(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2012-113459
(P2012-113459A)
(43)公開日　平成24年6月14日(2012.6.14)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 17/28 | (2006.01) | G06F | 17/28 | Z | 5B075 |
| G06F 17/30 | (2006.01) | G06F | 17/30 | 170Z | 5B091 |
| | | G06F | 17/30 | 350C | |

審査請求 有　請求項の数 6　OL　（全 13 頁）

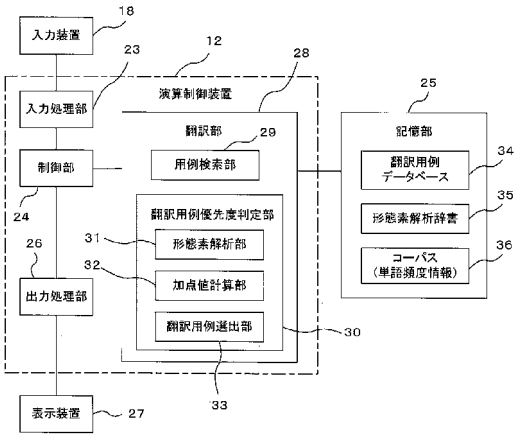
(21)出願番号　特願2010-260845 (P2010-260845)
(22)出願日　平成22年11月24日 (2010.11.24)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(71)出願人　301063496
東芝ソリューション株式会社
東京都港区芝浦一丁目１番１号
(74)代理人　100100516
弁理士　三谷　恵
(72)発明者　中村　寛爾
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内
(72)発明者　澁谷　貴志
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内

最終頁に続く

(54)【発明の名称】用例翻訳システム、用例翻訳方法及び用例翻訳プログラム

(57)【要約】
【課題】類似度の計算方法に翻訳対象の分野情報を指標の一つとして加えることで、利用者の求める翻訳結果により近い用例訳文を提供することである。
【解決手段】用例検索部は翻訳対象原文と翻訳用例データベースの翻訳用例原文との類似度を計算し類似度が予め定めた閾値以上の翻訳用例を翻訳用例データベースから検索する。形態素解析部は用例検索部により複数の翻訳用例が検索されたとき複数の翻訳用例のそれぞれの訳文を形態素解析辞書の形態素解析情報を参照して形態素解析し単語を抽出する。加点値計算部は形態素解析部で抽出された前記単語につきコーパスの単語頻度情報を参照し単語の出現頻度に応じて翻訳用例の類似度の加算値を計算する。翻訳用例選出部は用例検索部で計算された類似度に加点値計算部で計算された加算値を加算して最も大きい類似度の翻訳用例を選出する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
翻訳対象の第１言語の原文と翻訳目的の第２言語の訳文とを対にした翻訳用例を格納した翻訳用例データベースと、
文を形態素解析する際に参照する形態素解析辞書と、
所定の分野の文書で用いられた単語の出現頻度を格納したコーパスとを記憶した記憶装置と、
前記翻訳対象原文と前記翻訳用例データベースの翻訳用例原文との類似度を計算し、その類似度が予め定めた閾値以上の翻訳用例を前記翻訳用例データベースから検索する用例検索部と、
前記用例検索部により複数の翻訳用例が検索されたとき、当該複数の翻訳用例の各翻訳用例訳文を前記形態素解析辞書を参照して形態素解析し単語を抽出する形態素解析部と、
前記形態素解析部で抽出された単語につき前記コーパスに格納された当該単語の出現頻度に応じて前記類似度に加点する加点値計算部と、
前記加点後の類似度に基づいて翻訳用例を選出する翻訳用例選出部と、
を備えた用例翻訳システム。
【請求項２】
前記コーパスは単語の出現頻度の更新日時情報をさらに格納し、前記加点値計算部は前記加点について前記更新日時情報の新しいものほど大きな重み付けをする請求項１記載の用例翻訳システム。
【請求項３】
翻訳対象の第１言語の原文と翻訳目的の第２言語の訳文とを対にした翻訳用例を格納した翻訳用例データベースと、文を形態素解析する際に参照する形態素解析辞書と、所定の分野の文書で用いられた単語の出現頻度を単語頻度情報として格納したコーパスとを予め記憶装置に記憶しておき、
前記翻訳対象原文と前記翻訳用例データベースの翻訳用例原文との類似度を計算し、その類似度が予め定めた閾値以上の翻訳用例を前記翻訳用例データベースから検索し、
複数の翻訳用例が検索されたとき、当該複数の翻訳用例の各翻訳用例訳文を前記形態素解析辞書を参照して形態素解析し単語を抽出し、
抽出された前記単語につき前記コーパスに格納された当該単語の出現頻度に応じて前記類似度に加点し、
前記加点後の類似度に基づいて翻訳用例を選出して翻訳用例とする用例翻訳方法。
【請求項４】
前記コーパスに前記単語頻度情報に加え単語の出現頻度の更新日時情報を予め格納しておき、前記単語の出現頻度に応じて計算した加算値に前記更新日時情報の新しいものほど大きな重み付け係数を乗算した加算値を計算する請求項３記載の用例翻訳方法。
【請求項５】
前記用例翻訳プログラム、翻訳対象の第１言語の原文と翻訳目的の第２言語の訳文とを対にした翻訳用例を格納した翻訳用例データベース、翻訳用例訳文を形態素解析する際に参照する形態素解析辞書、所定の分野の文書で用いられた単語の出現頻度を単語頻度情報として格納したコーパスを予め記憶した記憶装置と、前記翻訳対象原文を入力するとともに操作に必要な情報を入力する入力装置と、前記翻訳対象原文や前記翻訳用例を表示する表示装置と、前記用例翻訳プログラムを演算実行する演算制御装置とを備えた用例翻訳システムとして機能させるためのコンピュータにおいて、
前記コンピュータを、
前記翻訳対象原文と前記翻訳用例データベースの翻訳用例原文との類似度を計算し、その類似度が予め定めた閾値以上の翻訳用例を前記翻訳用例データベースから検索する用例検索手段と、
前記用例検索部により複数の翻訳用例が検索されたとき、当該複数の翻訳用例の各翻訳用例訳文を前記形態素解析辞書を参照して形態素解析し単語を抽出する形態素解析手段と

、
前記形態素解析部で抽出された前記単語につき前記コーパスに格納された当該単語の出現頻度に応じて前記類似度に加点する加点値計算手段と、
前記加点後の類似度に基づいて翻訳用例を選出する翻訳用例選出手段として機能させるための用例翻訳プログラム。
【請求項６】
前記コーパスに前記単語頻度情報に加え単語の出現頻度の更新日時情報を予め格納しておき、前記加点値計算手段は単語の出現頻度に応じて計算した加算値に、前記更新日時情報の新しいものほど大きな重み付け係数を乗算した加算値を計算する請求項５記載の用例翻訳システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、翻訳対象の第１言語の原文と翻訳目的の第２言語の訳文とを対にした翻訳用例を用いて原文を訳文に翻訳する用例翻訳システム、用例翻訳方法及び用例翻訳プログラムに関する。
【背景技術】
【０００２】
外国語を用いた情報交換のツールとして機械翻訳の重要性が高まっている。機械翻訳の技術の一つとして翻訳用例を用いた翻訳方法が知られている。これは予め原文と訳文とを対にした複数の翻訳用例を翻訳用例データベースに登録しておき、翻訳対象原文が入力された際に、翻訳対象原文と類似した翻訳用例原文を翻訳用例データベースから検索し、得られた翻訳用例原文を、対となる翻訳用例訳文とともに利用者に提示するものである。利用者は必要に応じて提示された翻訳用例訳文を部分的に修正することで、希望する訳文を比較的少ない作業量で得ることができる。
【０００３】
ここで翻訳対象原文の類似文を検索する際、翻訳対象原文と各々の翻訳用例原文との類似度が計算される。類似度計算においては、翻訳対象原文と翻訳用例原文との間の一致する単語の割合を計算する方法が一般的な計算方法として知られている。類似度が同じ翻訳用例が複数見つかった場合には、検索で先に見つかったものを優先したり、登録時期が新しいものを優先したりすることで順位付けがなされることが多い。例として”I pass by the house every day.”という原文が与えられた場合、いま、翻訳用例データベースから類似度の高いものが検索され、下記のような用例が得られたとする。
【０００４】
用例１
原文：I pass by the shop every day.(類似度：86%)
訳文：私は毎日その店のそばを通る
用例２
原文：I pass behind the shop every day.(類似度：71%)
訳文：私は毎日その店の後ろを通る
この結果、翻訳者は最も類似度の高い用例１の訳文を部分的に修正することで、希望の訳文を得ることができる。
【０００５】
この方法により類似度が計算された場合、類似度が高いからといって必ずしも翻訳者の望む訳文に近い用例が優先的に検出されるとは限らない。例として、下記の用例１、２が翻訳用例データベースに登録されており、双方の原文と似た次の翻訳対象が入力されたとする。
【０００６】
用例１（登録日：2009/08/11）
原文：The stocks of this brand bring about profits.

訳文：この銘柄の株式は利益をもたらします
用例２（登録日：2009/11/30）
原文：The fans of this brand bring about profits.
訳文：このブランドのファンたちは利益をもたらします
翻訳対象
原文：The shares of this brand bring about damage.
ここで、この従来の計算方法により翻訳対象原文と各々の翻訳用例原文との類似度を計算した場合、用例１、用例２ともに８単語中６単語が原文と一致するため同じ類似度になり、登録日の新しい用例２の訳文が類似文の訳文として利用者に提示される。
【０００７】
しかし、もし、この翻訳対象原文が株式関連の文書中に現れた文であった場合、推測される訳文は「この銘柄の株式は損害をもたらします」となり、提示された用例２より用例１の訳文の方が近いということになる。このように、従来の計算方法により選ばれた類似文の訳文は必ずしも最適というわけではなく、類似度が同じ、またはやや低い別の用例の訳文の方が有用である場合も少なくない。
【先行技術文献】
【特許文献】
【０００８】
【特許文献１】特開２００６－２４１１４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００９】
従来の類似度の計算方法に翻訳対象の分野情報を指標の一つとして加えることで、利用者の求める翻訳結果により近い用例訳文を提供することである。
【課題を解決するための手段】
【００１０】
実施形態の翻訳用例システムは、翻訳対象の第１言語の原文と翻訳目的の第２言語の訳文とを対にした翻訳用例を格納した翻訳用例データベースと、文を形態素解析する際に参照する形態素解析辞書と、所定の分野の文書で用いられた単語の出現頻度を格納したコーパスとを記憶した記憶装置を備える。用例検索部は翻訳対象原文と翻訳用例データベースの翻訳用例原文との類似度を計算し、その類似度が予め定めた閾値以上の翻訳用例を翻訳用例データベースから検索する。形態素解析部は用例検索部により複数の翻訳用例が検索されたとき、当該複数の翻訳用例のそれぞれの各翻訳用例訳文を形態素解析辞書を参照して形態素解析し単語を抽出する。加点値計算部は形態素解析部で抽出された前記単語につきコーパスに格納された当該単語の出現頻度に応じて類似度に加点する。翻訳用例選出部は加点後の類似度に基づいて翻訳用例を選出する。
【図面の簡単な説明】
【００１１】
【図１】実施形態に係る用例翻訳システムの機能ブロック構成図。
【図２】実施形態に係る用例翻訳システムのハードウェア構成を示すブロック構成図。
【図３】実施形態に係るコーパスの一例の説明図。
【図４】実施形態でコーパスを作成する場合の処理内容を示すフローチャート。
【図５】実施形態でコーパスを作成する場合の特定分野の文書及びコーパスの説明図。
【図６】実施形態に係る用例翻訳システムに翻訳対象原文が入力されてからコーパスを利用して翻訳用例を選出するまでの処理内容を示すフローチャート。
【図７】実施形態に係るコーパスの他の一例の説明図。
【発明を実施するための形態】
【００１２】
以下、実施形態を図面に基づいて説明する。図１は、実施形態に係る用例翻訳システムの機能ブロック構成図、図２は実施形態に係る用例翻訳システムのハードウェア構成を示

すブロック構成図である。

【００１３】

図２において、用例翻訳システム１１は、例えば一般的なコンピュータに用例翻訳プログラムなどのソフトウェアプログラムがインストールされ、そのソフトウェアプログラムが演算制御装置１２のプロセッサ１３において実行されることにより実現される。

【００１４】

演算制御装置１２は機械翻訳に関する各種演算を行うものであり、演算制御装置１２はプロセッサ１３とメモリ１４とを有し、メモリ１４にはプログラム１５が記憶され、プロセッサ１３により処理が実行される際には作業エリア１６が用いられる。演算制御装置１２の演算結果等は表示装置１７に表示出力される。

【００１５】

入力装置１８は演算制御装置１２に情報を入力するものであり、例えば、マウス１９、キーボード２０、読み取り装置２１ａ、読み込み装置２１ｂから構成される。読み取り装置２１ａは、例えばＯＣＲ（光学式文字読み取り装置）等であり、読み込み装置２１ｂは、例えば磁気テープ、磁気ディスク、光ディスク等、コンピュータ可読媒体からの読み込み装置である。

【００１６】

例えば、マウス１９やキーボード２０は表示装置１７を介して演算制御装置１２に各種指令を入力し、キーボード２０、読み取り装置２１ａ、読み込み装置２１ｂは、翻訳対象の文書を入力する。すなわち、読み取り装置２１ａ、読み込み装置２１ｂは、翻訳対象の文書のファイルを記憶媒体に入出力するものである。さらに、演算制御装置１２の演算結果や用例翻訳に必要な知識・規則を蓄積した辞書等を記憶するハードディスクドライブ(ＨＤＤ)２２が設けられている。

【００１７】

図１において、演算制御装置１２内の各機能ブロックは、用例翻訳プログラムを構成する各プログラム１５の機能に対応する。すなわち、プロセッサ１３が用例翻訳プログラムを構成する各プログラム１５を実行することで、演算制御装置１２は、各機能ブロックとして機能することとなる。また、記憶装置２５の各ブロックは、演算制御装置１２内のメモリ１４及びハードディスクドライブ２２の記憶領域に対応する。

【００１８】

入力装置１８は、翻訳対象原文の文書の電子データを入力するものであり、利用者の入力操作に基づく文書の入力が可能である。また、入力装置１８は、入力処理部２３を介して制御部２４に対して各種コマンドを与える。入力装置１８によって入力された翻訳対象原文の文書は、演算処理部１２の入力処理部２３により入力処理されて取り込まれ、制御部２４を介して記憶装置２５の図示省略の文書記憶エリアに記憶される。制御部２４は、入力処理部２３、出力処理部２６、翻訳部２８を制御するとともに、記憶装置２５とのデータの授受の制御も行う。そして、演算制御装置１２の演算結果は表示装置２７に表示出力される。

【００１９】

翻訳部２８は、用例検索部２９及び翻訳用例優先度判定部３０を有し、翻訳用例優先度判定部３０は、形態素解析部３１、加点値計算部３２、翻訳用例選出部３３を有している。これらの詳細については、後述する。

【００２０】

また、記憶部２５には、複数の翻訳用例が予め登録されている翻訳用例データベース３４が格納されている。翻訳用例は第１言語の翻訳用例原文と第２言語の翻訳用例訳文とが対となって格納されている。また、記憶部２５には、翻訳対象原文や翻訳用例の原文及び訳文を形態素解析をする際に参照される形態素解析辞書３５が格納されている。形態素解析辞書３５には、形態素解析の対象となる第１言語や第２言語の文法の知識（文法のルールの集まり）や辞書（品詞等の情報付きの単語リスト）が形態素解析情報として格納されている。

【００２１】
さらに、記憶部２５には、翻訳に関連する分野の単語頻度情報が登録されているコーパス３６が格納されている。コーパス３６は、大量のテキストデータを翻訳システムで利用可能な形式にして登録したものであり、本実施形態では、翻訳対象原文と同じ分野の文書で用いられた名詞単語及びその出現頻度が単語頻度情報として格納され、また出現頻度の更新日時情報が格納されている。コーパス３６の詳細は後述する。
【００２２】
翻訳部２８の用例検索部２９は、翻訳対象原文と翻訳用例データベース３４の翻訳用例原文との類似度を計算し、類似度が予め定めた閾値以上の翻訳用例を翻訳用例データベース３４から検索するものである。
【００２３】
翻訳用例優先度判定部３０は、用例検索部２９により複数の翻訳用例が検索されたとき、検索された複数の翻訳用例のうち、どの翻訳用例を優先して選出するかを判定するものである。
【００２４】
翻訳用例優先度判定部３０の形態素解析部３１は、用例検索部２９により複数の翻訳用例が検索されたときは、複数の翻訳用例のそれぞれの翻訳用例訳文を形態素解析辞書３５の形態素解析情報を参照して形態素解析し、名詞単語を抽出する。
【００２５】
翻訳用例優先度判定部３０の加点値計算部３２は、形態素解析部３１で抽出された名詞単語につきコーパス３６の単語頻度情報を参照し、名詞単語の出現頻度に応じて翻訳用例の類似度の加算値を計算する。
【００２６】
翻訳用例優先度判定部３０の翻訳用例選出部３３は、用例検索部２９で計算された翻訳用例の類似度に、加点値計算部３２で計算された加算値を加算して、類似度の合計値が最も大きい翻訳用例を選出し、出力処理部２６を介して表示装置２７に表示出力する。
【００２７】
図３はコーパス３６の説明図である。コーパス３６は特定分野の文書で用いられた名詞単語の出現頻度を単語頻度情報として格納するとともに、出現頻度の更新日時を更新日時情報として格納している。図３では株式関連分野の場合のコーパスを示している。
【００２８】
例えば、株式という名詞単語は出現頻度が３０で更新日時は２０１０年１０月２９日であり、証券という名詞単語は出現頻度が２７で更新日時は２０１０年１０月０９日であり、以下、同様に株式関連分野の文書に用いられた名詞単語の出現頻度と更新日時とを情報として格納している。
【００２９】
図４は実施形態でコーパスを作成する場合の処理内容を示すフローチャートである。これは、図示は省略するが、コーパス作成プログラムをコンピューターにインストールし、そのソフトウェアプログラムを演算制御装置１２のプロセッサ１３において実行することにより実現される。
【００３０】
いま、図５（ａ）に示す株式分野の文書が入力装置１８から入力処理部２３を介して記憶部２５の図示省略の文書記憶エリアに記憶されたとする。図４に示すように、まず、コーパスの作成機能は、読み込まれた文書を文単位に切り出す（Ｓ１）。図５（ａ）の株式分野の文書の場合は、「株式とは、株式会社における社員権、持分のことである。」という文と、「通常の持分が社員の出資額などに応じて不均一な形態を取るのに対して、均一的な細分化された割合的な構成単位を取る点に特徴がある。」という文との二つの文からなっているので、この二つの文を切り出す。
【００３１】
次に、一つ目の文「株式とは、株式会社における社員権、持分のことである。」につき

、形態素解析により単語分割をする（Ｓ２）。そして、分割した単語から名詞を識別する（Ｓ３）。この場合の名詞は、図５（ａ）の下線を引いた単語であり、「株式」、「株式会社」、「社員権」、「持分」の４個の名詞である。

【００３２】

次に、変数ｉに「１」をセットし（Ｓ４）、ｉ個目の名詞を取り出し（Ｓ５）、ｉ個目の名詞の出現頻度に１を加算する（Ｓ６）。そして、更新日時を更新する（Ｓ７）。最初はｉ＝１であるから、１個目の名詞である「株式」が取り出される。１個目の名詞「株式」は、図５（ｂ）に示すように、「株式」の出現頻度２９に１を加算し、更新日時を本日の２０１０年１０月２９日に更新する。

【００３３】

次に、すべての名詞を取り出したか否かを判定し（Ｓ８）、すべての名詞を取り出していないときは、変数ｉに１を加算し（Ｓ９）、ステップＳ５に戻る。一方、すべての名詞を取り出しているときは、次の文はあるかどうかを判定し（Ｓ１０）、次の文があるときはステップＳ２に戻り、次の文がないときは処理を終了する。

【００３４】

このように、一つ目の文につき、ステップＳ５～ステップＳ９の処理により、２個目～４個目の名詞「株式会社」、「社員権」、「持分」についても、出現頻度に１を加算し、更新日時を本日の２０１０年１０月２９日に更新する。４個目の「持分」について処理が終了すると、ステップＳ１０により、二つ目の文についてステップＳ２～Ｓ１０までの処理が開始される。

【００３５】

次に、二つ目の文「通常の持分が社員の出資額などに応じて不均一な形態を取るのに対して、均一的な細分化された割合的な構成単位を取る点に特徴がある。」につき、形態素解析により単語分割をし（Ｓ２）、分割した単語から名詞を識別する（Ｓ３）。この場合の名詞は、図５（ａ）の下線を引いた単語であり、「通常」、「持分」、「社員」、「出資額」、「不均一」、「形態」、「均一」、「割合」、「構成単位」、「点」、「特徴」の１１個の名詞である。

【００３６】

一つ目の文の場合と同様に、変数ｉに「１」をセットし（Ｓ４）、ｉ個目の名詞を取り出し（Ｓ５）、ｉ個目の名詞の出現頻度に１を加算する（Ｓ６）。そして、更新日時を更新する（Ｓ７）。

【００３７】

最初はｉ＝１であるから、１個目の名詞である「通常」が取り出され、「通常」の出現頻度５に１を加算し、更新日時を本日の２０１０年１０月２９日に更新する。以下同様に、２個目～１１個目の名詞「持分」、「社員」、「出資額」、「不均一」、「形態」、「均一」、「割合」、「構成単位」、「点」、「特徴」についても、出現頻度に１を加算し、更新日時を本日の２０１０年１０月２９日に更新する。２個目の「持分」については、一つ目の文にも出現しているので、１が２回加算されることになる。そして、１１個目の名詞「特徴」の処理が終了すると、図５（ａ）の場合には、次の文はないので処理を終了する。

【００３８】

このようにして、コーパス３６には特定分野の名詞単語の出現頻度や更新日時が更新されて格納される。

【００３９】

次に、図６は、実施形態に係る用例翻訳システムに翻訳対象原文が入力されてからコーパスを利用して翻訳用例を選出するまでの処理内容を示すフローチャートである。

【００４０】

用例翻訳システムの利用者により入力装置１８から翻訳対象原文が入力されると、入力処理部２３により入力処理されて取り込まれ、制御部２４を介して記憶装置２５の図示省略の文書記憶エリアに記憶される。そして、制御部２４は翻訳部２８を起動する。

【００４１】
翻訳部２８は起動がかけられると、まず翻訳対象原文に対し形態素解析を行う（Ｓ１１）。用例検索部２９は、その結果をもとに翻訳用例データベース３４から翻訳用例を検索する（Ｓ１２）。このとき翻訳対象原文と、翻訳用例データベース３４に登録されている翻訳用例原文との類似度を計算することになるが、この類似度は双方の文中に同じ単語がどれだけ含まれるかという割合で決定される。そして、用例検索部２９は、類似度が予め定めた閾値以上を満たす翻訳用例は検索できたか否かを判定し（Ｓ１３）、翻訳用例が検索できない場合は処理を終了する。この場合は、翻訳用例を用いない通常の翻訳処理を行うことになる。
【００４２】
一方、用例検索部２９は、翻訳用例が検索できたときは、複数の翻訳用例か否かを判定する（Ｓ１４）。複数の翻訳用例でない場合、つまり一つの翻訳用例である場合には、その翻訳用例を選出する（Ｓ１５）。
【００４３】
ステップＳ１４の判定で、複数の翻訳用例が検索されたときは、翻訳用例優先度判定部３０は変数ｊに「１」をセットする（Ｓ１６）。これにより、形態素解析部３１はｊ個目の翻訳用例訳文を形態素解析し名詞単語を抽出する（Ｓ１７）。
【００４４】
次に、加点値計算部３２は、ｊ個目の翻訳用例訳文の名詞単語の出現頻度に応じて加算値を計算する（Ｓ１８）。すなわち、加点値計算部３２は、コーパス３６の名詞単語の出現頻度を参照して名詞単語の出現頻度を取得し、その出現頻度が高いほど大きな加算値を算出する。加算値の算出の仕方については後述する。そして、加点値計算部３２はｊ個目の翻訳用例の類似度に加算値を加算し（Ｓ１９）、すべての翻訳用例を取り出したか否かを判定し（Ｓ２０）、すべての翻訳用例を取り出していないときは、変数ｊに１を加算し（Ｓ２１）、ステップＳ１７に戻る。
【００４５】
一方、すべての翻訳用例を取り出しているときは、翻訳用例選出部３３は、加点値計算部３２により、翻訳用例の類似度に加算値を加算して得られた各々の翻訳用例の類似度合計値を比較し、最も大きい類似度合計値の翻訳用例を選出する（Ｓ２２）。
【００４６】
図６に示した処理内容につき具体例を用いて説明する。いま、翻訳対象原文として、下記の文が与えられたとする。
【００４７】
The shares of this brand bring about damage.
この翻訳対象原文の類似文を検索する場合を考える。なお、この翻訳対象原文は株式関連の文章中に現れた文であり、用例翻訳システムには予め株式の単語情報を登録したコーパス３６を持っているものとする。
【００４８】
まず、翻訳部２８は上記の翻訳対象原文を形態素解析により単語分割し、用例検索部２９は、分割された単語をもとに翻訳用例データベース３４中の翻訳用例を検索する。その結果、類似度の高い翻訳用例として、下記の二つの翻訳用例が得られたとする。
【００４９】
用例１
原文：The stocks of this brand bring about profits.
訳文：この銘柄の株式は利益をもたらします
用例２
原文：The fans of this brand bring about profits.
訳文：このブランドのファンたちは利益をもたらします
翻訳対象原文とこれら二つの翻訳用例原文との類似度は、どちらも７５｛（一致する単語数／全単語数）×１００で計算）｝で同じである。この場合、二つの翻訳用例が得られ

たので、図６のステップＳ１６以降の処理に移ることになる。

【００５０】

形態素解析部３１は、用例１及び用例２のそれぞれの翻訳用例訳文に対し、形態素解析により単語分割を行い名詞単語を取り出す。

【００５１】

用例１から、「銘柄」、「株式」、「利益」を取り出し、用例２から「ブランド」、「ファン」、「利益」を取り出す。

【００５２】

これらすべての名詞単語について、加点値計算部３２はコーパス３６を参照し、出現頻度の高い名詞単語についてはポイントを加点する。ここでは、簡略化のため、出現頻度を１０で割り小数点以下を切り捨てたものを加点するポイントとする。

【００５３】

用例１の「銘柄」の出現頻度は、図３に示すように「１６」であり、「株式」の出現頻度は「３０」であり、「利益」は未登録の名詞単語であるので出現頻度は「０」である。従って、用例１の名詞単語の出現頻度の合計は、（１６＋４０＋０＝４６）であり、これを１０で割り小数点以下を切り捨てると加算値は「４」と計算される。用例１の類似度は７５であるので、これに加算値４を加算すると、用例１の類似度合計値は７９となる。

【００５４】

一方、用例２の「ブランド」、「ファン」、「利益」は、図３に示すように、すべて未登録の名詞単語であるので出現頻度は「０」である。従って、用例２の加算値は「０」と計算される。用例２の類似度は７５であるので、これに加算値０を加算すると、用例２の類似度合計値は７５となる。

【００５５】

翻訳用例選出部３３は、最も大きい類似度合計値の翻訳用例を選出する。この場合は、用例１の類似度合計値が７９で用例２の類似度合計値が７５であるので、用例１が選出される。翻訳対象原文の内容を考慮すると、用例１の方が用例２より有用である。以上のように、コーパス３６を利用することによって翻訳対象原文と同じ分野の翻訳用例の類似度を高くすることで、より有用な翻訳用例を類似文として利用者に提示することができる。

【００５６】

このように、類似度の近い翻訳用例が複数ある場合、翻訳者の指定する分野の単語の出現頻度を利用することにより、指定の分野に近い訳文を持つ翻訳用例ほど類似度が高くなるため、単語の出現頻度を利用しない場合に比べ、より翻訳者の希望に近い類似文が検出される。

【００５７】

以上の説明では、類似度に加点するポイントについて、出現頻度をもとに計算したが、出現頻度だけではなく更新日時の情報も合わせて利用してもよい。一般的に、長い期間をかけてコーパス３６を作成する場合、古い単語情報よりも新しい単語情報の方がより有用である場合が多い。そこで、更新日時が新しいものほど加点ポイントが大きくなるよう重みを付けることによって、単語の新鮮さを類似文検索における指標へ反映させる。更新日時と係数との例を以下に示す。

【００５８】

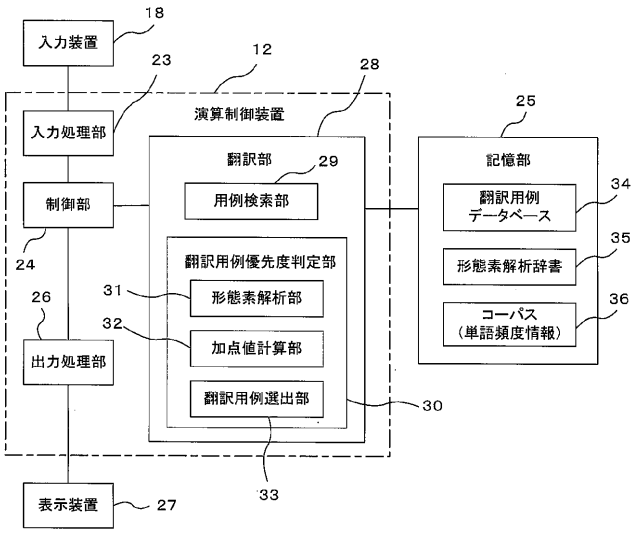
現在から　　係数
半年以内　　：1.0
１年以内　　：0.9
３年以内　　：0.8
３年以上経過　：0.7

具体例として以下の翻訳対象原文及び類似度の高い翻訳用例として検出された用例１、用例２を考える。また、利用するコーパス３６には、図７に示すような単語情報が登録されていたとする。

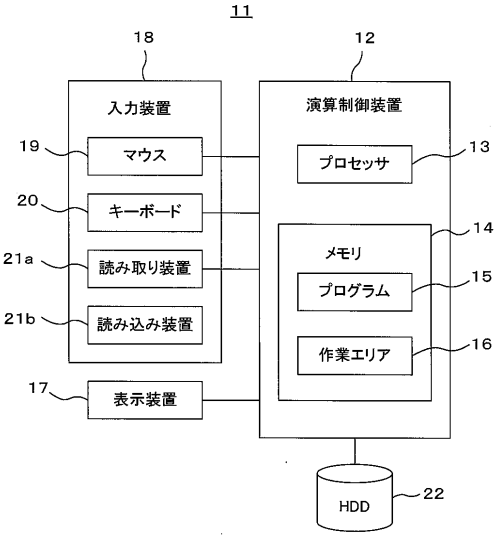
【００５９】

翻訳対象原文
The government must fix a safety net immediately.
用例１
原文：The city must fix a safety net immediately.
訳文：市は早急に安全網を整備しなければならない。
【００６０】
用例２
原文：The prefecture must fix a safety net immediately.
訳文：県は早急にセーフティーネットを整備しなければならない。
【００６１】
このとき、類似度に加点するポイントは、出現頻度に更新日時から計算した係数をかけたものを利用する。単純に出現頻度のみを参照した場合、「セーフティーネット」より「安全網」の方が重要な単語となり、用例２より用例１が優先される。
【００６２】
しかし、更新日時による係数をかけた場合、例えば、前述の係数を用いて「安全網」の出現頻度に０．７、「セーフティーネット」の出現頻度に１．０をかけると、加点ポイントは用例２の方が大きくなる。実際、現在では「安全網」という言葉より「セーフティーネット」の方が使われることが多く、用例２を優先的に利用者へ提示することは妥当な判断といえる。
【００６３】
以上のように、翻訳用例を用いて翻訳を行う場合に、類似した翻訳用例が複数検索された際、それら複数の翻訳用例に対して名詞単語の出現頻度や更新日時の情報をもとにポイントを加点することによって、翻訳する分野により近くより新しい訳し方を持つ翻訳用例が優先的に検出されるようになる。
【００６４】
本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。
【符号の説明】
【００６５】
１１…用例翻訳システム、１２…演算制御装置、１３…プロセッサ、１４…メモリ、１５…プログラム、１６…作業エリア、１７…表示装置、１８…入力装置、１９…マウス、２０…キーボード、２１ａ…読み取り装置、２１ｂ…読み込み装置、２２…ハードディスクドライブ、２３…入力処理部、２４…制御部、２５…記憶部、２６…出力処理部、２７…表示装置、２８…翻訳部、２９…用例検索部、３０…翻訳用例優先度判定部、３１…形態素解析部、３２…加点値計算部、３３…翻訳用例選出部、３４…翻訳用例データベース、３５…形態素解析辞書、３６…コーパス

【図１】



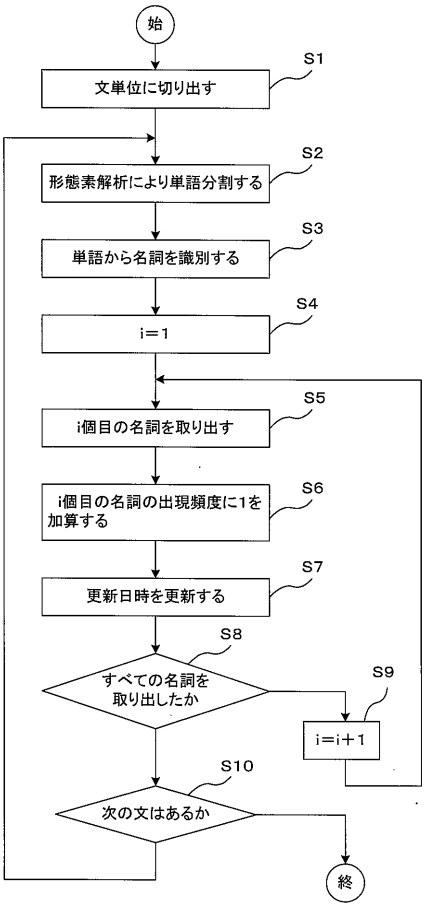

【図２】




【図３】

| 単語（名詞） | 出現頻度 | 更新日時 |
|---|---|---|
| 株式 | 30 | 2010-10-29 |
| 証券 | 27 | 2010-10-09 |
| 銘柄 | 16 | 2010-10-01 |
| 配当 | 9 | 2010-09-09 |
| 株式会社 | 11 | 2010-10-29 |
| 社員権 | 3 | 2010-10-29 |
| 持分 | 9 | 2010-10-29 |
| 通常 | 6 | 2010-10-29 |
| 社員 | 4 | 2010-10-29 |
| 出資額 | 5 | 2010-10-29 |
| 役員 | 5 | 2009-07-23 |
| 平均 | 21 | 2009-08-31 |
| 不均一 | 4 | 2010-10-29 |
| 形態 | 12 | 2010-10-29 |
| 均一 | 7 | 2010-10-29 |
| 単語 | 2 | 2009-11-09 |
| 割合 | 26 | 2010-10-29 |
| 構成単位 | 9 | 2010-10-29 |
| 会員 | 3 | 2009-11-09 |
| 点 | 9 | 2010-10-29 |
| 特徴 | 13 | 2010-10-29 |
| 経営 | 14 | 2010-05-24 |
| 巨額 | 9 | 2010-09-03 |
| 少額 | 6 | 2010-01-15 |
| ⋮ | ⋮ | ⋮ |

【図４】



【図５】

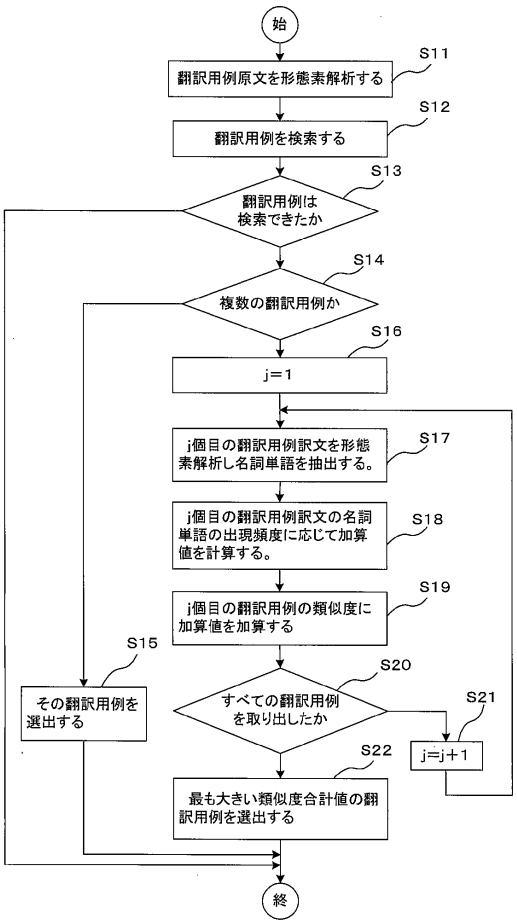
(a)

株式とは、株式会社における社員権、持分のことである。
通常の持分が社員の出資額などに応じて不均一な形態を取るのに対して均一的な細分化された割合的な構成単位を取る点に特徴がある。

(b)

| 単語(名詞) | 出現頻度 | 更新日時 |
|---|---|---|
| 株式 | 29+1 | 2010-10-29 |
| 証券 | 27 | 2010-10-09 |
| 銘柄 | 16 | 2010-10-01 |
| 配当 | 9 | 2010-09-09 |
| 株式会社 | 10+1 | 2010-10-29 |
| 社員権 | 2+1 | 2010-10-29 |
| 持分 | 7+1+1 | 2010-10-29 |
| 通常 | 5+1 | 2010-10-29 |
| 社員 | 3+1 | 2010-10-29 |
| 出資額 | 4+1 | 2010-10-29 |
| 役員 | 5 | 2009-07-23 |
| 平均 | 21 | 2009-08-31 |
| 不均一 | 3+1 | 2010-10-29 |
| 形態 | 11+1 | 2010-10-29 |
| 均一 | 6+1 | 2010-10-29 |
| 単語 | 2 | 2009-11-09 |
| 割合 | 25+1 | 2010-10-29 |
| 構成単位 | 8+1 | 2010-10-29 |
| 会員 | 3 | 2009-11-09 |
| 点 | 8+1 | 2010-10-29 |
| 特徴 | 12+1 | 2010-10-29 |
| 経営 | 14 | 2010-05-24 |
| 巨額 | 9 | 2010-09-03 |
| 少額 | 6 | 2010-01-15 |
| ⋮ | ⋮ | ⋮ |

【図６】




【図７】

| 単語(名詞) | 出現頻度 | 更新日時 |
|---|---|---|
| ⋮ | ⋮ | ⋮ |
| 安全網 | 85 | 2005-03-28 |
| ⋮ | ⋮ | ⋮ |
| セーフティーネット | 64 | 2010-02-19 |
| ⋮ | ⋮ | ⋮ |

フロントページの続き

(72)発明者　蔡　遠航
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内

Ｆターム(参考) 5B075 ND03　ND20　PR04
5B091 BA04　BA13