

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4041876号
(P4041876)

(45) 発行日 平成20年2月6日(2008.2.6)

(24) 登録日 平成19年11月22日(2007.11.22)

(51) Int.Cl. F I
G O 6 F 17/28 (2006.01) G O 6 F 17/28 P

請求項の数 3 (全 18 頁)

<p>(21) 出願番号 特願2001-268513 (P2001-268513) (22) 出願日 平成13年9月5日(2001.9.5) (65) 公開番号 特開2003-76687 (P2003-76687A) (43) 公開日 平成15年3月14日(2003.3.14) 審査請求日 平成13年9月5日(2001.9.5) 審判番号 不服2004-3567 (P2004-3567/J1) 審判請求日 平成16年2月24日(2004.2.24)</p> <p>特許法第30条第1項適用 平成13年3月30日 言語処理学会開催の「言語処理学会第7回年次大会」において文書をもって発表</p>	<p>(73) 特許権者 301022471 独立行政法人情報通信研究機構 東京都小金井市貫井北町4-2-1 (74) 代理人 100087848 弁理士 小笠原 吉義 (74) 代理人 100094662 弁理士 穂坂 和雄 (72) 発明者 村田 真樹 東京都小金井市貫井北町4-2-1 独立行政法人通信総合研究所内 (72) 発明者 井佐原 均 東京都小金井市貫井北町4-2-1 独立行政法人通信総合研究所内</p>
--	---

最終頁に続く

(54) 【発明の名称】 複数尺度の利用による言語変換処理システムおよびその処理プログラム

(57) 【特許請求の範囲】

【請求項1】

ある自然言語で記述された文字列を他の表現による文字列に変換するシステムであって

前記自然言語で記述された第1の文字列を同義の第2の文字列に言い換える変形の規則であって、コンピュータにより、複数の異なる辞書における同じ項目の定義文を照合し、該定義文において一致する文字列部分と一致しない文字列部分とを抽出することにより、同義表現の文字列を獲得し、得られた同義表現の文字列から自動生成された前記第1の文字列と前記第2の文字列からなる変形規則を記憶する変形規則記憶手段と、

文字列を変形した結果の表現が目的とする表現になっているかどうかを評価するための数値情報、関数群もしくはサブルーチン群、評価方法を記述した規則、またはこれらを組み合わせた情報からなる評価尺度ごとの評価情報であって、該評価情報を規定する評価尺度がM種類存在し、それらのM種類の評価尺度の中に、変換の候補の文字列の長短に基づき変換の候補の文字列が短いものほど高い評価値を与えるように予め定められた評価尺度、または、所定の平易な文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度、または、口語もしくは文章語で表現された文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度、または、特定の個人の文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基

10

20

づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度，または，変換対象となっている複数の文字列の類似度に基づき類似度が高いものほど高い評価値を与えるように予め定められた評価尺度のいずれか少なくとも2種類を含む，M種類の各評価尺度に対応した評価情報を記憶する評価情報記憶手段と，

自然言語で記述された変換対象の文字列を入力する入力手段と，

前記入力された文字列を前記変形規則記憶手段に記憶された変形規則を用いて変形し，複数の変換の候補を生成する変形処理手段と，

前記変換の候補の評価に用いるN種類の評価尺度を，前記M種類(M > N)の評価尺度の中から外部からの指定により選択し，前記変形処理手段により生成された複数の変換の候補の評価値を，前記評価情報記憶手段に記憶された前記N種類の評価尺度に対応する評価情報を用いて前記評価尺度ごとに算出し，その算出されたN種類の評価尺度によるN個の評価値のすべてを用いて算出された総合評価値の最も高い表現を選択する評価処理手段と，

前記選択された表現の変換結果を，目的とする表現に変換された文字列として出力する出力手段とを備え，

前記評価処理手段は，前記出力手段により変換結果が出力された後に，前記N種類の評価尺度を変更する指示があると，前記総合評価値の最も高い表現を選択する処理を，変更された評価尺度を用いて繰り返す

ことを特徴とする複数尺度の利用による言語変換処理システム。

【請求項2】

前記評価処理手段が変換の候補の評価に用いるN種類の評価尺度に対して，各評価尺度の重要度に関する指定情報を入力する手段を備え，

前記評価処理手段は，入力された指定情報に基づいて，個々の評価尺度に対する評価値に前記重要度に応じた重み付けをし，前記総合評価値を算出する

ことを特徴とする請求項1記載の複数尺度の利用による言語変換処理システム。

【請求項3】

ある自然言語で記述された文字列を他の表現による文字列に変換するシステムを，コンピュータによって実現するためのプログラムであって，

前記コンピュータを，

前記自然言語で記述された第1の文字列を同義の第2の文字列に言い換える変形の規則であって，複数の異なる辞書における同じ項目の定義文を照合し，該定義文において一致する文字列部分と一致しない文字列部分とを抽出することにより，同義表現の文字列を獲得し，得られた同義表現の文字列から自動生成された前記第1の文字列と前記第2の文字列からなる変形規則を記憶する変形規則記憶手段と，

文字列を変形した結果の表現が目的とする表現になっているかどうかを評価するための数値情報，関数群もしくはサブルーチン群，評価方法を記述した規則，またはこれらを組み合わせた情報からなる評価尺度ごとの評価情報であって，該評価情報を規定する評価尺度がM種類存在し，それらのM種類の評価尺度の中に，変換の候補の文字列の長短に基づき変換の候補の文字列が短いものほど高い評価値を与えるように予め定められた評価尺度，または，所定の平易な文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度，または，口語もしくは文章語で表現された文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度，または，特定の個人の文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度，または，変換対象となっている複数の文字列の類似度に基づき類似度が高いものほど高い評価値を与えるように予め定められた評価尺度のいずれか少なくとも2種類を含む，M種類の各評価尺度に対応した評価情報を記憶する評価情報記憶手

10

20

30

40

50

段と、

自然言語で記述された変換対象の文字列を入力する入力手段と、

前記入力された文字列を前記変形規則記憶手段に記憶された変形規則を用いて変形し、複数の変換の候補を生成する変形処理手段と、

前記変換の候補の評価に用いるN種類の評価尺度を、前記M種類(M < N)の評価尺度の中から外部からの指定により選択し、前記変形処理手段により生成された複数の変換の候補の評価値を、前記評価情報記憶手段に記憶された前記N種類の評価尺度に対応する評価情報を用いて前記評価尺度ごとに算出し、その算出されたN種類の評価尺度によるN個の評価値のすべてを用いて算出された総合評価値の最も高い表現を選択する評価処理手段と、

10

前記選択された表現の変換結果を、目的とする表現に変換された文字列として出力する出力手段として機能させ、

さらに前記評価処理手段が、前記出力手段により変換結果が出力された後に、前記N種類の評価尺度を変更する指示があると、前記総合評価値の最も高い表現を選択する処理を、変更された評価尺度を用いて繰り返すものとして

機能させるための複数尺度の利用による言語変換処理プログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、ある自然言語で記述された文または文章などの文字列を、同一または他の自然言語で記述された他の表現による文字列に変換するシステムであって、特に、コンピュータによる自然言語処理において多種多様な言い換えを扱うことができるようにした複数尺度の利用による言語変換処理システムおよびその処理プログラムに関するものである。

20

【0002】

【従来の技術】

自然言語で記述された文または文章に関する表現の変換処理として典型的なものは、機械翻訳である。機械翻訳では、ある国の自然言語で記述された文または文章を他の国の自然言語で記述された文または文章に変換する。

【0003】

機械翻訳が他の国の言語に変換するのに対し、同一の自然言語間での文または文章の変換処理を行うシステムも用いられるようになってきている。例えば、要約文を自動生成したり、文章を推敲したりするシステムである。

30

【0004】

一般に同一自然言語間での文の変換処理では、変換前の語・句・文などのパターンと変換後の語・句・文などのパターンとの対からなる変換規則を大量に用意し、いわゆるパターン・マッチングによって入力文中に現れる変換前のパターンを探し出し、該当するパターンがあれば、それを変換後の語・句・文などのパターンに置き換える処理を行っている。

【0005】

【発明が解決しようとする課題】

従来の同一自然言語内での文または文章の変換処理では、一般に変換規則による一律な変換を行っており、変換結果の良し悪しについての評価は行われていなかった。また、平易文生成、要約文生成、文章の推敲といった変換の目的に応じて、各システムごとにそれぞれ個別に独自の変換の処理ロジックを用いているため、例えば口語で表現された文章の要約を作成するというような場合には、まず口語文章語変換システムにより口語文を通常の文章語による文に変換し、その結果について要約文生成システムにより要約を生成するというような処理が必要であった。

40

【0006】

また、与えられた文章を推敲し、それについて要約を作成するというような場合にも、まず推敲システムにより文章を推敲し、その後要約文生成システムによって要約を生成するか、最初に要約文生成システムによって要約を生成し、その結果を推敲システムを用い

50

て推敲するという処理が必要であった。このとき、文章の変換は一律に行われ、要約のほうを推敲よりも重視するとか、これとは反対に推敲のほうを要約よりも重視するということはできなかった。

【0007】

本発明は上記問題点の解決を図り、複数種類の言い換えが必要な文または文章を、目的とする文または文章に簡単に変換することができるシステムを提供することを目的とする。

【0008】

【課題を解決するための手段】

本発明は、上記課題を解決するため、ある自然言語で記述された文字列を他の表現による文字列に変換するシステムにおいて、

前記自然言語で記述された第1の文字列を同義の第2の文字列に言い換える変形の規則であって、コンピュータにより、複数の異なる辞書における同じ項目の定義文を照合し、該定義文において一致する文字列部分と一致しない文字列部分とを抽出することにより、同義表現の文字列を獲得し、得られた同義表現の文字列から自動生成された前記第1の文字列と前記第2の文字列からなる変形規則を記憶する変形規則記憶手段と、

文字列を変形した結果の表現が目的とする表現になっているかどうかを評価するための数値情報、関数群もしくはサブルーチン群、評価方法を記述した規則、またはこれらを組み合わせた情報からなる評価尺度ごとの評価情報であって、該評価情報を規定する評価尺度がM種類存在し、それらのM種類の評価尺度の中に、変換の候補の文字列の長短に基づき変換の候補の文字列が短いものほど高い評価値を与えるように予め定められた評価尺度、または、所定の平易な文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度、または、口語もしくは文章語で表現された文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度、または、特定の個人の文章集合からなる言語データ中に現れる出現頻度もしくは出現確率に基づき変換の候補の出現頻度もしくは出現確率が大きいものほど高い評価値を与えるように予め定められた評価尺度、または、変換対象となっている複数の文字列の類似度に基づき類似度が高いものほど高い評価値を与えるように予め定められた評価尺度のいずれか少なくとも2種類を含む、M種類の各評価尺度に対応した評価情報を記憶する評価情報記憶手段と、

自然言語で記述された変換対象の文字列を入力する入力手段と、

前記入力された文字列を前記変形規則記憶手段に記憶された変形規則を用いて変形し、複数の変換の候補を生成する変形処理手段と、

前記変換の候補の評価に用いるN種類の評価尺度を、前記M種類(M > N)の評価尺度の中から外部からの指定により選択し、前記変形処理手段により生成された複数の変換の候補の評価値を、前記評価情報記憶手段に記憶された前記N種類の評価尺度に対応する評価情報を用いて前記評価尺度ごとに算出し、その算出されたN種類の評価尺度によるN個の評価値のすべてを用いて算出された総合評価値の最も高い表現を選択する評価処理手段と、

前記選択された表現の変換結果を、目的とする表現に変換された文字列として出力する出力手段とを備え、

前記評価処理手段は、前記出力手段により変換結果が出力された後に、前記N種類の評価尺度を変更する指示があると、前記総合評価値の最も高い表現を選択する処理を、変更された評価尺度を用いて繰り返すことを特徴とする。

【0009】

また、評価処理手段が変換の候補の評価に用いるN種類の評価尺度は、あらかじめ用意されたM種類(M > N)の評価尺度の中から、ユーザまたはアプリケーションプログラム等からの指定により選択する手段を持つ。

【0010】

10

20

30

40

50

さらに、評価処理手段が変換の候補の評価に用いるN種類の評価尺度に対して、各評価尺度の重要度に関する指定情報を入力する手段を設け、評価処理手段は、入力された指定情報に基づいて、個々の評価尺度に対する重要度に応じた評価結果から総合的な評価を行うようにすることもできる。

【0011】

N種類の評価尺度としては、変形後の文字列の長短、大量の用例に関する言語データ(コーパス)中に現れる出現頻度もしくは出現確率、所定の平易な文章集合からなる言語データ中に現れる出現頻度もしくは出現確率、口語もしくは文章語で表現された大量の文章集合からなる言語データ中に現れる出現頻度もしくは出現確率、特定の個人の文章集合からなる言語データ中に現れる出現頻度もしくは出現確率、または、変換対象となっている複数の文字列の類似度などを用いることができる。

10

【0012】

評価尺度として、変形後の文字列の長短を用い、変形後の文字列が短いものに高い評価を与えれば、冗長な表現を短くした変換文字列を生成することができる。また、評価尺度として、大量の用例に関する言語データ中に現れる出現頻度または出現確率を用い、その出現頻度または出現確率が大きいものに高い評価を与えれば、一般によく使われる文または文章になるように推敲した変換文字列を得ることができる。

【0013】

評価尺度として、所定の平易な文章集合からなる言語データ中に現れる出現頻度または出現確率を用い、その出現頻度または出現確率が大きいものに高い評価を与えれば、法律文などの難解な文を平易な文に変換した文字列を得ることができる。

20

【0014】

また、評価尺度として、口語で表現された大量の文章集合からなる言語データ中に現れる出現頻度または出現確率を用い、その出現頻度または出現確率が大きいものに高い評価を与えれば、文章語を口語表現に変換した文字列を生成することができる。これとは逆に、評価尺度として、文章語で表現された大量の文章集合からなる言語データ中に現れる出現頻度または出現確率を用い、その出現頻度または出現確率が大きいものに高い評価を与えれば、口語表現を文章語の表現に変換した文字列を生成することができる。

【0015】

さらに、評価尺度として、例えば夏目漱石とか芥川龍之介といった特定の個人の文章集合からなる言語データ中に現れる出現頻度または出現確率を用い、その出現頻度または出現確率が大きいものに高い評価を与えれば、与えられた文章を夏目漱石の文体もしくは芥川龍之介の文体といった特定の個人の文体に変換することができる。

30

【0016】

また、評価尺度として、変換対象となっている複数の文字列の類似度を用い、与えられた複数の文を比較する場合に、単なる表現形式ではなく、実質的な内容も考慮に入れた類似度を比較できるような文に変換することができるようになる。

【0017】

本発明では、特に以上のような評価尺度を複数組み合わせ用いることができる。したがって、与えられた文章を推敲し、それについて要約を作成するというような文字列の変換や、難解な文を易しい文に変換し、しかもそれを特定の個人の文体で表現するというような変換を、一度で行うことができるようになる。また、各評価尺度の重要度を随時変えることにより、目的とする変換文字列が得られるように調整することができる。

40

【0018】

以上の手段は、コンピュータと、そのコンピュータにインストールされ実行されるソフトウェアプログラムとによって実現することができ、そのプログラムは、コンピュータが読み取り可能な可搬媒体メモリ、半導体メモリ、ハードディスク等の適当な記録媒体に格納することができる。

【0019】

【発明の実施の形態】

50

図 1 は、本発明のシステム構成例を示す。図中、1 は CPU およびメモリなどからなる言語変換処理装置、2 はディスプレイ、キーボードその他の入出力装置を表す。

【0020】

変換対象文 10 は、本システムにおける入力となる自然言語文である。以下、特に断らないが変換対象文 10 は必ずしも一文に限られるわけではなく、文章または句、節のようなものであってもよい。変換結果文 17 は、本システムの出力であって、変換対象文 10 を同一の種類または異なる種類の自然言語で言い換えたものである。

【0021】

言語変換処理装置 1 のモジュールは、基本的に変形処理部 11 と評価処理部 12 とから構成される。変形処理部 11 は、変形規則記憶部 15 に格納されている変形規則を用いて、変換の候補を獲得するモジュールである。評価処理部 12 は、変換の候補のよさを、あらかじめ評価情報記憶部 16 に記憶されている複数の評価尺度（評価関数など）によって評価し、最もふさわしい変換の候補を選択するモジュールである。

10

【0022】

評価処理部 12 は、変換候補の評価に用いる評価尺度を入出力装置 2 からの指定によって選択する評価尺度選択部 13 と、選択された各評価尺度の重要度に関する指定情報を入出力装置 2 から入力し設定する評価重要度設定部 14 とを持ち、評価尺度選択部 13 により選択された評価尺度と、評価重要度設定部 14 によって設定された各評価尺度の重要度とから、個々の評価尺度に対する重要度に応じた変換候補の総合的な評価を行い、総合的な評価結果のよい変換候補を選択する。

20

【0023】

言語変換処理装置 1 の動作は、以下のとおりである。変換対象文 10 が入力されると、変形処理部 11 は、変形規則を用いて変換の候補を挙げ、評価処理部 12 は、変換の候補の妥当性をチェックして、最も妥当であると判断されたものを選択し、その結果を変換結果文 17 として出力する。

【0024】

変形規則記憶部 15 に記憶する変形規則は、コンピュータによって大量の言語データから獲得して自動的に生成する。例えば、同義性を満足する変形規則を自動獲得する方法の例としては、次のような方法を挙げることができる。異なる複数の辞書の同じ項目の定義文を照合し、その照合結果から変形規則を得る。例えば「あべこべ」という語の定義文を考えてみる。大辞林（三省堂）では、「あべこべ」の説明文が、「順序・位置などの関係がさかさまに入れかわっていること。」となっており、岩波国語辞典では、「順序・位置・関係がひっくり返っていること。」となっている。これを適当に照合すると、「関係が」と「こと。」が一致し、その間の「さかさまに入れかわっている」と「ひっくり返っている」が同義表現として機械的に獲得される。

30

【0025】

変換の候補を評価する評価尺度（評価関数）の評価情報は、扱う問題ごとに適正なものが複数種類、あらかじめ評価情報記憶部 16 に用意される。評価尺度としての評価情報は、評価のための数値情報であってもよいし、関数群もしくはサブルーチン群などによる手続き的なものであってもよい。また、評価方法を記述した規則（ルール）であってもよい。これらの組み合わせで実現することも可能である。評価処理部 12 で用いる評価尺度の例としては、以下のようなものが考えられる。

40

【0026】

(1) 長さ

例えば、要約の一つの分野の文圧縮のように、なるべく意味を変えずに文を圧縮したいとする。このとき、変形処理部 11 が使用する変形規則はすべて意味をほとんどかえずに変形するものであるとする。この場合、長さを評価の尺度とし、この長さが短くなるように変形を繰り返すと文圧縮が実現される。

【0027】

(2) 類似度

50

例えば、AとBの類似度を調べたいとする。このとき、変形処理部11が使用する変形規則がすべて同義性を満足するものであるとする。この場合、AとBの類似度が大きくなるように、変形規則でA、Bを変形し、A、Bをよく似た状態にしてから類似度を求める。こうすることにより、意味が同じなのに異なる表現で記述されているような場合でも正しく類似度を計算することができる。なお、類似度の値は、A、Bをそれぞれ構文解析し、一致する単語数、文節数、係り受け距離（構文木における二つの文節の間の枝の数）、文節距離などを考慮して定めることができる。

【0028】

(3) 出現頻度（または出現確率）

例えば、文章の表現を改善する推敲を考える。このとき、変形処理部11が使用する変形規則がすべて同義性を満足するものであるとする。この場合、推敲したいデータを、そのデータの出現（生起）確率が高くなるように変形すると非常に洗練された文章となる。

【0029】

もう少し簡単な例でこれを説明すると、例えば入力したデータに「データー」とあったとしよう。また、変形規則に「データー」を「データ」とする規則があったとしよう。新聞記事やコーパスなどのデータベースにより、「データー」と「データ」の出現回数を数え、「データ」の出現回数のほうが数が多い場合、「データ」のほうの評価を「データー」より高くする。

【0030】

また、出現頻度（または出現確率）を調べるコーパスをいろいろと変えることにより、さまざまな変換の結果を得ることができる。例えば、入力データが書き言葉のときに、コーパスとして話し言葉を用いると書き言葉の話し言葉への変換が実現される。

【0031】

また、入力データが法律関係の文のときに、コーパスとして平易な文章の集合を与えておくと、法律関係の難解な文章を平易な文章に変形させることが期待できる。

【0032】

さらにまた、ここで入力データとして適当に誰かが書いた小説の文章を入れて、コーパスとしてシェークスピアの小説をいれると、シェークスピアの文体の小説が新たに完成することになる。同様に、芥川龍之介の小説を夏目漱石の文体に変形するなどといったことも可能になる。

【0033】

上記の出現（生起）確率に基づく尺度は、文の正当性のチェックに使うこともできる。さらに、評価尺度として、所定の文章集合での出現頻度や出現確率に限らず、他の何らかの尺度を用いることもできる。例えば、あらかじめ単語の結び付きや、構文解析結果から得られる文法上の言い回しに対して、評価ポイントを定めておき、それを用いて評価するようなことも可能である。

【0034】

評価尺度を条件のようにして用いることもできる。条件のような尺度として、例えば「21世紀」というような特定の語を使うことに高い評価を与えたり、起承転結を満足する文章構成をとる変換に高い評価を与えたり、係り先未決定文節数が8程度以上である変換は評価を低くするということが考えられる。また、英語文でRやLを含む発音しにくい単語をあまり使わないというような尺度も考えられる。

【0035】

以上の評価尺度を複数組み合わせるにより、多種多様な文字列の変換を実現することができる。本発明は、複数の評価尺度を組み合わせる望ましい変換結果を得ることができるようにしたものであるが、以下では、本発明の理解を容易にするために、各種の評価尺度を単独で用いた場合の具体例について説明する。

【0036】

(A) 文内圧縮の変換例

図2は、文内圧縮（要約文生成）の変換例を示している。図2に示す変換では、要約文の

10

20

30

40

50

作成などのために、与えられた文をできるだけ元の文の意味を保存した形で、冗長な文を短く圧縮する処理を行う。例えば、新聞記事の要約を考えた場合、評価の尺度としては、入力されたデータがより短くなるような変形をよしとする尺度が考えられる。さらに条件として、新聞記事での出現が1個以上というような条件を付加してもよい。以下、具体例に従って説明する。

【0037】

例えば図2の例のように、変換対象文10として、「次の参議院選挙でA氏を擁立することを決めた」という文が入力されたとする。変形処理部11は、この変換対象文10を、変形規則記憶部15にあらかじめ用意された変形規則を用いて、異なる表現に言い換える。ここで、変形規則として、

「XでYを擁立すること」 「XでのYの擁立」

という規則があったとすると、変形処理部11は、変換対象文10に変形規則を適用することにより、「次の参議院選挙でA氏を擁立することを決めた」という文から「次の参議院選挙でのA氏の擁立を決めた」という文を生成する。この他にも、種々の変形規則が存在し、多くの変形された文が候補として生成されることになる。これらの文を評価処理部12に渡す。

【0038】

評価処理部12は、文内圧縮用の評価尺度(評価関数)を用いて、変形処理部11が変形した文を評価する。ここで評価の尺度が、入力した文の長短であり、文の長さが短いほど評価が高いとすると、多くの変形の中から文が最も短い文が選ばれることになる。この例では、「次の参議院選挙でのA氏の擁立を決めた」の評価が高く、変換結果文17としてこの文が出力されている。

【0039】

(B) 文章推敲の変換例

図3は、文章推敲のための変換例を示している。図3に示す文章推敲では、入力した文または文章を推敲して、より良いと考えられる表現の文または文章に改善する処理を行う。

【0040】

例えば図3の例のように、変換対象文10として、「世界の平和・安定に貢献する」という文が入力されたとする。変形処理部11は、この変換対象文10を、変形規則記憶部15にあらかじめ用意された変形規則を用いて、異なる表現に言い換える。ここで、変形規則として、

「・」 「と」

という規則があったとすると、変形処理部11は、変換対象文10に変形規則を適用することにより、「世界の平和・安定に貢献する」という文から「世界の平和と安定に貢献する」という文を生成する。この他にも、種々の変形規則が存在し、多くの変形された文が候補として生成されることになる。これらの文を評価処理部12に渡す。なお、変形されなかった変換対象文10についても候補の一つとして評価処理部12に渡す。

【0041】

評価処理部12は、文章推敲用の評価尺度(評価関数)を用いて、変形処理部11が変形した文を評価する。ここで評価の尺度が、大量の言語データ(用例、つまり実際に人々によって用いられたことのある言語表現の集合)での出現頻度もしくは出現確率が大きくなる変換をよしとするものである場合に、評価処理部12は、大量の言語データにおける「世界の平和・安定に貢献する」と「世界の平和と安定に貢献する」の生起確率を求める。簡便な手法としては、変形した部分を含む小さい領域範囲の文字列が言語データで何回出現したかを数える。例えば「平和・安定」が134回、「平和と安定」が23823回現れたとすると、「平和と安定」のほうが出現頻度が大きくこの表現のほうがより自然な表現であるとわかる。これによりこの変形はよしとされ、変換結果文17として「世界の平和と安定に貢献する」が出力される。なお、出現頻度ではなく、出現(生起)確率を計算

10

20

30

40

50

してもよく、出現確率にしたほうが評価関数としては精度のよいものとなる。

【0042】

(C) 難解文の平易文への変換例

図4は、難解文を平易文に変換した変換例を示している。図4に示す難解文の平易文への変換では、法律文章を平易な文に書き換えたり、難しい新聞の記事を小学生向けの易しい文に書き換えたりする処理を行う。

【0043】

例えば図4の例のように、変換対象文10として、「大臣を罷免する」という文が入力されたとする。変形処理部11は、この変換対象文10を、変形規則記憶部15にあらかじめ用意された変形規則を用いて、異なる表現に言い換える。ここで、変形規則として、

10

「罷免する」 「やめさせる」

という規則があったとすると、変形処理部11は、変換対象文10に変形規則を適用することにより、「大臣を罷免する」という文から「大臣をやめさせる」という文を生成する。この他にも、種々の変形規則が存在し、多くの変形された文が候補として生成されることになる。これらの文を評価処理部12に渡す。なお、変形されなかった変換対象文10についても候補の一つとして評価処理部12に渡す。

【0044】

評価処理部12は、難解文変換用の評価尺度(評価関数)を用いて、変形処理部11が変形した文を評価する。ここで評価の尺度が、例えば小学生向けというような低年齢層向けの文章集合での出現頻度または出現確率が大きくなる変換をよしとするものである場合に、評価処理部12は、あらかじめ定められた範囲での低年齢層向けの文章集合における「大臣を罷免する」と「大臣をやめさせる」の出現頻度を求める。簡便な手法としては、変形した部分を含む小さい領域範囲の文字列が言語データで何回出現したかを数える。「大臣をやめさせる」のほうが出現頻度が大きい場合、この表現のほうが高年齢層向けの易しい表現であるとわかる。これによりこの変形はよしとされ、変換結果文17として「大臣をやめさせる」が出力される。なお、出現頻度ではなく、出現(生起)確率を計算してもよいことは、前述した例と同様である。

20

【0045】

(D) 特定個人文体への変換例

図5は、特定の個人文体への変換例を示している。図5に示す特定個人文体への変換では、例えば芥川龍之介の小説を、夏目漱石の文体の小説に書き換えたり、ある無名の作家の小説をシェークスピアの文体の小説に書き換えたりする処理を行う。

30

【0046】

例えば図5の(1)の例のように、変換対象文10として、「大臣を罷免するなどを行った」という文が入力されたとする。変形処理部11は、この変換対象文10を、変形規則記憶部15にあらかじめ用意された変形規則を用いて、異なる表現に言い換える。ここで、変形規則として、

「するなど」 「するといったこと」

40

という規則があったとすると、変形処理部11は、変換対象文10に変形規則を適用することにより、「大臣を罷免するなどを行った」という文から「大臣を罷免するといったことを行った」という文を生成する。この他にも、種々の変形規則が存在し、多くの変形された文が候補として生成されることになる。これらの文を評価処理部12に渡す。なお、変形されなかった変換対象文10についても候補の一つとして評価処理部12に渡す。

【0047】

評価処理部12は、特定個人文体への変換用の評価尺度(評価関数)を用いて、変形処理部11が変形した文を評価する。ここで評価の尺度が、変換目的である特定個人の文章集合での出現頻度または出現確率が高くなるような表現をよしとするものである場合に、評価処理部12は、その特定個人の文章集合における「大臣を罷免するなどを行った」とい

50

う文や、「大臣を罷免するといったことを行った」という文の出現頻度を求める。なお、出現頻度は、必ずしも文全体の出現回数でなくてもよく、変形した部分を含む小さい領域範囲の文字列が文章集合の中で何回出現したかでもよい。「大臣を罷免するといったことを行った」という文の出現頻度が大きい場合、評価処理部 12 は、変換結果文 17 として「大臣を罷免するといったことを行った」を出力する。

【0048】

また、例えば変形規則として、
「と思われる」「であろう」
・・・

という規則をあったとする。ある文章を、「であろう」を多用する人の文体に変換することを考える。この場合、評価の尺度として、その「であろう」を多用する特定個人の文章集合での出現頻度または出現確率が高くなるような表現をよしとするものを用いる。

10

【0049】

変形処理部 11 は、図 5 の(2)のように「大臣を罷免すると思われる」という変換対象文 10 を入力すると、この入力に対して変形規則を適用することにより、この文を「大臣を罷免するであろう」という表現に変形する。評価処理部 12 による評価によって、「大臣を罷免するであろう」という表現の評価値が最も高いことがわかると、評価処理部 12 はこの文を変換結果文 17 として出力する。

【0050】

(E) 質問応答システムのための変換例

20

図 6 は、質問応答システムのための変換例を示している。図 6 に示す変換では、与えられた質問文の答えが書いてありそうな文を、新聞記事、各種論文、百科事典その他の所定の知識データベースから探し出し、その答えが書いてありそうな文と質問文との類似度が大きくなるように双方を書き換えて照合し、答えが書いてありそうな文での、質問文の疑問詞に対応している箇所を答えとして出力するといったことを行う。

【0051】

この質問応答システムでは、類似度を尺度として言い換えを行っていることになる。類似度が高くなるように言い換えを行うことで質問文と回答を含むデータとの照合がしやすくなる。

【0052】

30

本システムに入力される変換対象文 10 は、質問文と、回答が含まれる文の候補となる知識データベースの文である。ユーザからの質問文が、例えば「日本国の首都はどこであるか」であり、知識データベース中にある文が、「東京は日本の首都である」であったとする。

【0053】

変形処理部 11 は、これらの二つの変換対象文 10 を、それぞれ変形規則記憶部 15 にあらかじめ用意された変形規則を用いて、異なる表現に言い換える。ここで、変形規則として、図 6 に示すように、

1 「XはYである」「YはXである」

2 「日本国」「日本」

40

・・・

があったとする。

【0054】

1 の変形規則により「東京は日本の首都である」から「日本の首都は東京である」という文が生成される。また、2 の変形規則により「日本国の首都はどこであるか」から「日本の首都はどこであるか」が生成される。ここでは、簡単な変形規則を例示したが、通常の変形処理では、変形された文のさらなる変形というように、多段に変形が繰り返されることになる。これらの変形した文の結果が評価処理部 12 に引き渡される。

【0055】

評価処理部 12 では、質問応答システム用の評価尺度(評価関数)を用いて、変形処理部

50

11が変形した文を評価する。ここで評価の尺度が、入力した二つの文の類似度であり、類似度が大きくなる変換が評価が高いとすると、多くの変形の中から二つの文の類似度が高いものが選ばれることになる。

【0056】

変形された質問文と知識データベースの文の中で類似度が最も高いものが、「日本の首都はどこであるか」と「日本の首都は東京である」であったとすると、この変換はよしとされ、これらの二つの文が変換結果文17として出力される。これらの文から「どこ」と「東京」が対応することがわかり、質問応答システムから質問文に対する回答として、「東京」または「日本国の首都は東京である」がユーザに出力されることになる。

【0057】

質問応答システムにおいて類似度を尺度として言い換えを行った例を説明したが、同様に情報検索においても類似度を尺度とした変換を利用することができる。この場合、検索のクエリと検索される記事との類似度が高くなるように言い換えてから、クエリと記事との類似度を求める。

【0058】

照応の問題でも、「近くの大きな杉の木の根元にある穴」と「杉の木の根元の穴」の同一性もしくは包含関係が判定できないと照応を解決できないというのがあるが、類似度を尺度として両者を言い換え、例えば「近くの大きな杉の木の根元の穴」と「杉の木の根元の穴」になった場合、後者が前者に含まれることがわかり、後者が前者を指示可能となる。

【0059】

以上の変換例の他に、例えば入力データが書き言葉（文章語）のときに、出現頻度や出現確率などによる変換候補の評価に用いるコーパスとして話し言葉（口語）の言語データを用いると、書き言葉を話し言葉へ変換するシステムが実現され、またこの逆に、変換候補の評価に用いるコーパスとして書き言葉の言語データを用いると、話し言葉を書き言葉に変換するシステムを実現することもできる。

【0060】

また、ある自然言語で記述された文を他の自然言語で記述された文に変換する機械翻訳にも、次のように適用することができる。機械翻訳への適用の場合、変形規則記憶部15に記憶する変形規則として翻訳規則を入れ、評価処理部12では、ターゲット側の自然言語の言語コーパスを用いて、その言語コーパスにおける出現頻度または出現確率などを評価尺度として用いる。

【0061】

本発明では、以上説明したような個々の評価尺度を同時に複数利用して与えられた文字列を変換することを可能にする。さらに、どの評価尺度を重視した変換を行うかについても指定可能にする。

【0062】

例えばある文書を文内圧縮（要約）し、かつ推敲することを考える。文内圧縮（要約）の評価尺度は、文の長短であった。また、推敲の評価尺度は、大量な用例の言語コーパスでの出現頻度（出現確率）であった。

【0063】

このとき、ユーザは、要約のほうを推敲よりも重視したいとする。ユーザは、評価尺度として文の長短と、大量な用例の言語コーパスでの出現頻度（出現確率）を選択し、文の長短の評価尺度としての重みを大きく設定する。例えば評価尺度を、1 「長さを短くする」、2 「長さが同じ場合には言語コーパスでの出現確率を高くする」というように設定してもよい。

【0064】

また、ユーザが推敲のほうを要約よりも重視したいというような場合には、例えば評価尺度を、1 「言語コーパスでの出現確率を高くする」、2 「言語コーパスでの確率が同じ場合には長さを短くする」と設定すればよい。

【0065】

10

20

30

40

50

さらに、はっきりとどちらを重視するという指定ではなく、適当な比率 t を用いて、評価尺度を「(文の長さの逆数) × (コーパスでの出現確率)^t」と設定するというような実施も可能である。

【0066】

同様に、機械翻訳し、その要約を生成するといった変換や、口語表現を文章語表現に変換し、さらにその要約を生成するといった変換についても、それぞれに適した評価尺度の選択によって、容易に実現することができる。前者の変換の場合、機械翻訳における目的言語のコーパスでの出現確率、要約での文の長さを評価尺度として用いればよい。また、後者の変換では、文章語のコーパスでの出現確率と、文の長さを評価尺度として選択すればよい。

10

【0067】

「要約」と「口語文章語変換」と「推敲」を同時に行う変換の例を説明する。口語文章語変換と推敲は、文章語のコーパスの確率を評価尺度として用いる。これらは、同じ評価尺度を利用することも可能である。要約は、文の長さを評価尺度とする。ここでは、両方の尺度を、以下のような重みづけで用いることにする。 t はユーザが指定することができる重要度のパラメータである。

【0068】

・総合評価尺度：(文の長さの逆数) × (コーパスでの出現確率)^t
 変換対象文 10 として、講演会での次のような話し言葉が入力されたとする。
 (入力)「今日はえー単語を意味でソートすることについてお話しします。」
 変形規則としては、次のような規則が変形規則記憶部 15 に登録されていたとする。

20

【0069】

「えー」 * (削除)
 「AをBでCする」 「AのBのC」
 「お話しします」 「お話しする」
 「お話しする」 「述べる」
 「Aについて述べる」 「Aを述べる」
 「AのB」 「AB」
 「今日は」 「今日」

... ..

30

これらの変形規則を用いることで、要約と口語文章語変換を同時に行うことができる。これらの変形規則により多くの変換の候補が作成されることになり、その変換の候補の選択は、要約と口語文章語変換の各評価尺度を用いて行われ、総合的に評価結果のよい候補が変換結果文 17 として出力される。

【0070】

「今日はえー単語を意味でソートすることについてお話しします。」という変換対象文 10 に対して、上記変形規則を適用することにより、多数の変換の候補が得られるが、ここで、要約の評価尺度だけを使った場合には、
 「今日単語意味ソートを述べる。」
 という変換の候補が、変換結果文 17 として選択される。この文は「今日」「単語意味ソート」の部分が読みにくい文となっている。

40

【0071】

一方、口語文章語変換と推敲の評価尺度だけを使った場合には、
 「今日は単語を意味でソートすることについて述べる。」
 という変換の候補が、変換結果文 17 として選択される。この文は、それほど短い文になっていない。

【0072】

ここで、要約と口語文章語変換・推敲の両方の評価尺度を同時に用いることで
 「今日は単語の意味ソートを述べる。」
 という変換の候補が、変換結果文 17 として選択され、簡潔で適切な表現が得られること

50

になる。

【 0 0 7 3 】

図 7 は、図 1 に示す言語変換処理装置 1 の処理フローチャートである。変形処理部 1 1 は、まずステップ S 1 0 により、言い換え対象として指定された変換対象文 1 0 を入力する。キーボードなどからの入力、ファイルからの入力、アプリケーションプログラムからの入力など、入力方法は問わない。

【 0 0 7 4 】

ステップ S 1 1 では、変形規則記憶部 1 5 から変換に必要な変形規則を読み込む。既に読み込まれている場合には、ここでの読み込みは不要である。次に、ステップ S 1 2 では、入力した変換対象文 1 0 を変形規則を用いて変形する。ここで変形規則は、適用可能なものを繰り返し適用することができ、変形規則の適用によって多数の変換の候補が生成されることになる。それらの変換の候補を作業用メモリに蓄積する。

10

【 0 0 7 5 】

ステップ S 1 3 では、ユーザに使用する評価尺度を選択させるために、評価尺度の選択画面を表示する。図 8 に、評価尺度選択画面の例を示す。評価尺度の選択項目として、「短い文」「よく使われる表現」「平易な文」「著名な作家の文体」「口語の表現」「文章語の表現」「二つの文の類似度」...などがあり、これらは複数の選択が可能である。「著名な作家の文体」が選択された場合には、さらに作家名の入力が必要とされる。

【 0 0 7 6 】

ステップ S 1 4 では、選択終了のボタンがマウス等によりクリックされると、そのときに選ばれている評価尺度の選択情報を入力する。図 8 の例では、「短い文」「著名な作家（夏目漱石）の文体」が選択されており、ユーザは、変換対象文 1 0 を夏目漱石の文体で短い文（要約）に変換することを指定している。

20

【 0 0 7 7 】

次に、ステップ S 1 5 では、選択された各評価尺度の重要度指定画面を表示する。図 9 に、評価尺度の重要度指定画面の例を示す。図 9 (A) の画面は、評価尺度として「短い文」「よく使われる表現」「平易な文」が選択された場合の例であり、これらの評価尺度に対応するつまみをマウス等によりスライドさせる操作によって、各評価尺度の重要度を指定することができるようになっている。

【 0 0 7 8 】

また、図 9 (B) の画面は、評価尺度として「短い文」と「平易な文」の二つが選択されたときの重要度指定画面であり、つまみを左側へ動かせば、相対的に「短い文」であることが重視され、つまみを右側へ動かせば、「平易な文」であることが重視される。

30

【 0 0 7 9 】

ステップ S 1 6 では、図 9 (A) または (B) の画面において「OK」のボタンが押されたときに、その時のつまみの位置から各評価尺度に対する重要度を決定する。

【 0 0 8 0 】

続いてステップ S 1 7 では、ステップ S 1 2 で生成された変形後の表現の各候補について、ステップ S 1 4 で入力した各評価尺度（評価関数）を用いて評価する。ここでは、評価値が数値として算出されることになる。

40

【 0 0 8 1 】

次に、ステップ S 1 8 では、ステップ S 1 6 で入力した各評価尺度の重要度に従って総合的な評価結果を算出する。例えば各評価尺度による評価値が v_1, v_2, \dots, v_n であり、各評価尺度の重要度が t_1, t_2, \dots, t_n であったときに、総合的な評価値 V を、
$$V = t_1 \times v_1 + t_2 \times v_2 + \dots + t_n \times v_n$$
 というような式によって計算してもよいし、
$$V = v_1^{t_1} \times v_2^{t_2} \times \dots \times v_n^{t_n}$$
 というような式によって計算してもよい。

【 0 0 8 2 】

ステップ S 1 9 では、総合的な評価値 V の最も大きい変形後の表現（変換の候補）を選択

50

し、それを変換結果文17として出力する。その後、ステップS20では、評価尺度を変更するかどうかをユーザに問い合わせる。ユーザは、出力された変換結果文17が満足できる表現になっていれば、評価尺度の変更は指示しないで、処理を終了させる。もし、出力された変換結果文17が満足できる表現になっていなければ、評価尺度の変更を指示することができる。

【0083】

ステップS21では、評価尺度の変更の指示があると、ステップS13へ制御を戻し、評価尺度の選択処理から変換結果文17の出力まで同様に処理を繰り返す。これにより、ユーザは望ましい結果が得られるまで、評価尺度またはその重要度を変えて、対話的に変換の処理を進めることができる。

【0084】

【発明の効果】

以上説明したように、本発明によれば、種々の言語変換処理を同義表現の変形規則と評価尺度を用いて行い、各種の文(文章)の言い換えを行うことができるようになる。このとき、言い換えの目的に応じて変える部分は、評価尺度の部分だけである。このため複数の変換内容を含む変換、例えば要約と推敲を同時に行うというような変換を、複数の評価尺度を選択し、複合した意味合いの尺度を用いることで簡単に実現することができる。これはユーザにとって好ましい変換を行うことができるという利点があり、システム設計においても柔軟で多様な言語変換機能を簡易な構成で提供することができるという大きな利点がある。また、出力の精度も、要約生成システムにより要約し終わったものを推敲システムにより推敲するという2段階の処理構成をとる場合よりも高くなることが期待できる。

【図面の簡単な説明】

【図1】本発明のシステム構成例を示す図である。

【図2】文内圧縮の変換例を示す図である。

【図3】文章推敲の変換例を示す図である。

【図4】難解文を平易文に変換した変換例を示す図である。

【図5】特定の個人文体への変換例を示す図である。

【図6】質問応答システムのための変換例を示す図である。

【図7】言語変換処理の処理フローチャートである。

【図8】評価尺度選択画面の例を示す図である。

【図9】評価尺度の重要度指定画面の例を示す図である。

【符号の説明】

- 1 言語変換処理装置
- 2 入出力装置
- 10 変換対象文
- 11 変形処理部
- 12 評価処理部
- 13 評価尺度選択部
- 14 評価重要度設定部
- 15 変形規則記憶部
- 16 評価情報記憶部
- 17 変換結果文

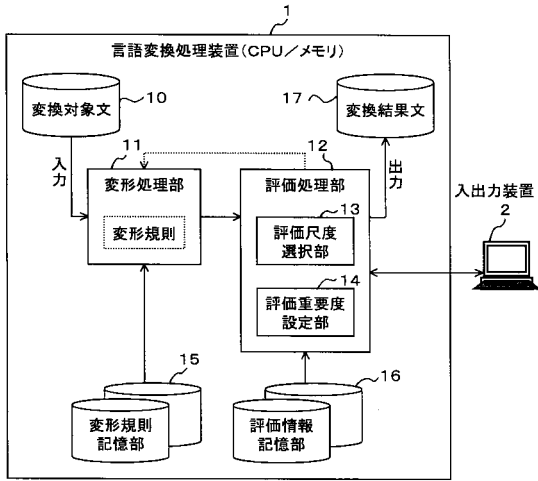
10

20

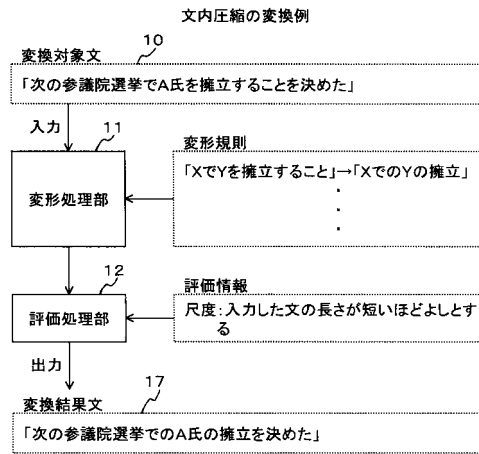
30

40

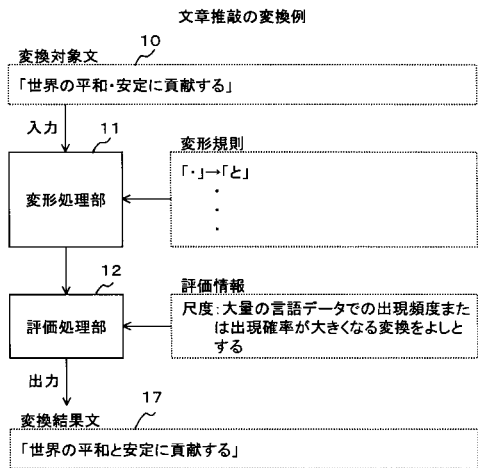
【図1】



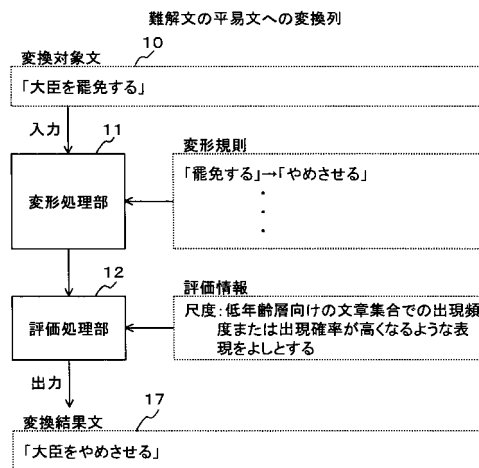
【図2】



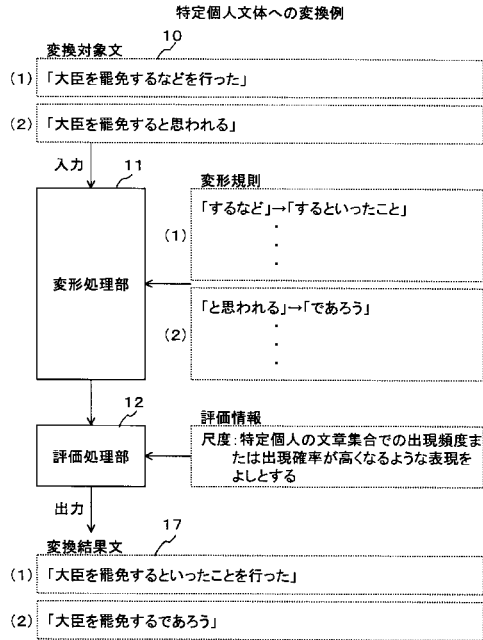
【図3】



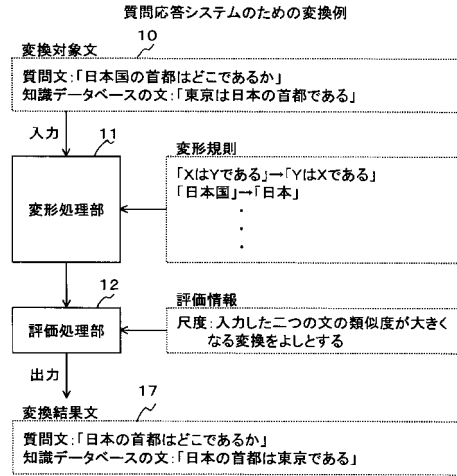
【図4】



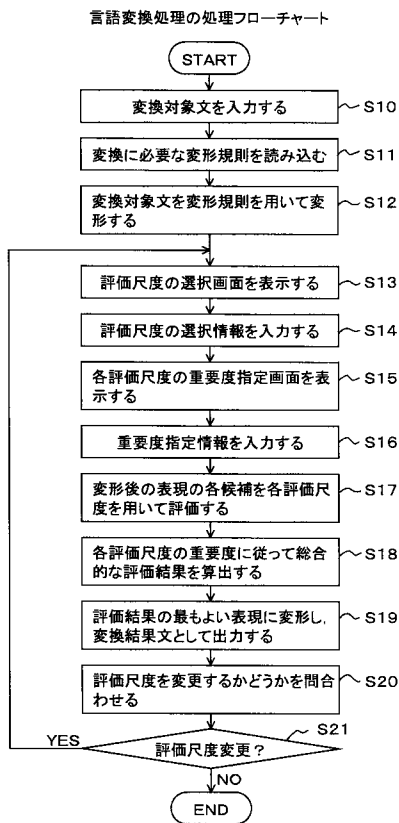
【図5】



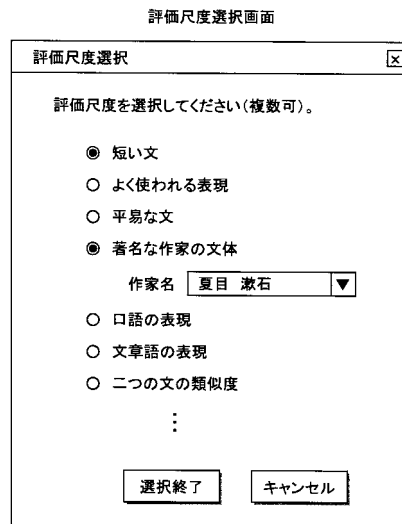
【図6】



【図7】



【図8】



【 図 9 】

評価尺度の重要度指定画面

(A) 評価尺度の重要度

	小	中	大
短い文	----- ----- -----	----- ----- -----	----- ----- -----
よく使われる表現	----- ----- -----	----- ----- -----	----- ----- -----
平易な文	----- ----- -----	----- ----- -----	----- ----- -----

OK キャンセル

(B) 評価尺度の重要度

短い文	←-----→	平易な文
-----	---------	------

OK キャンセル

フロントページの続き

合議体

審判長 立川 功

審判官 野崎 大進

審判官 長 由紀子

- (56)参考文献 特開昭62-78674(JP,A)
特開平3-122760(JP,A)
特開平5-12329(JP,A)
特開平6-139231(JP,A)
特開平9-274615(JP,A)
特開平10-254877(JP,A)
特開平11-238060(JP,A)
特開2000-250907(JP,A)

- (58)調査した分野(Int.Cl., DB名)

G06F17/21-17/28