

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3757289号

(P3757289)

(45) 発行日 平成18年3月22日(2006.3.22)

(24) 登録日 平成18年1月13日(2006.1.13)

(51) Int. Cl.	F I
G06F 17/28 (2006.01)	G06F 17/28 Z
G06F 17/30 (2006.01)	G06F 17/30 180A
	G06F 17/30 330C

請求項の数 3 (全 16 頁)

(21) 出願番号	特願2004-17846 (P2004-17846)	(73) 特許権者	301022471
(22) 出願日	平成16年1月27日(2004.1.27)		独立行政法人情報通信研究機構
(62) 分割の表示	特願2001-311329 (P2001-311329) の分割	(74) 代理人	100103827 弁理士 平岡 憲一
原出願日	平成13年10月9日(2001.10.9)	(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
(65) 公開番号	特開2004-178615 (P2004-178615A)	(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
(43) 公開日	平成16年6月24日(2004.6.24)		
審査請求日	平成16年1月27日(2004.1.27)		
特許法第30条第1項適用 平成13年7月10日 社 団法人電子情報通信学会発行の「電子情報通信学会技術 研究報告 信学情報 Vol. 101 No. 190」 に発表		審査官	和田 財太

最終頁に続く

(54) 【発明の名称】 言語処理システム

(57) 【特許請求の範囲】

【請求項1】

質問文を入力する入力手段と、

前記入力された質問文を平叙文に変換する質問文変換部と、

順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムを用いて前記平叙文に変換された質問文と知識データとの共通部分と差分部分を検出する差分検出手段と、

前記検出した共通部分と差分部分により、前記平叙文に変換された質問文と知識データとの類似度を求める類似度演算手段とを備え、

前記類似度演算手段は、前記類似度の最も大きい知識データの前記平叙文に変換された質問文の疑問詞と対をなして差分部分として得られた部分を出力することを特徴とした言語処理システム。

【請求項2】

前記差分検出手段で検出した差分の共通部分と不一致部分の数により、前記二つの言語データの類似度を求める前記類似度演算手段を備えることを特徴とした請求項1記載の言語処理システム。

【請求項3】

前記順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムとしてdiffコマンドを用いることを特徴とした請求項1又は2記載の言語処理システム。

10

20

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、差分検出を行うdiff(ディフ)コマンドのような順序情報を保持したまま一致部分が最大になるように複数のデータの対応づけを行うシステムを用いた言語処理システムに関する。

【背景技術】

【0002】

従来、例えば、日本語文と英語文の対応付け、講演と予稿の対応付け、質問文と知識データの対応付け等は、判断処理としてのプログラムが複雑なものであった。

10

【発明の開示】

【発明が解決しようとする課題】

【0003】

上記従来の対訳コーパスの対応付け、従来の講演と予稿の対応付け及び質問文と知識データの対応付けを行うものは、複雑なプログラムが必要であり、簡単に対応付けを行うのが困難なものであった。

【0004】

本発明は上記問題点の解決を図り、diffコマンドのような順序情報を保持したまま一致部分が最大になるように複数のデータの対応づけを行うシステムを用い、対訳コーパスの対応付け、講演と予稿の対応付け、質問文と知識データ等の複数の言語データの対応付けを簡単に行えるようにすることを目的とする。

20

【課題を解決するための手段】

【0005】

図1は本発明の言語処理システムである。図1中、1は入力手段、2は処理部、3は形態素解析手段、4は差分検出手段、5は類似度演算手段である。

【0006】

本発明は、前記従来の課題を解決するため次のような手段を有する。

【0007】

(1)：質問文を入力する入力手段1と、前記入力された質問文を平叙文に変換する質問文変換部と、順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムを用いて入力された前記平叙文に変換された質問文と知識データとの共通部分と差分部分を検出する差分検出手段4と、前記検出した共通部分と差分部分により、前記平叙文に変換された質問文と知識データとの類似度を求める類似度演算手段5とを備え、前記類似度演算手段5は、前記類似度の最も大きい知識データの前記平叙文に変換された質問文の疑問詞と対をなして差分部分として得られた部分を出力する。このため、質問文に対する解を容易に出力することができる。

30

【0008】

(2)：前記(1)の言語処理システムにおいて、前記差分検出手段4で検出した差分の共通部分と不一致部分の数により、前記二つの言語データの類似度を求める前記類似度演算手段を備える。このため、知識データ等の多数の言語データから類似度の大きいものを容易に選択することができる。

40

【0009】

(3)：前記(1)又は(2)の言語処理システムにおいて、前記順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムとしてdiffコマンドを用いる。このため、質問文と知識データとの対応付けを簡単に行うことができる。

【0010】

(4)：質問文を入力する処理と、順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムを用いて入力された前記質問文と知識データとの共通部分と差分部分を検出する処理と、前記検出した共通部分と差分部分により、前記質問文と知識データとの類似度を求める処理と、類似度の大きい知識データの前記質問文

50

の疑問詞に対応する部分を入力する処理とを、コンピュータに実現させるためのプログラムとする。このため、このプログラムをコンピュータにインストールすることで、質問文に対する解を容易に出力することができる言語処理システムを容易に提供することができる。

【発明の効果】

【0011】

本発明によれば次のような効果がある。

【0012】

(1)：類似度演算手段で、類似度の最も大きい知識データの質問文の疑問詞と対をなして差分部分として得られた部分を入力するため、質問文に対する解を容易に出力することができる。

10

【0013】

(2)：差分検出手段で検出した差分の共通部分と不一致部分の数により、二つの言語データの類似度を求める類似度演算手段を備えるため、知識データ等の多数の言語データから類似度の大きいものを容易に選択することができる。

【0014】

(3)：順序情報を保持したまま一致部分を最大になるように複数のデータの対応づけを行うシステムとしてdiffコマンドを用いるため、質問文と知識データとの対応付けを簡単に行うことができる。

【発明を実施するための最良の形態】

20

【0015】

(1)：言語処理システムの説明

図1は本発明の言語処理システムの説明図である。図1において、入力手段1は、差分検出を行う言語データを入力するものである。処理部2は、入力されたデータの処理を行うものである。形態素解析手段3は、入力された言語データを辞書と文法を用いて最適な単語列に分割するものである。差分検出手段4は、対応関係のある複数の言語データをdiffコマンドを用いて差分の検出を行うものである。類似度演算手段5は、差分検出手段4を用いて質問文と知識データとの類似度を求め、類似度の大きい知識データの質問文の疑問詞に対応する部分を入力するものである。

【0016】

30

(2)：diff、mdiff、mdiffcの説明

(1) diffの説明

diff(ディフ)とは、UNIX(ユニックス)(登録商標)のファイル比較ツールdiffのことである。このdiffコマンドは、与えられた二つのファイルの差分を順序情報を保持したまま行を単位として出力するものである。

【0017】

例えば、

今日
学校へ
いく

ということが書いてあるファイル(1)と

今日
大学へ
いく

ということが書いてあるファイル(2)があるとする。これらのdiffをとると、差分が

<学校へ
>大学へ

40

のような形で出力される。

【0018】

(2) mdiffの説明

50

ところで、diffコマンドには、-D オプションという便利なオプションがある。このオプションを付けてdiffコマンドを使うと差分部分だけでなく共通部分も出力される。つまり、ファイルのマージが実現される。また、差分部分は、C (プログラム言語) のプリプロセッサなどで使われるifdef 文などで表現されるが、ここでは、ifdef 文は、見にくいので差分部分は以下のように表示することにする。

【0019】

```

;
(一つ目のファイルにだけある部分)
;
(二つ目のファイルにだけある部分)
;

```

10

ここでは、“ ; ” は差分部分の始まりを、“ ; ” は差分部分の終わりを意味し、“ ; ” は差分を構成する二つのデータの境界を意味する。

【0020】

本実施の形態では、-D オプションを付けて更にifdef の部分を上記のように表示(整形)して、ファイルのマージを行う場合のdiffをmdiff (エムディフ)と呼ぶ(mはmergeのmである)。

【0021】

実際に、先ほどのデータ(ファイル(1)とファイル(2))に対して、mdiff をかけると、以下のような結果になる。

20

【0022】

```

今日
;
学校へ
;
大学へ
;
いく

```

これは「今日」が一致し、「学校へ」と「大学へ」が差分となり、「いく」がまた共通部分となっている。このように、mdiff の出力はdiffと異なり一致部分も出力されるために分かりやすい。

30

【0023】

また、mdiff の結果からは元の二つのファイルのデータを完全に復元することができる。共通部分と、差分部分の黒丸(;)の上側だけを取り出すと、

```

今日
学校へ
いく

```

のように一つ目のファイルの情報が取り出される。また、共通部分と、差分部分の黒丸(;)の下側だけを取り出すと、

```

今日
大学へ
いく

```

40

のように二つ目のファイルの情報が取り出される。このように、もとの情報を完全に復元できることになる。

【0024】

また、mdiff では、一致部分は片方のデータにあったものだけを表示し、不一致部分のみ両方のデータのものを表示するために、元の二つのデータよりもデータ量は削減できるが、上記のように元の情報を完全に復元できるために、復元できる状態でデータ量を削減するという意味ではmdiff はデータ圧縮を実現しているものといえる。

【0025】

50

図2はmdiffの言語処理システムの説明図である。図2において、mdiffの言語処理システムには、UNIXdiff処理部11、整形部12が設けてある。UNIXdiff処理部11は、入力された二つのファイルのdiffによる差分部分と共通部分を出力するものである。整形部12は、UNIXdiff処理部11の出力を見やすい表現に整形するものである。

【0026】

図3はmdiffによるフローチャートである。以下、図3の処理S1～S4に従って説明する。

【0027】

S1：UNIXdiff処理部11に、入力として二つのファイルが与えられる。

【0028】

S2：UNIXdiff処理部11で、二つのファイルの一致・不一致部分を検出する(ここのdiffはオプション-Dを付けておき一致部分の検出も行う)。

【0029】

S3：整形部12において、UNIXdiff処理部11の出力を整形する。具体的には“diff-D記号”の出力のifdef文を“ ; ”などの見やすい表現に直す処理を行う。

【0030】

S4：整形部12は、整形された結果を出力する。

【0031】

(3) mdiffcの説明

次に、文字を単位としたmdiffを考える。言語処理の場合は、文字単位を差分で取りたい場合が多い。そのようなときは一度ファイルの中身の情報を、一文字ずつ改行をして出力したファイルでmdiffをとればよい。例えば先のファイル(1)の情報だと、

今

【0032】

日
学
校
へ
い
く

という形にしてから、mdiffをとればよい。この一文字単位でmdiffをかけることをmdiffcと呼ぶ(mdiffcのcはcharacterのc)。

【0033】

diffの表示は見にくく、mdiffはdiffで表示される情報を完全に含むので以降の説明は、mdiffを用いて行う。

【0034】

(3)：差分検出及び書き換え規則の獲得の説明

(1) 複数システムの出力の差分検出の説明

以前、jumanのシステムのバージョンが複数乱立しているとき、この複数のjumanの出力をmdiffによりマージして形態素解析結果の品質を向上させるようなことをしていた(参考文献、村田真樹、日本語文章における名詞の指示対象の推定、京都大学工学部博士論文、(1995)、石間衛、藤井敦、石川徹也、日本語形態素・構文解析システムJEMONIの開発と評価について、情報処理学会自然言語処理研究会 98-NL-127、(1998)、参照)。ここでは「といったこと」の例で説明する。

【0035】

「といったこと」を解析し、jumanのAというバージョンの出力が

と	と	助詞
いった	言う	動詞
こと	こと	名詞

10

20

30

40

50

となっていて、Bのバージョンの出力が

と	と	助詞
いった	行く	動詞
こと	こと	名詞

となっているとする。「いった」という語は「行く」と「言う」の曖昧性があり、Bのバージョンではこれを誤って「行く」の方の語であると出力していたとする。ここでmdiffをとると以下のような結果となる。

【0036】

と	と	助詞
;		
いった	言う	動詞
;		
いった	行く	動詞
;		
こと	こと	名詞

10

mdiffをとることで複数のシステムの出力の差異を容易に検出することができる。この場合、「いった」の部分が出力に差異があることが分かる。ここで、出力修正の作業者は、このような差分が検出された箇所においてどちらが正しいかを判断し、上が正しいければ何もせず下が正しいければ「 ; 」の先頭に“ x ”を付けるなどと決めておく。そのようにすると、“ x ”がなければ差分の下を、あれば差分の上の情報と区切り、記号を消すことで、その作業結果のデータから自動的にそれぞれの差分からよい結果の方を選び、それぞれのバージョンのものより高い精度の結果を生成できる。また、差分の両方が誤っている場合がよくある。このときは「 ; 」の上の方のデータを実際書き直す

20

【0037】

この方法を用いると、修正できないものは両方のバージョンで同じように誤るものだけであり、多くの形態素誤りを修正できる。ここで注意すべきことは異なる性質のシステムを複数用意しないといけないということである。誤り方が同じシステムの場合だと多くの誤りを見逃すことになる。

【0038】

また、システムが三つある場合は、diff3 コマンドを使うとよい。diff3 は、三つのファイルの差分を検出することができる。

30

【0039】

上記では、形態素解析を例にあげたが、他の解析でも解析結果を行単位にすることでmdiffで差分をとることができる。また、文字単位が必要ならばmdiffcを使えばよい。

【0040】

ここでは、複数のシステムの出力の差分をとる話をしたが、一つをタグ付きコーパスとし、それを何かのシステムで解析した結果と比較することで、そのタグ付きコーパスの誤りを検出し修正することもできる。

【0041】

(2) 差分の考察と書き換え規則の獲得の説明

ここでは、話し言葉と書き言葉のdiffの研究について記述する。対応のとれた話し言葉と書き言葉のデータを使い、それらの差分から話し言葉と書き言葉の違いを考察したり、話し言葉から書き言葉への言い換え規則、また、その逆のための規則を獲得するものである。データとしては、学会の口頭発表を話し言葉データとし、その口頭発表の内容が記されたその学会の予稿原稿を書き言葉として用いた。

40

【0042】

図4は書き言葉データと話し言葉データの例の説明図である。例えば、書き言葉と話し言葉のデータが図4のような形で与えられたとする。ここでは、差分がとりやすいように形態素解析システム(形態素解析手段3)などで1行に1単語がはいるような形に変換し

50

である。このような書き言葉と話し言葉のデータが与えられたとき、mdiff をとると、図 5 のような結果を得る。図 5 は書き言葉データと話し言葉データのdiffの結果の説明図である。図 5 の結果から、差分部分だけを抽出すると、図 6 のような結果が得られる。図 6 は差分部分の抽出の説明図である。

【 0 0 4 3 】

図 6 の結果から、話し言葉には「え」などが挿入されること、また話し言葉では「っていうの」という表現をいれて発話をなめらかにすることなどが分かる。また、「述べる」が「述べます」と言い換えられていることが分かる。以上のように、mdiff を使うことで話し言葉と書き言葉の差異を検出でき、また、それを考察することで、話し言葉と書き言葉の違いのようなものを調査できることが分かる。また、これらの差分は、話し言葉と書き言葉の言い換え規則としてみることもできる。

10

【 0 0 4 4 】

例えば、「え」の部分は、書き言葉に何も無いところに話し言葉に変換する場合「え」をいれるという規則のように見ることが出来る。また、「述べる」と「述べます」の部分は、話し言葉に変換する場合は「述べる」を「述べます」に言い換える規則のように見ることが出来る。その意味でmdiff を用いることで言い換え規則、もしくは、変換規則のようなものを検出できることが分かる。

【 0 0 4 5 】

ここでは、話し言葉と書き言葉のデータを例にとったが、このようなことは様々なところで可能である。例えば、英文校閲前のテキストと英文校閲後のテキストで、mdiff をとると、どのような違いをどのように直せばよいか分かるし、また英文校閲用の規則のようなものが獲得できる。また、要約前のテキストと要約後のテキストで、mdiff をとると、どのように要約されているかを如実に見ることが出来るし、要約用の規則のようなものが獲得できる。その他にも対応のとれた性質の異なるデータに対してmdiff をとると、様々な考察と、言い換え規則の獲得ができる。

20

【 0 0 4 6 】

(4) : データのマージの説明

(1) 対訳コーパスの対応付けの説明

ここでは、対訳コーパスの対応付けを考える。条件として、それぞれのコーパスには、対応する箇所と同じ記号が入っていることを前提とする。また、対応付けの単位は、この記号で区切られた部分であるとする。

30

【 0 0 4 7 】

図 7 はコーパスの構成の説明図である。ここでは、日本語のコーパスと英語のコーパスがまだ、ばらばらに存在し、対応付けられていないとする。図 7 の例のように両方とも Section 1 などの同じ形をしたセクション情報が与えられているとする。このとき、日本語と英語では、同じセクションのものは、同じ内容であるとする。

【 0 0 4 8 】

この場合、これらのデータのmdiff をとることで、図 8 のような結果を得ることが出来る。図 8 はmdiff によって対応付けられた対訳コーパスの説明図である。図 8 の結果では、Section 1 などが共通部分となり、その他の部分が不一致部分となる。この不一致部分では、日本語と英語が上下に分かれて格納されることになる。このようにすることで、mdiff を用いて対訳データが作成されることになる。

40

【 0 0 4 9 】

ここで示したものは、文ごとなどの細かい対応付けをするものではなく、セクションなどの大雑把なもので一見役に立たないように思われるかもしれないが、文の対応付けは難しい問題で、まず予め対応がとれていることがはっきりしている章、段落のレベルで対応付けをしてから細かい対応付けをするという考え方もあり、その意味ではこのような粗い対応付けも役立つものである。

【 0 0 5 0 】

また、ここで示したものは、Section 1 などの情報を認識させて区分するだけでそのよ

50

うなことをするプログラムを書くことでも同じように対訳データの対応付けを行うことができる。しかし、mdiff を使うとそのような複雑なプログラムを書くこともなく対応付けを容易に実現できるものである。

【 0 0 5 1 】

図 9 は対訳コーパスの言語処理システムの説明図である。図 9 において、対訳コーパスの言語処理システムには、mdiff 処理部 2 1 が設けてある。mdiff 処理部 2 1 は、入力された原文データと翻訳データとの二つのファイルのmdiff を出力するものである。

【 0 0 5 2 】

図 1 0 は対訳コーパスのmdiff によるフローチャートである。以下、図 1 0 の処理 S 1 1 ~ S 1 3 に従って説明する。

【 0 0 5 3 】

S 1 1 : mdiff 処理部 2 1 に、入力として二つのファイルが与えられる。ここではこの二つのファイルは、それぞれ英語文章、日本語文章を格納したものである。

【 0 0 5 4 】

S 1 2 : mdiff 処理部 2 1 で、この二つのファイルのmdiff をとる。

【 0 0 5 5 】

S 1 3 : mdiff 処理部 2 1 は、mdiff の結果を出力する。

【 0 0 5 6 】

(2) 講演と予稿の対応付けの説明

講演と予稿の対応付けを考える。この講演と予稿は、先の書き換え規則の獲得でも述べた書き言葉データと話し言葉データに対応する。即ち、講演は学会の口頭発表で、予稿はその口頭発表に対応する論文のことである。このような講演と予稿が与えられたとき、講演の各部分と、予稿の各部分の対応がとれると、講演を聞いている時だと、それに対応する予稿の部分を参照できるし、予稿を読んでいるときだと、それに対応する講演の部分を参照できて便利である。ここでは、この講演と予稿の対応付けをmdiff で行う説明をする。

【 0 0 5 7 】

ここでは、特に予稿の各章が講演のどこの部分に対応するかをmdiff でもとめることにする。ここで予稿と講演とは、同じ順序でなされると仮定する。また、予稿の章が認識しやすいように予稿データには、図 1 1 のように “ <Chapter 1> ” のような記号を挿入しておく。図 1 1 は予稿データの構成の説明図である。この形にしてにおいて、予稿と講演のデータに対して、形態素解析をして各行に単語がくる状態でmdiff を使うことで、もしくは、mdiffcを使うことで、図 1 2 のような結果を得る。図 1 2 は予稿と講演のmdiff の結果の説明図である。ここで、差分部分で予稿に対応する上半分の方を、“ <Chapter 1> ” のような記号を除いてすべて消し去ると図 1 3 のような結果を得る。図 1 3 は講演データへの章の情報の挿入結果の説明図である。図 1 3 では、元の講演データに対して “ <Chapter 1> ” のような記号だけが挿入された形になる。つまり、講演のどの部分が予稿のどの章にあたるかが分かることになる。

【 0 0 5 8 】

これは簡単にいうと、mdiff の照合能力を用いて予稿と講演を照合し、章の情報だけ残して予稿の情報を消し去ることにより、講演データに章の情報を挿入するということを行っていることを意味する。このような予稿と講演の対応付けもmdiff を用いると簡単にすることができる。

【 0 0 5 9 】

図 1 4 は講演と予稿の対応付けの言語処理システムの説明図である。図 1 4 において、mdiff の言語処理システムには、mdiff 処理部 2 1、予稿削除部 2 2 が設けてある。mdiff 処理部 2 1 は、入力された二つのファイルのmdiff をとり出力するものである。予稿削除部 2 2 は、予稿側の差分部分で<Chapter 1> などの章情報のみを残して予稿データをすべて削除するものである (mdiff 記号の “ ; ” など削除する)。

【 0 0 6 0 】

10

20

30

40

50

図 1 5 は講演と予稿の対応付けの処理フローチャートである。以下、図 1 5 の処理 S 2 1 ~ S 2 4 に従って説明する。

【 0 0 6 1 】

S 2 1 : mdiff 処理部 2 1 に、入力として二つのファイルが与えられる。この二つのファイルは、それぞれ予稿、講演の文章を格納したものである。また、予稿データの方は、<Chapter 1> などの章の範囲を示す記号が付されているものとする。

【 0 0 6 2 】

S 2 2 : mdiff 処理部 2 1 で、この二つのファイルの mdiff をとる。

【 0 0 6 3 】

S 2 3 : 予稿削除部 2 2 は、予稿側の差分部分で<Chapter 1> などの章情報のみを残して予稿データをすべて削除するものである。また、mdiff 記号の “ ; ” なども削除する。 10

【 0 0 6 4 】

S 2 4 : 予稿削除部 2 2 は、予稿側の差分部分で章情報のみを残して予稿データをすべて削除した結果を出力する。

【 0 0 6 5 】

(5) : 最適照合能力を用いた質問応答システムの説明

ここでは mdiff の最適照合能力を用いた質問応答システム (質問応答言語処理システム) について記述する。質問応答システムとは、例えば、「日本の首都はどこですか」と聞くと「東京」と答えそのものをずばり返すシステムである。 20

【 0 0 6 6 】

知識が自然言語で書かれていると仮定すると、基本的には質問文と知識の文を照合し、その照合結果で疑問詞に対応するところを答えとして出力すればよい。例えば先の問題だと、「日本の首都は東京です」という文を探してきて、この文で疑問詞に対応する「東京」を解として出力するのである。ここではこれを mdiff で行なうことを考える。

【 0 0 6 7 】

まず、質問文の疑問詞の部分を X に置き換え、また文末を平叙文に変換し、「日本の首都は X です」を得る。また、知識ベースから「日本の首都は東京です」を得る。ここでこの二つの mdiffc をとると以下のような結果を得る。

【 0 0 6 8 】

日
本
の
首
都
は
;
X
;
東
京
;
で
す

30

40

ここで X と差分部分で組になっているものを解とすると、「東京」を正しく取り出せることになる。

【 0 0 6 9 】

ところで mdiffc を使う場合少々文に食い違いがあっても、答えを正しく取り出すことができる。例えば、知識ベースの文が「日本国の首都は東京です」であったとする。この場合は mdiffc の結果は以下ようになる。

50

【 0 0 7 0 】

日
本
；
；
国
；
の
首
都
は
；
X
；
東
京
；
で
す

10

差分部分は少し増えるが X に対応する箇所は「東京」のままで、解を正しく抽出できる

20

【 0 0 7 1 】

ところで、われわれが提案する質問応答システムでは、類似度を尺度として用いた変形をくりかえし、質問文と知識データの文がより一致した状態で上記のような照合を行なう。このために類似度を定義する必要がある。

【 0 0 7 2 】

mdiff を用いた場合は、一致部分と不一致部分が認定できるので、類似度は、（一致部分の文字数） / （全文字数）のような形で定義できる。ここではmdiff により類似度を求めるようなことをしている。このように mdiff は文の類似性 / 類似度を求めることにも役に立つ。

30

【 0 0 7 3 】

ここで、「日本国」と「日本」を言い換える規則があれば「日本の首都は X です」を「日本国の首都は X です」と言い換えて照合し、不一致部分を減らすことで、より確実に解を得ることができる。

【 0 0 7 4 】

このように、mdiff を使うだけで簡単な質問に回答する言語処理システムを、容易に構築できることは簡便さの観点から価値がある。

【 0 0 7 5 】

図 1 6 は質問応答システムの説明図である。図 1 6 において、質問応答システムには、質問文変換部 3 1、質問文保存部 3 2、キーワード抽出部 3 3、データベース文検索部 3 4、データベース文保存部 3 5、類似度演算部 3 6、mdiff 処理部 3 7、質問文変形部 3 8、データベース文変形部 3 9、対応部出力部 4 0 が設けてある。質問文変換部 3 1 は、疑問文を平叙文にまた疑問詞を X に変換するものである。質問文保存部 3 2 は、質問文を保存するものである。キーワード抽出部 3 3 は、質問文からキーワードを抽出するものである。データベース文検索部 3 4 は、キーワードを多く含むデータベース文を検索するものである。データベース文保存部 3 5 は、データベース文検索部 3 4 で検索したデータベース文を保存するものである。類似度演算部 3 6 は、質問文とデータベース文とを比較し類似度を求めるものである。mdiff 処理部 3 7 は、質問文とデータベース文とを 1 文字単位で mdiff をとるものである。質問文変形部 3 8 は、質問文の変形を行うものである。データベース文変形部 3 9 は、データベース文の変形を行うものである。対応部出力部 4 0 は

40

50

、mdiff 処理部 37 の出力結果のうち “ X ” に対応するデータベース側の表現を抽出し出力するものである。

【 0 0 7 6 】

図 17 は質問応答の処理フローチャートである。以下、図 17 の処理 S 3 1 ~ S 3 9 に従って説明する。

【 0 0 7 7 】

S 3 1 : 質問応答システムの類似度 S 0 の値を適当に与え (通常最初は “ 0 ” とする) 処理 S 3 2 に移る。

【 0 0 7 8 】

S 3 2 : 質問応答システムの入力として、質問文が与えられ処理 S 3 3 に移る。 10

【 0 0 7 9 】

S 3 3 : 質問文変換部 3 1 は、与えられた質問文の疑問文を平叙文に、また疑問詞を X に変換し、この変換結果の文を質問文保存部 3 2 に渡す。

【 0 0 8 0 】

S 3 4 : キーワード抽出部 3 3 にも、入力として与えられた質問文が渡され、質問文からキーワードが抽出される。この抽出されたキーワードはデータベース文検索部 3 4 に渡され処理 S 3 5 に移る。

【 0 0 8 1 】

S 3 5 : データベース文検索部 3 4 で、キーワードを多く含むデータベース文が、知識ベース (図示しないデータベース) からいくつか検索され、その結果はデータベース文保存部 3 5 に格納され処理 S 3 6 に移る。 20

【 0 0 8 2 】

S 3 6 : 類似度演算部 3 6 で、質問文とデータベース文とを比較し類似度を求める。この類似度の算出にはmdiff 処理部 3 7 を用いる。即ち、類似度を求めたい質問文とデータベース文をmdiff 処理部 3 7 に入力し、このmdiff 結果により一致部分の文字数、不一致部分の文字数が求まる。類似度は、(一致部分の文字数) / (全文字数) と予めきめておく。そして、この類似度を質問文保存部 3 2 とデータベース文保存部 3 5 にあるすべての文の組に対して求め、このとき最も類似度が高かったときの類似度を S とする。次に、類似度演算部 3 6 は、類似度 S が S 0 より大きいかどうか判断する。この判断で、類似度 S が S 0 より大きい場合は処理 S 3 7 に移り、類似度 S が S 0 と同じとき処理 S 3 8 に移る 30

。

【 0 0 8 3 】

S 3 7 : 質問文変形部 3 8 とデータベース文変形部 3 9 では、それぞれ質問文とデータベース文の変形を行い、その変形結果をそれぞれの保存部に格納する。類似度演算部 3 6 は S 0 に類似度 S の値をセットし、処理 S 3 6 に戻る。なお、ここでの変形とは、「日本の首都は X です」を「日本の首都は X です」にいいかえるようなことを意味している。

【 0 0 8 4 】

S 3 8 : 類似度演算部 3 6 は、このときの類似度 S の値を求めるときに使ったmdiff 処理部 3 7 の出力結果を対応部出力部 4 0 に渡し処理 S 3 9 に移る。

【 0 0 8 5 】

S 3 9 : 対応部出力部 4 0 では、mdiff 処理部 3 7 の出力結果のうち “ X ” に対応するデータベース側の表現を抽出して出力する。 40

【 0 0 8 6 】

なお、上記実施の形態では、差分検出を行うのにdiffコマンドを用いたが、予め定めた単位で、順序情報を保持したまま一致部分を最大にする対応付けを行うシステムであれば他の差分検出手段を用いることができる。

【 0 0 8 7 】

(6) : プログラムインストールの説明

入力手段 1、処理部 2、形態素解析手段 3、差分検出手段 4、類似度演算手段 5、UNIXdiff処理部 11、整形部 12、mdiff 処理部 21、予稿削除部 22、質問文変換部 3 50

1、質問文保存部32、キーワード抽出部33、データベース文検索部34、データベース文保存部35、類似度演算部36、mdiff処理部37、質問文変形部38、データベース文変形部39、対応部出力部40等は、プログラムで構成でき、主制御部(CPU)が実行するものであり、主記憶に格納されているものである。このプログラムは、一般的な、コンピュータで処理されるものである。このコンピュータは、主制御部、主記憶、ファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。

【0088】

このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピィ、光磁気ディスク等の可搬型の記録(記憶)媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記録媒体に対して、アクセスするためのドライブ装置を介して、或いは、LAN等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。

【図面の簡単な説明】

【0089】

【図1】本発明の言語処理システムの説明図である。

【図2】実施の形態におけるmdiffの言語処理システムの説明図である。

【図3】実施の形態におけるmdiffによるフローチャートである。

【図4】実施の形態における書き言葉データと話し言葉データの例の説明図である。

【図5】実施の形態における書き言葉データと話し言葉データのdiffの結果の説明図である。

【図6】実施の形態における差分部分の抽出の説明図である。

【図7】実施の形態におけるコーパスの構成の説明図である。

【図8】実施の形態におけるmdiffによって対応付けられた対訳コーパスの説明図である。

【図9】実施の形態における対訳コーパスの言語処理システムの説明図である。

【図10】実施の形態における対訳コーパスのmdiffによるフローチャートである。

【図11】実施の形態における予稿データの構成の説明図である。

【図12】実施の形態における予稿と講演のmdiffの結果の説明図である。

【図13】実施の形態における講演データへの章の情報の挿入結果の説明図である。

【図14】実施の形態における講演と予稿の対応付けの言語処理システムの説明図である。

【図15】実施の形態における講演と予稿の対応付けの処理フローチャートである。

【図16】実施の形態における質問応答システムの説明図である。

【図17】実施の形態における質問応答の処理フローチャートである。

【符号の説明】

【0090】

- 1 入力手段
- 2 処理部
- 3 形態素解析手段
- 4 差分検出手段
- 5 類似度演算手段

10

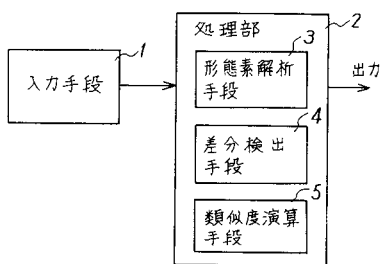
20

30

40

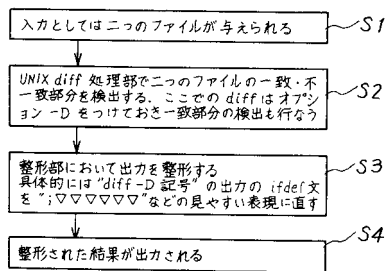
【 図 1 】

言語処理システムの説明図



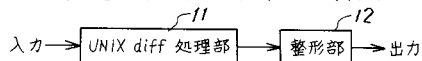
【 図 3 】

mdiffによるフローチャート



【 図 2 】

mdiffの言語処理システムの説明図



【 図 4 】

書き言葉データと話し言葉データの例の説明図

書き言葉データ	話し言葉データ
本	今日
論文	は
で	え
は	意味
意味	ソート
ソート	に
に	ついて
ついて	述べ
述べる	ます
。	一般に
一般に	ソート
ソート	って
は	いう
50	の
音	は
順	だいたい

【 図 5 】

書き言葉データと話し言葉データのdiffの結果の説明図

;▽▽▽▽▽	ついて
本	;▽▽▽▽▽
論文	述べる
で	。
;●●●●●	;●●●●●
今日	述べ
;△△△△△	ます
は	;△△△△△
;▽▽▽▽▽	一般に
;●●●●●	ソート
え	;▽▽▽▽▽
;△△△△△	;●●●●●
意味	って
ソート	いう
に	の
(右欄につづく)	;△△△△△

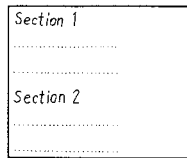
【 図 6 】

差分部分の抽出の説明図

書き言葉データ	話し言葉データ
本論文で	今日
述べる。	え
	述べます
	っていうの

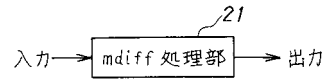
【 図 7 】

コーパスの構成の説明図



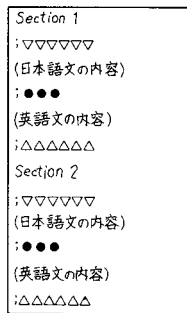
【 図 9 】

対訳コーパスの言語処理システムの説明図



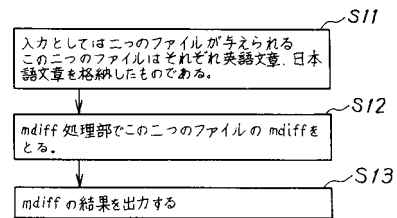
【 図 8 】

mdiffによって対応付けられた対訳コーパスの説明図



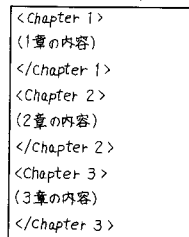
【 図 10 】

対訳コーパスのmdiffによるフローチャート



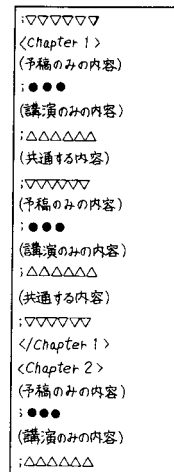
【 図 11 】

予稿データの構成の説明図



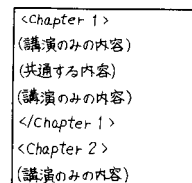
【 図 12 】

予稿と講演のmdiffの結果の説明図



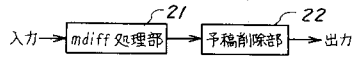
【 図 13 】

講演データへの章の情報の挿入結果の説明図



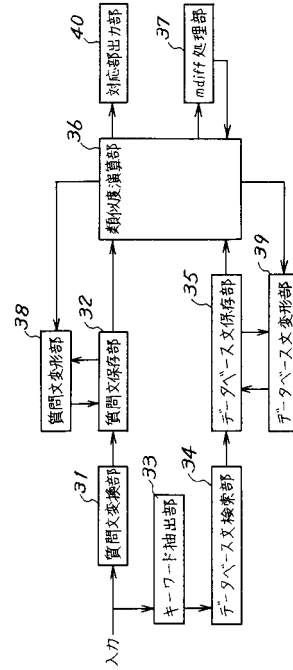
【図14】

講演と予稿の対応付けの言語処理システムの説明図



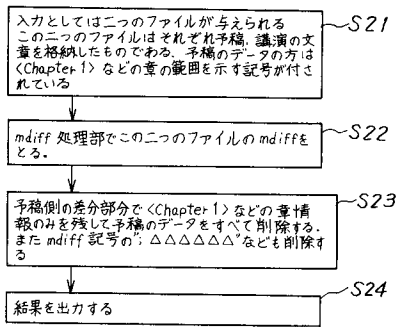
【図16】

質問応答システムの説明図



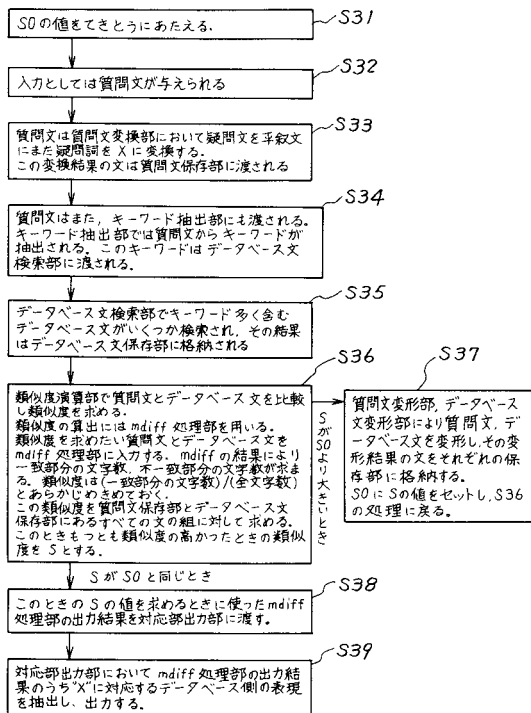
【図15】

講演と予稿の対応付けの処理フローチャート



【図17】

質問応答の処理フローチャート



フロントページの続き

(56)参考文献 特開平07-244669(JP,A)

特開2001-22763(JP,A)

村田真樹・井佐原均, diffと言語処理, 情報処理学会研究報告2001-NL-144, 日本, 社団法人情報処理学会, 2001年 7月17日, Vol.2001, No.69, p.127 - p.134, 電子情報通信学会(言語理解とコミュニケーション研究会)と共催

(58)調査した分野(Int.Cl., DB名)

G06F 17/21 - 17/30