施形態に係る検索装置300は、クエリ文の内容に類似する検索対象文を、記憶部30に予め記憶される検索対象文DB30Aから検索する。この場合、検索

40

50

装置 300 は、クエリ文を受け付けた場合に、クエリ文に対してのみ係り受け解析器を用いて言語構造に関する情報を取得すると共に、クエリ文に含まれるキーワードに対応するキーワードベクトル、文節ベクトル、及び係り受けベクトルを生成する。すなわち、検索対象文に関するキーワードベクトル、文節ベクトル、及び係り受けベクトルは、検索対象文 DB 30A に予め格納されているため、クエリ文を受け付ける毎に検索対象文に関する各種ベクトルを生成する場合と比較して、高速にスコア $\text{Score}(S_1, S_2)$ を計算することができる。

【0168】

なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

10

【0169】

例えば、上述の類似度評価装置 100、キーワード評価装置 200、及び検索装置 300 は、内部にコンピュータシステムを有しているが、「コンピュータシステム」は、WWW システムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。

【0170】

また、本願明細書中において、プログラムが予め ROM にインストールされている実施形態を説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。

20

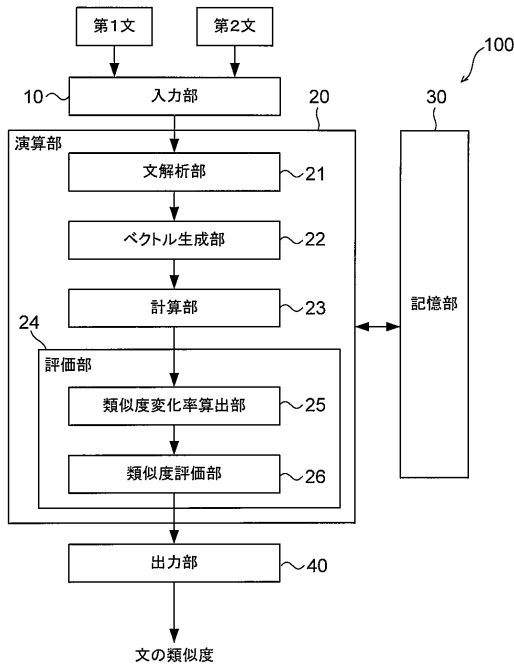
【符号の説明】

【0171】

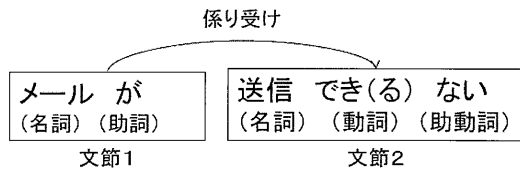
10・・・入力部
 20・・・演算部
 21・・・文解析部
 22・・・ベクトル生成部
 23・・・計算部
 24（24A、24B）・・・評価部
 25・・・類似度変化率算出部
 26・・・類似度評価部
 26A・・・キーワード重要度評価部
 26B・・・クエリ文類似度評価部
 27・・・検索部
 30・・・記憶部
 40・・・出力部
 100・・・類似度評価装置
 200・・・キーワード評価装置
 300・・・検索装置
 d_{sd} 、 d_{ws} ・・・類似度変化率
 30A・・・検索対象文 DB

30

【図 1】



【図 2】



【図 4】

第1文:「PWの変更をしたい」
第2文:「パスワードを変えたらログインできない」

1) 第1文、第2文からキーワード(内容語)を抽出する。
第1文: PW、変更
第2文: パスワード、変える、ログイン

2) 第1文の各キーワード毎に、キーワードベクトルを用いて、第2文内で最も類似度が高いキーワードを1つ抽出する。
PW → パスワード (類似度:0.90)
変更 → 変える (類似度:0.95)

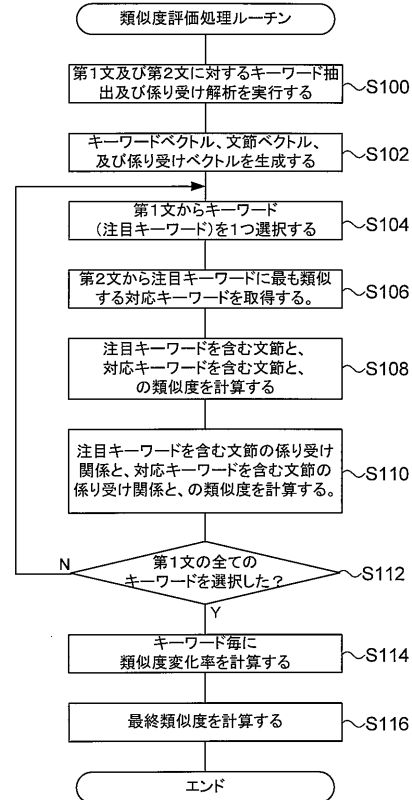
3) 上記2)で抽出したキーワードが含まれる文節の類似度を計算する。
PWの → パスワードを (類似度:0.75)
変える → 変えたら (類似度:0.32)

4) 上記3)の文節が含まれる係り受け関係を1つ選択し、類似度を計算する。
PWの変更を → パスワードを変えたら (類似度:0.15)
変更をしたい → 変えたらログインできない (類似度:0.04)

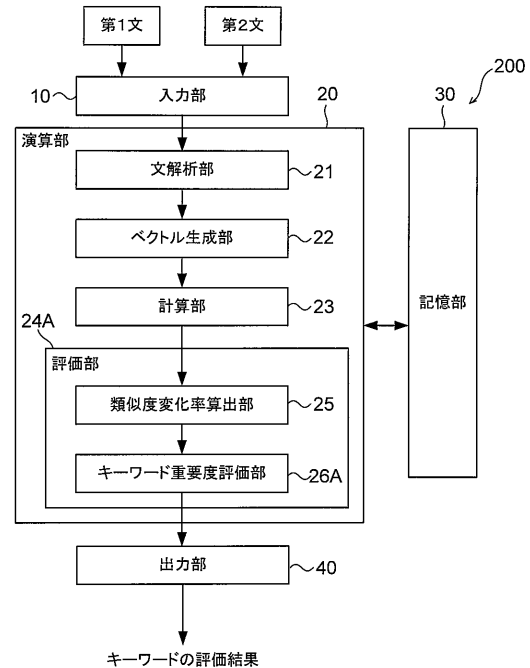
5) 上記2)～4)で計算した類似度を使用し、類似度変化率を計算する。
 $d_{ws}(PW) = \text{sim}(PWの, パスワードを) - \text{sim}(PW, パスワード)$
 $= 0.75 - 0.90 = -0.15$
 $d_{sd}(PW) = \text{sim}(PWの変更を, パスワードを変えたら) - \text{sim}(PWの, パスワードを)$
 $= 0.15 - 0.75 = -0.60$
 $d_{ws}(変更) = \text{sim}(変更を, 変えたら) - \text{sim}(変更, 変える)$
 $= 0.32 - 0.95 = -0.63$
 $d_{sd}(変更) = \text{sim}(変更をしたい, 変えたらログインできない) - \text{sim}(変更を, 変えたら)$
 $= 0.04 - 0.32 = -0.28$

6) 上記5)の類似度変化率を用いて最終的な類似度を計算する。
 $\text{SIM}(PWの変更をしたい, パスワードを変えたらログインできない)$
 $= (1/2) * [\text{sim}(PW, パスワード) * (1/3) * (1.0 + (1.0 - |d_{ws}(PW)|) * (1.0 + (1.0 - |d_{sd}(PW)|)))]$
 $+ \text{sim}(変更, 変える) * (1/3) * (1.0 + (1.0 - |d_{ws}(変更)|) * (1.0 + (1.0 - |d_{sd}(変更)|)))]$
 $= (1/2) * [0.3 * (1.0 + 0.85 + 0.34) + 0.32 * (1.0 + 0.36 + 0.259)]$
 $= 0.59$

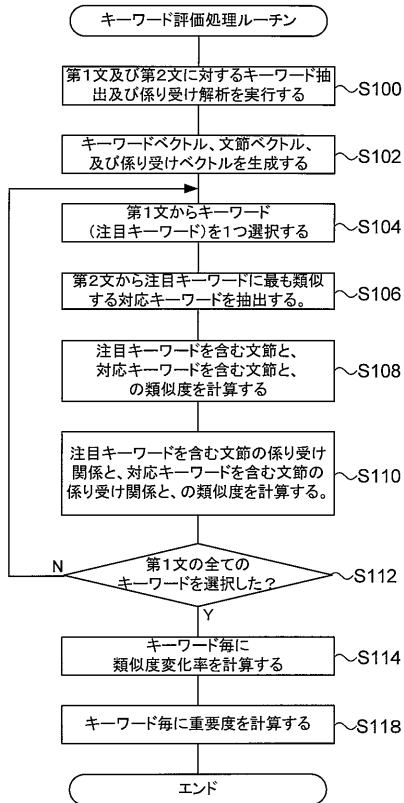
【図 3】



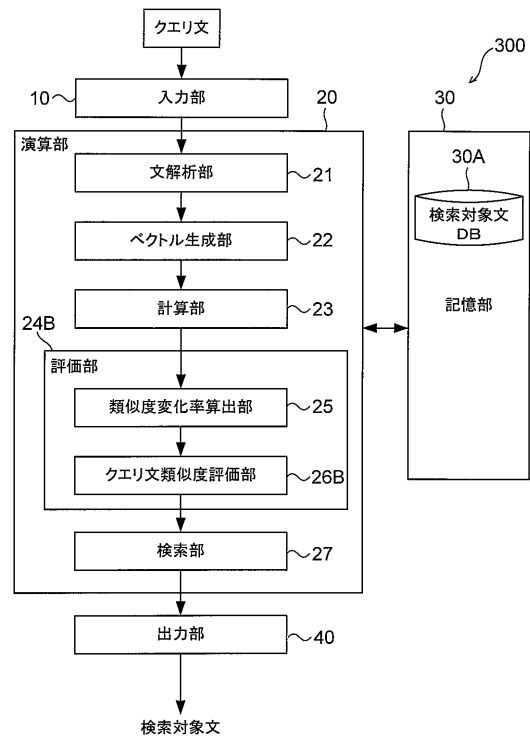
【図 5】



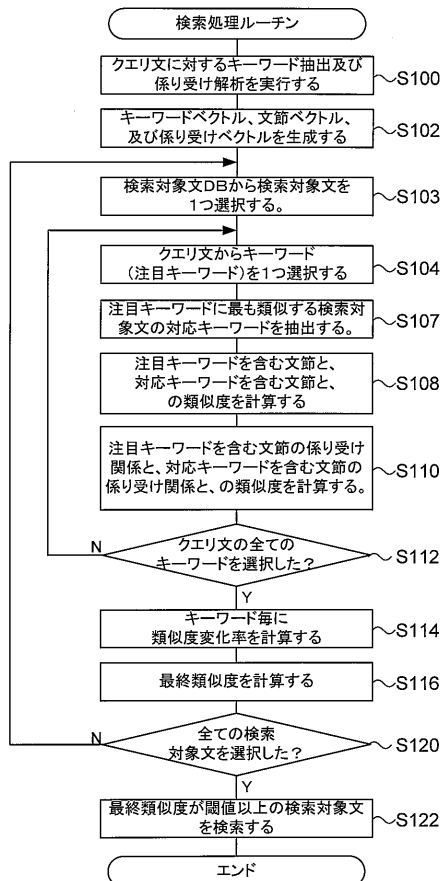
【図 6】



【図 7】



【図 8】



フロントページの続き

- (72)発明者 浅野 久子
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
- (72)発明者 松尾 義博
東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内