(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第3721397号

(P3721397)

(45)発行日　平成17年11月30日(2005.11.30)　(24)登録日　平成17年9月22日(2005.9.22)

(51)Int.Cl.[7]	F I		
G06F 17/21	G06F	17/21	550L
G06F 3/16	G06F	3/16	320H
G06F 17/28	G06F	17/28	X
G10L 15/00	G10L	3/00	551B
G10L 15/18	G10L	3/00	537D

請求項の数 4　(全 13 頁)　最終頁に続く

(21)出願番号　特願2001-324096(P2001-324096)
(22)出願日　平成13年10月22日(2001.10.22)
(65)公開番号　特開2003-132047(P2003-132047A)
(43)公開日　平成15年5月9日(2003.5.9)
審査請求日　平成13年10月22日(2001.10.22)

(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　100130111
弁理士　新保　斎
(74)代理人　100090893
弁理士　渡邊　敏
(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

審査官　水野　恵雄

最終頁に続く

(54)【発明の名称】話し言葉の書き言葉への変換装置

(57)【特許請求の範囲】

【請求項１】

話し言葉の入力手段と、

前記入力手段により入力された話し言葉を書き言葉に変換する言葉変換手段と、

前記言葉変換手段により変換された書き言葉を出力する出力手段とを有する話し言葉の書き言葉への変換装置であって、

入力手段は、

（１）話し言葉のデータを入力すると共に、話し言葉データを所定のファイル形式で記録媒体に保存する動作を行い、

言葉変換手段は、

（２）データ読み出し部において該記録媒体から話し言葉データを読み出し、

（３）次いで形態素列分解部において該話し言葉データを形態素の羅列に分解し、

（４）次いで文頭から順次着目する形態素で始まる形態素列について、（４－１）変換候補抽出部が、予め記録媒体に備えた書き言葉・話し言葉間の言い換えテーブル中の話し言葉文字列と一致する被変換候補の形態素列を抽出すると共に、該言い換えテーブルにおいて話し言葉文字列に対応する書き言葉文字列を変換候補として読み出し、メモリ上に保持する処理と、（４－２）頻度測定部において、当該着目中の形態素列の前後所定個数の形態素が予め記録媒体に備えた書き言葉データベース中で、当該被変換候補を挟んで並ぶ頻度と、当該変換候補を挟んで並ぶ頻度とを測定して比較する処理と、（４－３）変換処理部が、該比較において後者の頻度が大きい時に着目中の形態素列をメモリ上の変換候補

で置き換える処理、の各処理（４－１）～（４－３）を文末の形態素まで行い、
出力手段は、
（５）少なくとも所定のファイル形式によって置き換えられた書き言葉データを記録媒体に保存する動作を行う
ことを特徴とする話し言葉の書き言葉への変換装置。
【請求項２】
前記言葉変換手段における言い換えテーブルを、
コンピュータのプロセッサにおけるデータ照合部により、記憶媒体内に予め記録された対を成す話し言葉データベース及び書き言葉データベースから不一致部分及び一致部分の検出を行うと共に、
確率演算部により、
前一致部分、不一致部分、後一致部分となる文字列の組合せにおいて、前一致部分については後方所定の文字数以内に後一致部分の文字列が現れる確率と、後一致部分については前方所定の文字数以内に前一致部分の文字列が現れる確率とのそれぞれ余事象同士を乗じ、当該不一致部分が差分として確からしい確率を算出する第１のステップ及び、
不一致部分が同一の全ての前一致部分及び後一致部分について該第１のステップにおける確率の余事象の直積を演算し、その余事象を当該不一致部分が、差分として確からしい確率として算出する第２のステップ
を処理し、
所定の確率値を超えた差分部分だけを備えることにより構成した
ことを特徴とする請求項１記載の話し言葉の書き言葉への変換装置。
【請求項３】
前記話し言葉の書き言葉への変換装置における言い換えテーブルが、確率演算部の第１のステップにおける処理において、
前一致部分、不一致部分、後一致部分となる文字列の組合せを取り出す際に、該不一致部分がその内部に一致部分である文字列を含んで構成される
ことを特徴とする請求項２記載の話し言葉の書き言葉への変換装置
【請求項４】
前記入力手段が、
マイク又は音声再生装置から話し言葉の音声波形をコンピュータのプロセッサにおける音声処理部に入力する音声入力部と、
入力された音声波形を音素認識して記号化データを記憶し、該記号化データをセグメンテーション処理により音声単位データに分割し、該音声単位データから単語認識して話し言葉データに変換する処理を少なくとも行う音声処理部と、
該話し言葉データをファイルとして記録媒体に保存する保存部とを備える
請求項１ないし３記載の話し言葉の書き言葉への変換装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
この発明は、入力された話し言葉を書き言葉に変換し、変換後の書き言葉により表された文章等の保存、表示、プリント等の出力機能を備えた話し言葉の書き言葉への変換装置に関するものである。
【０００２】
【従来の技術】
話し言葉と書き言葉には違いがあり、例えば、書き言葉の「データ」、「え」と「＝」が、話し言葉でそれぞれ「データー」、「えー」と「は」となったり、また話し言葉では「という」をいれて柔らかく言う場合等がある。話し言葉と書き言葉の差分を抽出し、前記のような話し言葉と書き言葉の違いを調べ、その差分結果により作成した書き言葉から話し言葉への変形規則を用い、書き言葉から話し言葉へ自動で言い換える方法に関しては、既に論文等で発表された周知の技術である。また、前記変形規則を用い、話し言葉から

書き言葉へ自動で言い換えることも同様な方法により可能である。
【０００３】
前記話し言葉から書き言葉への変換は、様々な人の話し言葉を誰でも容易に理解できる書き言葉として残したり、話し言葉で入力したものを書き言葉の原稿として容易に得たり、テレビ、映画等において話す言葉を逐次画面の端等に表示する等に利用される。
【０００４】
また、話し言葉の入力手段としては、パソコン等に搭載された音声認識機能等があり、マイクを通して入力された話し言葉の音声から、単語、文章等の話し言葉データに変換する方法等、他にも多種多数の従来技術がある。前記変換後の単語、文章等の話し言葉データは、パソコン等のメモリに一時記憶あるいはハードディスク等の記録媒体に保存され、必要に応じた処理が行われる、尚、話し言葉は、マイクを通してのみではなく、磁気テープ等の記録媒体に記録された音声を再成する装置からの出力をパソコン等に入力し、前記同様に話し言葉データとして得られる場合もある。
【０００５】
また、パソコン等のキーボードにより、音声を聞きながらあるいは映画の台本、原稿等に話し言葉で記述された文章等をパソコン等に入力する方法もある。前記入力された文章等の話し言葉データは、パソコン等のハードディスク等の記録媒体に保存され、必要に応じて処理される。
【０００６】
尚、前記パソコン等では、前記記録媒体に記録する出力形態ばかりでなく、前記記録された文章等の話し言葉データを、専用のソフトウェアあるいは市販のワードプロセッサー等のソフトウェアを用いてＣＲＴ等の表示装置に表示したり、プリンターによりプリントとして出力することができる。
【０００７】
【発明が解決しようとする課題】
しかしながら、前記従来のパソコン等では、入力された単語、文章等の話し言葉データが、そのまま話し言葉の単語、文章等として出力されるものであり、入力された話し言葉が、書き言葉に変換されて出力されるものがないという問題点があった。
【０００８】
本発明は、前記従来の問題点を解決するためになされたもので、話し言葉で入力した文章等を書き言葉の文章等で出力することができる話し言葉の書き言葉への変換装置を提供することである。
【０００９】
【課題を解決するための手段】
本発明の請求項１に記載の発明によれば、話し言葉の入力手段と、前記入力手段により入力された話し言葉を書き言葉に変換する言葉変換手段と、前記言葉変換手段により変換された書き言葉を出力する出力手段とを有する話し言葉の書き言葉への変換装置を提供する。そして、入力手段は、（１）話し言葉のデータを入力すると共に、話し言葉データを所定のファイル形式で記録媒体に保存する動作を行う。
言葉変換手段は、（２）データ読み出し部において該記録媒体から話し言葉データを読み出し、（３）次いで形態素列分解部において該話し言葉データを形態素の羅列に分解し、（４）次いで文頭から順次着目する形態素で始まる形態素列について、（４－１）変換候補抽出部が、予め記録媒体に備えた書き言葉・話し言葉間の言い換えテーブル中の話し言葉文字列と一致する被変換候補の形態素列を抽出すると共に、該言い換えテーブルにおいて話し言葉文字列に対応する書き言葉文字列を変換候補として読み出し、メモリ上に保持する処理と、（４－２）頻度測定部において、当該着目中の形態素列の前後所定個数の形態素が予め記録媒体に備えた書き言葉データベース中で、当該被変換候補を挟んで並ぶ頻度と、当該変換候補を挟んで並ぶ頻度とを測定して比較する処理と、（４－３）変換処理部が、該比較において後者の頻度が大きい時に着目中の形態素列をメモリ上の変換候補で置き換える処理、の各処理（４－１）～（４－３）を文末の形態素まで行う。

出力手段は、（５）少なくとも所定のファイル形式によって置き換えられた書き言葉データを記録媒体に保存する動作を行うことを特徴とする。

また、請求項２に記載の発明は、前記言葉変換手段における言い換えテーブルを、コンピュータのプロセッサにおけるデータ照合部により、記憶媒体内に予め記録された対を成す話し言葉データベース及び書き言葉データベースから不一致部分及び一致部分の検出を行うと共に、確率演算部により、前一致部分、不一致部分、後一致部分となる文字列の組合せにおいて、前一致部分については後方所定の文字数以内に後一致部分の文字列が現れる確率と、後一致部分については前方所定の文字数以内に前一致部分の文字列が現れる確率とのそれぞれ余事象同士を乗じ、当該不一致部分が差分として確からしい確率を算出する第１のステップ及び、不一致部分が同一の全ての前一致部分及び後一致部分について該第１のステップにおける確率の余事象の直積を演算し、その余事象を当該不一致部分が、差分として確からしい確率として算出する第２のステップを処理し、所定の確率値を超えた差分部分だけを備えることにより構成する。

さらに、請求項３に記載の発明は、前記話し言葉の書き言葉への変換装置における言い換えテーブルが、確率演算部の第１のステップにおける処理において、前一致部分、不一致部分、後一致部分となる文字列の組合せを取り出す際に、該不一致部分がその内部に一致部分である文字列を含んで構成されることを特徴とするものである。

請求項４に記載の発明によると、前記入力手段が、マイク又は音声再生装置から話し言葉の音声波形をコンピュータのプロセッサにおける音声処理部に入力する音声入力部と、入力された音声波形を音素認識して記号化データを記憶し、該記号化データをセグメンテーション処理により音声単位データに分割し、該音声単位データから単語認識して話し言葉データに変換する処理を少なくとも行う音声処理部と、該話し言葉データをファイルとして記録媒体に保存する保存部とを備える話し言葉の書き言葉への変換装置を提供することができる。

【００１０】

【発明の実施の形態】

以下、図面を参照して、本発明の実施形態について説明する。図１は、本発明による話し言葉の書き言葉への変換装置の実施形態を示す図である。

【００１１】

図１に示すように、本発明の実施形態の話し言葉の書き言葉への変換装置１０は、入力手段１１と、言葉変換手段１２と、出力手段１３とを有するコンピュータとして構成されている。コンピュータには公知のように、各種プロセッサ（特にＣＰＵ）やそれと連動するメモリ、ハードディスクなどが備えられており、これらがソフトウェアにより動作する。

入力手段１１は、話し言葉で記述された文章等、例えば、映画の台本、講演記録、会話の記録等を直接入力するキーボード、あるいはマイクから音声として入力若しくは磁気テープ等の記録媒体１４に記録された話し言葉の単語、文章等を再生する再生装置の出力を前記マイクと同様に音声として入力し、前記入力された音声又は再生装置の出力からの音声を単語、文章等の話し言葉データに変換するソフトウェアからなるものである。

【００１２】

前記入力された音声又は再生装置の出力からの音声を話し言葉データに変換する手段として、専用の変換ソフトウェアを用いる場合は、既知の技術である次に示す音声認識による音声処理技術を利用することができる。

入力手段１１に、例えば一般的な音声処理技術を実装し、マイクや公知の入力端子である音声入力部により入力された音声波形が、ＣＰＵの音声処理部において音素認識（記号化）、音声単位へのセグメンテーションする。さらに、音響分析等の音響処理を行い、次に予測単語の認識、重要な自立語などのキーワードをボトムアップ的に抽出するワードスポッティング等の単語認識を行い、最後に意味解析、構文解析等の言語処理を行い話し言葉データを生成するものである。

そして、ＣＰＵの保存部の作用によって、話し言葉データは所定のファイル形式で記録

媒体１４に保存される。

【００１３】

尚、既存の音声認識システム等の音声から話し言葉データに変換するものを利用することも可能であり、この場合、言葉変換手段１２で用いるファイル形式として、前記既存の音声認識システムにおける文章等のファイル形式と共通のものを用いればよい。

【００１４】

言葉変換手段１２では、話し言葉と書き言葉の対応関係を規定する言い換え（以下パラフレーズという）テーブル１４ａを予め作成しておく。パラフレーズの作成は、日本語の講演発表の音声を書き起こし、形態素情報を付与した日本語話し言葉集成（以下コーパスという）を話し言葉データベース１４ｃとして、講演発表の元となる論文の電子化データを書き言葉データベース１４ｂとして、それぞれ用いる。即ち、話し言葉と書き言葉のデータを用いる。次に、話し言葉データベース１４ｃと書き言葉データベース１４ｂの一致部分及び不一致部分を調べ、話し言葉データから書き言葉データへの変換規則を獲得しておく。即ち、パラフレーズテーブル１４ａの作成に、話し言葉と書き言葉の差分部分のデータを用いる。

【００１５】

言葉変換手段１２では、図２に示すアルゴリズムを用いて話し言葉から書き言葉への言い換え（変換）を行う。このアルゴリズムは、大雑把に言うと、前記書き言葉データベース１４ｂでの頻度が大きくなるように書き換える、つまり、書き言葉データベース１４ｂで出現し易い表現に書き換えるものである。

【００１６】

先ず、入力された話し言葉データをＣＰＵのデータ読み出し部１２ａでハードディスク１４から読み出し、形態素列分解部１２ｂで形態素解析する。形態素解析処理については公知であり、例えば、形態素解析プログラムであるＪＵＭＡＮを用いて形態素解析して、形態素列に分解する。そして、文頭の形態素から順に、形態素ごとに次の処理を行う。

【００１７】

先ず、変換候補抽出部１２ｃにおいて、現在の形態素で始まる形態素列Ｓ（形態素を一つも持たない場合、つまり空文字列も含む）と、前記変換規則を得たときの差分データの話し言葉の文字列A_iが一致するものを抽出し、その場合における差分部分のデータR_iが規則として用いられる。そして、前記パラフレーズテーブル１４ａ中の差分部分データの書き言葉の部分（文字列B_i）を、書き言葉の候補としてメモリ１５上に保持する。また、Ｓの前節ｋ－ｇｒａｍの形態素列$S1_i$、Ｓの後節ｋ－ｇｒａｍの形態素列を$S2_i$とする。尚、前記ｋは、定数である。

【００１８】

次に、前記B_iに対して、頻度測定部１２ｄにより前記書き言葉データベースでの$S1_i B_i S2_i$の文字列の頻度を求め、該頻度が最も大きかった時のｉをｍとする。

【００１９】

次に、同じく頻度測定部１２ｄにより書き言葉データベースでの$S1_m A_m S2_m$の文字列の頻度を求め、この値よりも$S1_m B_m S2_m$の文字列の頻度の方が大きいか否か比較する。後者の頻度が大きい時には変換処理部１２ｅが、A_mをB_mに置き換え、処理を次の形態素に移す。

【００２０】

以上の処理を言葉変換手段１２の各処理部が文末の形態素まで行うことにより、話し言葉データから書き言葉データへの書き換え（変換）を行うことができる。

【００２１】

出力手段１３では、言葉変換手段１２で変換して得られた書き言葉データを、所定のファイル形式でハードディスク等の記録媒体１４に保存したり、ＣＲＴ等の表示手段に書き言葉データの表す文章等を表示したり、プリンターで前記文章等をプリントする等の出力を行う。

【００２２】

話し言葉の文章等の入力から書き言葉の文章等の出力まで、専用の処理プログラムを用いて行うように構成することもできるし、専用の処理プログラムと市販のワードプロセッサー等の処理プログラムとを用いて行うように構成することも可能である。

【００２３】

尚、前記表示手段として、テレビ画面、映画画面等であってもよい。また、前記所定のファイル形式を、市販のワードプロセッサーで取り扱えるファイル形式とすることにより、容易に変換後の書き言葉での文章等を表示したり、プリントしたりすることができる。

【００２４】

以上、本発明の実施形態の話し言葉の書き言葉への変換装置を用いると、映画あるいはテレビにおける台詞、解説等をほぼ同時に書き言葉として読みやすく、理解しやすい文章等で映画あるいはテレビの画面の端に表示できる等、気軽に使っている話し言葉から、形式的である故誰でも理解し易い書き言葉の文章等に容易に変換できる。

【００２５】

【実施例】

次に、言葉変換手段１２において用いる話し言葉と書き言葉の差分を調べた実施例について説明する。

【００２６】

差分をとる話し言葉と書き言葉のデータの形態素解析を行い、図３に示すように、各形態素が各行に分かれた形にデータを変形する。前記形態素解析は、形態素解析プログラムであるＪＵＭＡＮを用いて行う。

【００２７】

次に、前記データを照合し、話し言葉と書き言葉のデータの差分部分と一致部分の検出を行う。ここでは、ＵＮＩＸコマンドのｄｉｆｆを用いて行う。得られた結果を図４に示す。

【００２８】

図４において、セミコロンで始まる行は一致部分、差分部分を示すためのもので、「；●●●」から「；▼▼▼▼▼▼」までの部分は、話し言葉データでのみ出現したもの、「；▲▲▲▲▲▲」から「；●●●」までの部分は、書き言葉のみ出現したもの、「；▼▼▼▼▼▼」から「；▲▲▲▲▲▲」までの部分は、話し言葉と書き言葉でともに出現したものを意味する。ここで取り出したいものは、話し言葉と書き言葉の違いであるので、「；▲▲▲▲▲▲」から「；▼▼▼▼▼▼」までの部分となり、差分部分は図５のようになる。

【００２９】

前記図５の「え今日は」「本論文では単語の羅列を」は、話し言葉で「え今日は」とあったのが、書き言葉では「本論文では単語の羅列を」となったということを意味する。

この結果では、差分部分のデータとして精度が悪すぎるので、ｄｉｆｆの結果から、ある程度よさそうな話し言葉と書き言葉の差分部分を抽出する。ここでは、１．珍しい（出現頻度の低い）文字列に囲まれた不一致部分ほど、パラフレーズとしては確からしい、２．複数箇所に出現した不一致部分ほど、パラフレーズとしては確からしい、という二つの特徴を利用する。

【００３０】

先ず、前記特徴１「珍しい文字列に囲まれた不一致部分ほど、パラフレーズとしては確からしい」の方を考える。ここでは、差分部分（不一致部分）が、図６に示すように、一致部分である文字列Ｓ１、Ｓ２にはさまれていて、Ｓ１とＳ２の間がｄ文字以内に図７中の方向にＳ２及びＳ１が現れる確率を、Ｐ（Ｓ１）、Ｐ（Ｓ２）とすると、Ｐ（Ｓ１）とＰ（Ｓ２）は、それぞれ近似的に数１と数２で表される。

【００３１】

【数１】

$$P(S1) \simeq (d+1) * \frac{\text{文字列}S1\text{の出現数}}{\text{文字総数}}$$

【００３２】
【数２】

$$P(S2) \simeq (d+1) * \frac{\text{文字列}S2\text{の出現数}}{\text{文字総数}}$$

【００３３】
この時、差分部分が確からしい確率をＰ（差分、Ｓ１、Ｓ２）とすると、Ｐ（差分、Ｓ１、Ｓ２）はＳ１、Ｓ２がともに図６に示すような形で現れ難い確率であると仮定すると、Ｓ１とＳ２が独立であることを仮定して、数３のようになる。
【００３４】
【数３】

$$P(\text{差分}, S1, S2) \simeq (1 - P(S1))(1 - P(S2))$$

【００３５】
次に、前記特徴２「複数箇所に出現した不一致部分ほど、パラフレーズとしては確からしい」の方を考える。これは、複数箇所での確率をうまく組み合わせればよい。複数箇所のうち一箇所でも正しければ、その差分部分は正しいものとして抽出できると考える。即ち、差分部分が正しい事象は、任意のＳ１、Ｓ２に対してＳ１、Ｓ２に囲まれる差分部分が全て確からしくない場合の余事象なので、差分部分が確からしい確率をＰ（差分）とすると、Ｐ（差分）は、各差分部分が独立であることを仮定して、数４で表される。
【００３６】
【数４】

$$P(\text{差分}) \simeq 1 - \prod_{S1,S2} (1 - P(\text{差分}, S1, S2))$$

【００３７】
差分部分の取り出しは，ｄｉｆｆの結果を前記数４のＰ（差分）の値でソートし、その値の大きいものから取り出すことによって行われる。
【００３８】
尚、最初の差分部分の候補の取り出しについては、次に示す改良を行うことができる。図７に示すように、一致部分と差分部分が出現している時に、“「差分部分１」「一致部分１」「差分部分２」”、“「差分部分１」「一致部分１」「差分部分２」「一致部分２」「差分部分３」”といったものも差分部分の候補とする。
【００３９】
前記改良は、単に「差分部分１」だけでは、「一致部分０」「一致部分１」から求まるＰ（差分）の値が小さくて取り出せないような時も、“「差分部分１」「一致部分１」「差分部分２」を差分部分と考えることで、「一致部分０」「一致部分２」から求まるＰ（差分）の値が大きくなって取り出しうるという効果を持つ。ここでは、この連結によって生成する差分部分は、元の差分部分を３個以下しか含まないものに限る。
【００４０】
次に、話し言葉、書き言葉のデータとして、開放融合プロジェクトにおける８２編の学会講演の論文の電子版を利用した。話し言葉データは、開放融合コーパス（集成）の内、前記論文データに対応するもの（３３０６７９文字）である。書き言葉データは、前記８２編の論文データ（打ち込み、３５２６６０文字）である。
【００４１】
話し言葉データには、図８に示すタグが埋め込まれていたので、次の処理を施した。基本的に各タグのリストの第二要素をタグの代りに本分に埋め込む。例えば、“（Ｆ　あの

）”の場合、「あの」を本分の該当箇所に挿入する。但し、セミコロンで区切られているものについては、最後のものを、カンマで区切られているものについては、最初のものを用いる。

【００４２】

フィラーや言い直しなどは省いた方がよいとも考えられるが、あえてそのような表現も差分部分として抽出することを目的として残した。

【００４３】

書き言葉データとして用いる論文データには、表題や著者名、所属なども含まれているが、これらはそのまま残して利用した。

【００４４】

以上の条件で話し言葉データと書き言葉データの差分部分を抽出した。抽出総数を図９に示す。

図９における確率値は、数４で算出された値である。差分部分を前記数４で算出される値で分類した結果の上位５０個を図１０に示す。図１０における頻度は差分部分の出現回数である。

【００４５】

前記図１０における「データー」「データ」の食い違いは、コーパス（集成）の定義によるもので、書き言葉で「データ」と書くが、話し言葉で「データー」と伸ばして発音し易いということを意味しているものではない。また、話し言葉で“＜Ｃ＞”が得られているが、これはコーパスにおいて単語の途中を意味するタグで、これが得られてもあまり意味はない。その他目立つものとしては、「え」「えー」などのフィラーが検出できていたり、「＝」は「は」と読むということが分かったり、話し言葉では「という」をいれて柔らかくいう場合があることが分かる。

【００４６】

抽出された差分結果を分析したところ、主に次のものがあった。

１．表記ゆれ

表記ゆれの例を図１１に示す。これはコーパスの定義にも関係するところであるが、ここでは大規模コーパスを使用してコーパスで多く出現するものを自然な表現として定義した。

【００４７】

２．表記・読みを与えるもの

この例を図１２に示す。図１２により「＝」は「は」と読めばよいとか、「ｓ」は「秒」を意味するときと「Ｓ」を意味する時があるなどが分かる。

【００４８】

３．同義関係のもの

この例を図１３に示す。論文に書いていたことを少し違えて言ったりするために、図１３のような同義な意味を示す言い換え表現を獲得することができる。ここでは、研究がらみの同義表現が得られている。

【００４９】

４．口語調のもの

この例を図１４に示す。話し言葉で丁寧語にするものから、「。」と書いているところを「訳ですが」と文をつなげるものなども得られている。また、最後の行に「これ」が得られているが、これは「明瞭に発声したもの（これ）を」という形で使われていた。

【００５０】

５．省略をしているもの

この例を図１５に示す。話し言葉の方では「処理」を省いて言ってみたり、データの値を「１１．２５」を「１１．３」に丸めていってみたり、はしょっているところがある。

【００５１】

６．補完をしているもの

この例を図１６に示す。これは、前記「省略しているもの」の逆の例である。書き言葉

では「損失の平均」となっているが、「損失の値の平均」と「値」をいれて分かりやすいように言い換えている。また、値も正確に「七十五五デシベル」と言っている場合もある。一般に、話し言葉の方が書き言葉よりも省略が多いとされており、この場合は逆の現象である。

【００５２】

７．コーパスの誤り検出に関わるもの

この例を図１７に示す。もともと、話し言葉データ、書き言葉データ自体に誤りがあった場合、その部分が差分として得られることがある。１行目のデータは、「速報」を「速記」と誤ったものと思われる。この誤りは論文データをオンライン化する時に生じたものと思われる。また、話し言葉データの方にも誤りが見受けられる。最後の行のデータは、「死活」と「生活」どちらでも正しそうな感じがする。

【００５３】

以上、話し言葉と書き言葉の差分を取り出した。図１に示す言葉変換手段１２において、以上で得られたデータをデータベースとして利用する。

【００５４】

【発明の効果】

本発明によれば、話し言葉で入力した文章等を書き言葉の文章等で出力することができる話し言葉の書き言葉への変換装置が得られる。

【図面の簡単な説明】

【図１】　本発明による話し言葉の書き言葉への変換装置の実施形態を示す図である。

【図２】　図１の言葉変換手段で用いる話し言葉から書き言葉への言い換え（変換）のアルゴリズムである。

【図３】　書き言葉データと話し言葉データの形態素への分割を示す説明図である。

【図４】　書き言葉データと話し言葉データのdiffの結果を示す説明図である。

【図５】　差分部分の抽出結果を示す説明図である。

【図６】　差分の出現模式図である。

【図７】　差分部分の拡張を示す図である。

【図８】　話し言葉データに使用されているタグの説明図である。

【図９】　差分部分の抽出数の結果を示す図である。

【図１０】　話し言葉データと書き言葉データの照合結果の上位５０個を示す図である。

【図１１】　表記のゆれの例である。

【図１２】　表記・読みを与えるものの例である。

【図１３】　同義関係のものの例である。

【図１４】　口語調のものの例である。

【図１５】　省略しているものの例である。

【図１６】　補完しているものの例である。

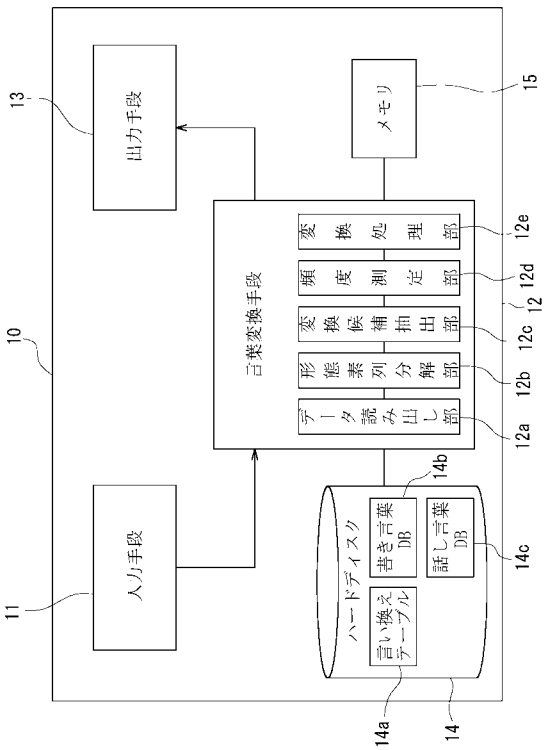
【図１７】　誤り検出の例である。

【符号の説明】

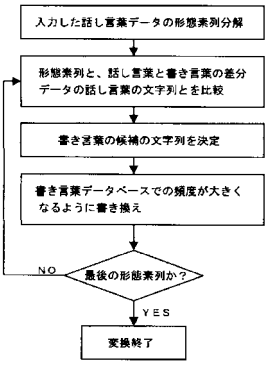
１０　話し言葉の書き言葉への変換装置
１１　入力手段
１２　言葉変換手段
　１２ａ　データ読み出し部
　１２ｂ　形態素列分解部
　１２ｃ　変換候補抽出部
　１２ｄ　頻度測定部
　１２ｅ　変換処理部
１３　出力手段
１４　記憶媒体
　１４ａ　言い換え（パラフレーズ）テーブル
　１４ｂ　書き言葉データベース

１４ｃ　話し言葉データベース
１５　メモリ

【図１】



【図２】



【図３】

話し言葉データ	書き言葉データ
え	本
今日は	論文
意味	で
ソート	は
に	単語
ついて	の
述べ	羅列
ます	を
え	意味
普通の	で
ソート	ソート
って	する
いう	と
の	いろいろな
は	とき
だいたい	に
え	便利である
５０	と
音	いう
順	こと
と	に
か	ついて
ＥＵＣ	記述
コード	する
順	。

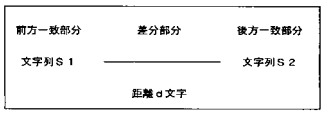
【図４】

；▲▲▲▲▲▲	；▼▼▼▼▼▼
本	ソート
論文	；▲▲▲▲▲▲
で	する
は	と
単語	いろいろな
の	とき
羅列	；●●●
を	；▼▼▼▼▼▼
；●●●	に
え	；▲▲▲▲▲▲
今日は	便利である
；▼▼▼▼▼▼	と
意味	いう
；▲▲▲▲▲▲	こと
で	に
；●●●	；●●●
（右欄につづく）	；▼▼▼▼▼▼

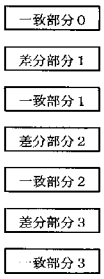
【図５】

話し言葉データ	書き言葉データ
え今日は	本論文では単語の羅列を で するといろいろなとき 便利であるということに

【図６】



【図７】



【図８】

タグ	説明	使用例
(F)	フィラー・感情表出系感動詞	(F あの)
(D)	言い直し	(D こ）これ
(W)	いい間違い	(W ミダリ；ヒダリ)
(?)	聞き取りなどに自身がない	(? タオンダー)
		(? あのー、あんのー)
(M)	音や言葉の引用	(M わ）は（M は）と表記
(O)	外国語や古語、方言など	(O ザッツファイン)
(R)	個人名など	(R 小林）さんが
(A)	基本型で漢字仮名以外の文字を使用する場合	(A イーユー；ＥＵ)
(K)	何らかの原因で漢字表記できなくなった場合	(K たち（F んー）ばな：橘)
(S)	未登録の口語表現が出現した場合	(S こりゃ)
(笑)、(泣)、(咳)、(あくび)	非言語音との共起	(笑 なにそれ)
(L)	さとやき声や独り言などの小さな声	(L アレコレナンダッケ)

【図９】

	抽出数
確率 99.99％以上	1,011
確率 99.9％以上	3,245
確率 99％以上	7,846
確率 90％以上	14,777
総数	72,835
頻度２以上	421

【図１０】

頻度	前方一致部分の例	話し言葉	書き言葉	後方一致部分の例
182	ＩＰＡＬの形容詞		、	形容動詞の
72	を母国語	の		キーワードで効率良く検索
43	ばたばた			ばたばた
49	の知識である	え		識別情報抽出法
56	ＬＲ表への複数	えー		接続制約の組み込みによる一般化ＬＲ法の
54	引き上げでＮＴＴ株は百万円になる		Ｊ	と予言したとか
39	に属するベクトル		の	和の記載に
28	本文間のハイパーリンク	を		自動生成
19	修飾要素		を	取り込む
21	ロストファイル群から構出した鉄込み語は		「	米映画
22	名詞の繰り返しの場合	えー	、	先頭の名詞のみを余剰語の挿入
20	Ｖは		）	Ｘと事情が違う
21	経験的損失は全学習	データー	データ	に対する損失の
11	および用言の	が	ガ	格情報を付与したコーパスの作成
19	計算言語学の学会発表論文	くの	２５５４元題材からなる	
13	大量のコーパスを用いて	え	、	機械翻訳により
15	空間数	に		より切り出された
12	原言語と目標言語		と	の語の同時進行性が
10	数値を示す言葉を捕って	―	Ｉ	つの文を
10	雑音スペクトル	を	の	減算
14	その結果を入手	で		修正していくプロジェクトについて
20	また頻度も少なくない	という		ことから
10	審組	の	を	最初から最後まで
10	合計頻度がＮ		＿｛	ｘ
11	記事の	その		構造と特徴
16	ニュース番組自動字幕化の	為	ため	の音声認識システムを
10	割引い	クラスター	クラスタ	が意味
6	ベクトルを異なる	Ｋ	ｋ	個の集合にクラスタリングした
8	ＶＱコードブックの	ニ	２	種類の結合モデル
7	述主主	は	＝	最適な結果が
8	その文中で	ん		話題となっている要素
9	十分良い精度で	は		推定できないという問題が
6	連結学習と	お		混合数を倍増する操作を繰り返す
11	ピッチの上昇	というもの		が有声音
7	基礎的	な	の	検討を
5	生起制限	が	／	厳しい制約も
5	技術域			広帯域ＣＥＬＰ
7	深さＤの	おー		上位概念に抽象化
7	をコンテキストに持つ	ような		言語モデルに
14	手法	というの		を提案
4	かぎ括弧部分が	掛かる	係る	語の制述を示す言葉ではないが
7	状態遷移をセットにして遺伝的操作を	行なう	行う	ことを提案し
4	ｅ		｝	以下のリンクを検出
4	瞬間的に逆向きの遷移	が	は	起こらないという
5	科学論文における要旨		―	本文間のハイパーリンク
4	最も結果	の	が	良かった混合数
4	良い意味合い		］「	悪い意味合い
6	概念が	あー		支配する概念数
4	が特って	る	いる	常識を
6	性能	が		改善

【図１１】

話し言葉	書き言葉
データー	データ
クラスター	クラスタ
パラメーター	パラメータ
モダリティー	モダリティ
データーベース	データベース
掛かる	係る
越える	超える
全て	すべて
為	ため
行なう	行う
言い替え	言い換え

【図１２】

話し言葉	書き言葉
は	＝
ニ	２
ゼロ	零
グラム	－ｇｒａｍ
秒	ｓ
Ｓ	ｓ
ヘブ	Ｈｅｂｂ

【図１３】

話し言葉	書き言葉
と	および
とか	や
研究	論文
異なりで	・異なり
それぞれ	各
ターム１	ｉ番目のターム
動詞	述語
認識	識別
違えば	違えば、

【図１４】

話し言葉	書き言葉
という	
いたしました	した。
ですね	、
です	
られます	られる。
っていうか	
とか	や
こう	
非常にこう	
います	いる。
分かりました	分かった。
ません	ない。
訳ですが	。
っていうの	
れるんですが	れた。
であって	であり、
訳ですけれども	ことである。
ありますけれども	ある。
なくとも、	なくても、
ないっと	ないと
これ	

【図１５】

前方一致部分	話し言葉	書き言葉	後方一致部分
スームージング		処理	を
各		Ｃ（Ｖ）＿｛ｋ｝	素片
スポーツニュース	の	における	会話部分を
平均時間が	１１．３	１１．２５	分まで

【図１６】

前方一致部分	話し言葉	書き言葉	後方一致部分
に対する損失の	値の		平均として
会話に	関しましては全然		不便はない
音圧レベル	七十五五デシベル	７０ｄＢ	で表示

【図１７】

前方一致部分	話し言葉	書き言葉	後方一致部分
話し言葉データでの誤り			
かぎ括弧の種類	表装的な	表層的な	情報のみで
書き言葉データでの誤り			
ニュース	速報	速記	記事
日本語が	延べ	述べ	１７８０９１語
日本語複数	名詞構造	名詞構造	解析法
どちらかのデータが誤り、もしくは不明			
マイクロホン	および声帯	及び生体	アンプ
社会の	生活	死活	にかかわる問題

フロントページの続き

(51)Int.Cl.[7] F I
Ｇ１０Ｌ 15/22 Ｇ１０Ｌ 3/00 ５６１Ｅ

(56)参考文献 特開２０００－０５７１４２（ＪＰ，Ａ）
特開平０５－０１２２４６（ＪＰ，Ａ）
特開平０２－０３６４６２（ＪＰ，Ａ）
特開２００３－１２２７４７（ＪＰ，Ａ）
村田真樹，井佐原均，diffと言語処理，電子情報通信学会技術研究報告，日本，電子情報通信学会，２００１年 7月１０日，Vol101,No190，pp29-36

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 17/21 550
G06F 3/16 320
G06F 17/28
G10L 15/00
G10L 15/18
G10L 15/22