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

特許第4085156号
(P4085156)

(45)発行日　平成20年5月14日(2008.5.14)　(24)登録日　平成20年2月29日(2008.2.29)

(51)Int.Cl.　F I
G06F 17/28 (2006.01)　G06F 17/28　R
G06F 17/30 (2006.01)　G06F 17/30　170A
G06F 17/30　320D

請求項の数 2　(全 10 頁)

(21)出願番号　特願2002-74270(P2002-74270)
(22)出願日　平成14年3月18日(2002.3.18)
(65)公開番号　特開2003-271592(P2003-271592A)
(43)公開日　平成15年9月26日(2003.9.26)
審査請求日　平成14年3月18日(2002.3.18)
審判番号　不服2005-8361(P2005-8361/J1)
審判請求日　平成17年5月6日(2005.5.6)

(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　斉
(72)発明者　内元　清貴
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　関根　聡
アメリカ合衆国、ニューヨーク州１０００３、ニューヨーク、セブンスフロアー、ブロードウェイ、７１５、ニューヨークユニバーシティ、コンピュータサイエンスデパートメント

最終頁に続く

(54)【発明の名称】テキスト生成方法及びテキスト生成装置

(57)【特許請求の範囲】
【請求項１】
所定の言語の文又は文章のテキストを生成するテキスト生成装置であって、
複数のキーワードとなる単語を入力する入力手段と、
キーワードから文節や句の候補（以下、文節等候補と呼ぶ。）を生成する文節等候補生成手段と、
係り受けの方向についての修飾条件、係り受け関係が交差するか否かについての交差条件、及び、係り受け要素に対する受け要素の個数についての対応条件からなる３つの係り受け条件に従った係り受け関係を仮定してテキスト候補を生成するテキスト候補生成手段と、
該テキスト候補を評価付けする評価手段と、
評価付けされた少なくとも１つのテキスト候補を出力する出力手段と共に、
入力手段で入力されたキーワードを含む文・語句を、データベースから抽出する抽出手段と、
抽出された文・語句を形態素解析及び／又は構文解析を行い、該キーワードを含む主辞形態素及び、それに連続する任意の数の形態素とから成る形態素集合を抽出し、該キーワードと該形態素集合との対応を文節等候補の生成規則として自動獲得する生成規則獲得手段と
を少なくとも備え、
文節等候補生成手段においては、少なくとも１つの該キーワードの前又は後に、該キー

ワードに関連した文字列を付加して文節等候補を生成し、他の全ての該キーワードについても同様に文字列を付加し、或いは付加せずに文節等候補を生成する処理か、又はキーワードから該生成規則を用いて文節等候補を生成する処理かのいずれかの処理を行い、

テキスト候補生成手段においては、該文節等候補から係り要素と受け要素の組合せを生成して該組合せ数に相当するテキスト候補を生成する

ことを特徴とするテキスト生成装置。

【請求項２】

前記入力手段において、

入力された単語と係り受け関係を有する単語を当該言語のデータベースから抽出し、その単語を新たなキーワードとして入力する

請求項１に記載のテキスト生成装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は自然言語処理の方法及び装置に関する。特に、いくつかのキーワードからテキストを生成する手法に関わる。

【０００２】

【従来の技術】

テキスト生成は機械翻訳、要約、対話システムなど自然言語処理の様々な応用に利用される重要な要素技術の一つである。近年、大量のコーパスが利用可能となり、自然な表層文を生成する目的にもコーパスが利用されるようになってきた。その典型例の一つが原言語から目的言語への機械翻訳に用いられる言語モデルである。

【０００３】

例えば、本件出願人らが特願２００１－３９５６１８号で開示したテキスト生成のシステムでは、置き換えた単語や句を目的言語側で尤もらしい順序に並び替え、目的言語を生成する。言語モデルの入力は、一般に語の集合であり、言語モデルに要求されるのは、基本的にそれらの語の並べ換えである。

このような従来のシステムでは、与えられた語の集合を並べ換えると自然な文を生成できるという仮定がある。つまり、自然な文を生成するための語の集合は翻訳モデルにより過不足なく生成されることが前提となっている。

【０００４】

しかし、この前提のためには大規模な対訳コーパスが必要であり、日本語などの比較的コーパスが整備された言語が原言語であっても、対象言語との対訳コーパスの状況、対象言語におけるコーパスの状況によっては、上記従来の手法では十分なテキスト生成が行えない場合があった。

また、上記開示でもある程度の語句の補完は行うが、補助的な補完を行うのみで、効率的に関連する語句を補完することはできなかった。

【０００５】

この問題は機械翻訳に限らず、一般的にテキスト生成において生じる問題であり、原言語テキストが完全なものでなく、誤りのあるOCR 認識結果や音声認識結果などの場合には同様に高精度なテキスト生成ができない問題があった。

【０００６】

【発明が解決しようとする課題】

本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、入力するキーワードが十分でない場合にも、有意なテキストを生成するテキストの生成方法・生成装置を提供することである。

【０００７】

【課題を解決するための手段】

本発明は、上記の課題を解決するために、次のようなテキスト生成装置を創出する。

すなわち、所定の言語の文又は文章のテキストを生成するテキスト生成装置であって、

複数のキーワードとなる単語を入力する入力手段と、キーワードから文節や句の候補（本発明において文節等候補と呼ぶ）を生成する文節等候補生成手段と、係り受けの方向についての修飾条件、係り受け関係が交差するか否かについての交差条件、及び、係り受け要素に対する受け要素の個数についての対応条件からなる３つの係り受け条件に従った係り受け関係を仮定してテキスト候補を生成するテキスト候補生成手段と、該テキスト候補を評価付けする評価手段と、評価付けされた少なくとも１つのテキスト候補を出力する出力手段をまず備える。

【０００８】

さらに、本発明は、入力手段で入力されたキーワードを含む文・語句を、データベースから抽出する抽出手段と、抽出された文・語句を形態素解析及び／又は構文解析を行い、該キーワードを含む主辞形態素及び、それに連続する任意の数の形態素とから成る形態素集合を抽出し、該キーワードと該形態素集合との対応を文節等候補の生成規則として自動獲得する生成規則獲得手段とを備える構成である。

【０００９】

文節等候補生成手段においては、少なくとも１つの該キーワードの前又は後に、該キーワードに関連した文字列を付加して文節等候補を生成し、他の全ての該キーワードについても同様に文字列を付加し、或いは付加せずに文節等候補を生成する処理か、又はキーワードから該生成規則を用いて文節等候補を生成する処理かのいずれかの処理を行う。

また、テキスト候補生成手段においては、該文節等候補から係り要素と受け要素の組合せを生成して該組合せ数に相当するテキスト候補を生成することを特徴とするテキスト生成装置である。

【００１０】

本発明では、前記入力手段において、入力された単語と係り受け関係を有する単語を当該言語のデータベースから抽出し、その単語を新たなキーワードとして入力するようにすることもできる。

【００１１】

【発明の実施の形態】

以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。

図１には本発明におけるテキスト生成装置（以下、本装置）（１）の説明図を示す。最も単純な本装置（１）の機能として、例えば「彼女」「家」「行く」の３つのキーワード（２）が入力された時に、「彼女の家に行く」（３ａ）「彼女が家に行った」（３ｂ）などのテキストを生成する。

【００１２】

本装置（１）の具体的な構成例として図２に示す各部を備える。本装置（１）は例えば、ＣＰＵとメモリ、ハードディスクなどの外部記憶媒体を備えるパーソナルコンピュータなどにより構成することができ、主な処理をＣＰＵにおいて行い、処理の結果を随時ＲＡＭ、外部記憶媒体に記録する。

本発明において入力となるキーワード（２）は、パーソナルコンピュータに接続されたキーボードを用いて入力したり、他の言語処理システムから出力されたデータを用いることができる。

【００１３】

本実施例で、キーワード（２）は２つの処理に用いられる。その１つは文節生成規則獲得部（４）であり、もう１つは文節候補生成部（５）である。

ここでは、日本語を対象とし、文節を生成する。キーワードは文節の主辞となる語であると定義する。そして、文節の主辞となる語は、文末に一番近い内容語であるとする。ここで、内容語は、その語の品詞が、動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞、未定義語である形態素の見出し語であるとし、それ以外の形態素の見出し語を機能語とする。ただし、サ変動詞、動詞「なる」、形式名詞「の」については、文節内で他に内容語がない場合を除いて機能語として扱う。品詞の体系は京大コーパス(Version3.

0)（黒橋長尾1997）のものに従った。

【００１４】

文節生成規則獲得部（４）では、キーワード「彼女」、「家」、「行く」が与えられたとき、それぞれを含む文をコーパス（８）から検索し、形態素解析、構文解析(係り受け解析）をする。そして、そこからキーワード（２）を含む文節を抽出して、キーワードから文節を生成する規則「彼女」→「彼女の」、「彼女」→「彼女が」、「家」→「家に」、「行く」→「行く」、「行く」→「行った」などの文節生成規則（９）を獲得し、記録する。

【００１５】

ここで、生成規則の自動獲得には次の手法を用いる。キーワードの集合をVとし、キーワードk(∈V）から文節を生成する規則の集合をR_k とするとき、規則r_k(∈Rk）は次の形式で表現されるものと定義する。

k → h_k m*

ここで、h_k はキーワードを含む主辞形態素、m*は同じ文節内でh_kに連続する任意個の形態素とする。キーワードが与えられると、この形式を満たす規則を単言語コーパスから自動獲得する。

【００１６】

一方、文節候補生成部（５）では、文節生成規則（９）を参照しながら、入力されたキーワード（２）から出力するテキスト（３）を構成する文節の候補を生成する。

例えば、「彼女」では自然なテキストを構成する文節とはなりにくいが、「彼女の」あるいは「彼女が」のように「彼女」という単語と極めて密接な関連性を有する語句を付加し、後段の処理によるテキスト生成に備える。

【００１７】

本実施例のように、文節生成規則獲得部（４）によりコーパス（８）から入力するキーワード（２）の文節規則を生成することで、最小限の計算量で効果的に文節生成規則を得ることができ、処理速度の向上に寄与する。

【００１８】

しかしながら、本発明の実施においては必ずしもキーワード（２）に関連する語句をコーパスから抽出する構成を取る必然性はなく、計算能力に応じて任意の語句を入力されたキーワード（２）の前後に付加してもよい。

本発明では後述の評価部（７）により、任意の語句を付加しても当該文節候補について緻密な評価がされるため、これにより最も評価値の高くなる文節候補が生成できるようになる。

【００１９】

次に、テキスト候補生成部（６）でテキスト候補を生成する。テキスト候補はグラフあるいは木の形で表現する。

すなわち、図３に示すように、各文節候補（４ａないし４ｆ）の間に係り受けの関係を仮定して、テキスト候補１（１２）、テキスト候補２（１３）のような文節を単位とした依存構造木の形でテキスト候補を生成する。

【００２０】

このとき、次の条件を満たすように依存構造木の候補を生成する。

(i) 係り受けは前方から後方に向いている。(後方修飾)

(ii) 係り受け関係は交差しない。(非交差条件)

(iii) 係り要素は受け要素を一つだけ持つ。

例えば、キーワードが３個の場合、キーワードを含む文節候補がそれぞれb1、b2、b3であったとすると、順序を固定した場合には、(b1(b2 b3))、((b1 b2) b3) の２通り、固定しない場合には１６通りの候補ができる。

【００２１】

生成されたテキスト候補（１２・１３など）は、評価部（７）でコーパスから学習したキーワード生成モデル（１０）や言語モデル（１１）を用いて順序付けされる。

以下、キーワード生成モデル（１０）と、言語モデル（１１）として形態素モデル及び係り受けモデルについて説述する。

【００２２】

キーワード生成モデルでは、次の５種類の情報を素性として用いたモデル（ＫＭ１ないし５）を考える。以下で、キーワードの集合Vは、ある回数以上コーパスに出現した主辞単語の集合とし、文節は前記で表現されるものと仮定する。また、各キーワードは独立であり、与えられたテキストが単語列w1・・・wm からなるとき、キーワードki は単語wj(1≦j≦m) に対応していると仮定する。

【００２３】

【ＫＭ１】

前方の二単語を考慮(trigram)

ki は前方の二単語wj-1 とwj-2 のみに依存すると仮定する。

【式１】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{j-1},w_{j-2})$$

【００２４】

【ＫＭ２】

後方の二単語を考慮(後方trigram)

ki は後方の二単語wj+1 とwj+2 のみに依存すると仮定する。

【式２】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{j+1},w_{j+2})$$

【００２５】

【ＫＭ３】

係り文節を考慮(係り文節)

ki を含む文節に係る文節がある場合、ki はそのうち最も文末側の文節の末尾から二単語wl とwl-1 のみに依存すると仮定する(図４参照）。

【式３】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{l},w_{l-1})$$

【００２６】

【ＫＭ４】

受け文節を考慮(受け文節)

ki を含む文節を受ける文節がある場合、ki はその文節内の主辞単語から二単語ws とws+1 のみに依存すると仮定する(図４参照）。

【式４】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{s},w_{s+1})$$

【００２７】

【ＫＭ５】

係り文節を最大二文節考慮(係り二文節)

ki を含む文節に係る文節がある場合、ki は、そのうち最も文末側の文節の末尾から二単語wl 、wl-1 と、最も文頭側の文節の末尾から二単語wh 、wh-1 のみに依存すると仮定する(図４参照）。

【式５】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_l, w_{l-1}, w_h, w_{h-1})$$

【００２８】

次に、形態素モデル（ＭＭ）について示す。形態素に付与するべき文法的属性がl 個あると仮定する。テキストつまり文字列が与えられたとき、その文字列が形態素であり、かつj(1≦j≦l) 番目の文法的属性を持つとしたときの尤もらしさを確率値として求めるモデルを用いる。

テキストT が与えられたとき、順序付き形態素集合M が得られる確率は、各形態素mi(1≦i≦n) が独立であると仮定し、

【式６】

$$P(M|T) = \prod_{i=1}^{n} P(m_i|m_1^{i-1}, T)$$

と表す。ここで、mi は1 からl までのいずれかの文法的属性を表わす。

【００２９】

一方、係り受けモデル（ＤＭ）は、テキストTと順序付き形態素集合Mが与えられたとき、各文節に対する係り受けの順序付き集合Dが得られる確率は、各々の係り受けd1・・・dn が独立であると仮定し、

【式７】

$$P(D|M,T) = \prod_{i=1}^{n} P(d_i|M,T)$$

と表わす。

【００３０】

例えば、「彼女　公園　行った」の３つのキーワードから「(彼女は(公園へ行った))」と「((彼女の公園へ) 行った)」の２つの候補が生成されたとする。係り受けモデルにより、このうち尤もらしい係り受け構造を持つ候補が優先される。

【００３１】

以上に示すような各モデルを用い、本発明では評価部（７）においてテキスト候補（１２・１３など）に評価付けを行う。

そして、評価値が最大あるいは閾値を超えるテキスト候補、あるいは評価値の上位Ｎ個を表層文に変換して出力する。

【００３２】

出力方法としては、モニタによる表示の他、音声合成を用いた発声、翻訳システムなど他の言語処理システムへのデータ出力などが可能である。

これにより、例えばキーワード（２）として「彼女」「家」「行く」を入力したときに、「彼女の家に行く」（３ａ）「彼女が家に行った」（３ｂ）などのテキスト（３）を出力することができる。なお、出力は上記した通り、最も評価値の高いものを１つ選択してもよいし、評価値順に複数出力してもよく、例えば複数提示してキーワードを入力した者が最適なものを選択するようにしてもよい。

【００３３】

以上に示した実施例では、キーワードの前後に語句を付加する構成を主としているが、本発明の実施においてはキーワード（主辞単語に相当するもの）そのものを補完する構成をとることもできる。

例えば、「彼　本」から述語などを補完して「彼が本を読んだ」や「彼が本を書いた」、「彼が本を買った」などを生成するために、入力されたキーワードに関連する新たなキーワードを追加して入力することも可能である。

【００３４】

具体的な実施例としては、図２の構成に図５の要素を追加する。すなわち、キーワード

（２）を係り受け関係語抽出部（１４）にも入力し、該部（１４）ではコーパス（８）から該キーワード（２）と係り受け関係にある単語を抽出する。
そして、単語を新たなキーワードとして加え、もともと入力されたキーワード（２）と合わせて文節候補生成部（５）における処理を行う。
【００３５】
例えば、「（彼が（本を　読む））」そのものがコーパス（８）に無くとも、「（彼が　読む）」と「（本を　読む）」という係り受け関係がそれぞれコーパス（８）にあれば、それらに共通する単語「読む」を新たにキーワードとして追加することによって、文節候補生成部（５）によって「読んだ」が生成できるようになる。
【００３６】
本構成は、計算量が少なく高速なキーワードの追加が可能であるが、本発明では必ずしもコーパスから係り受け関係にある単語を抽出することに限らず、任意のキーワードの候補を追加し、その中から評価部（７）における評価が結果的に最も高くなるものを出力してもよい。
これによって、キーワードにテキストの意味を決定する重要な単語が一部欠落していたとしても、有意なテキストが出力できるようになる。
【００３７】
なお、上記実施例では生成するテキストを日本語とし、キーワードから文節を生成していたが、本発明は任意の言語に適用可能である。
例えば、英語の場合、複数の名詞句、動詞句が集まって別の名詞句や動詞句を形成し、階層をなす場合がある。そのような言語では、中でも最も小さな句、「基本句」を文節の代わりに用いることもできる。
【００３８】
最後に、本発明によるテキスト生成方法を用いた実験例を示す。実験は表１に示す３つの入力するキーワードで行い、出力されたテキストが主観的に正しいか否かで評価した。評価基準は以下の２つである。
【００３９】
基準１：　１位の候補が意味的、文法的に適切であればシステムの出力が正しいと判断する。
基準２：　上位１０位に意味的、文法的に適切な候補があればシステムの出力が正しいと判断する。
評価結果を表２に示す。
【００４０】
生成規則により生成されたテキスト候補の数は、２つのキーワードが入力の場合、キーワード一組あたり平均868.8 個(26,064/30) 、３つのキーワードが入力の場合、キーワード一組あたり平均41,413.5個(1,242,404/30) であった。
表２では、前述したキーワード生成モデル（ＫＭ１ないし５）と言語モデル（ＭＭ及びＤＭ）について、+ は各モデルの組み合わせを表わしている。
【００４１】
表２から分かるように、ＫＭ１やＫＭ３、ＫＭ５のモデルにＭＭとＤＭを組み合わせた場合が最も良い結果となった。ＭＭやＤＭを用いた場合、用いなかった場合と比べて基準１による評価結果が飛躍的に良くなっているが、その理由は、名詞と格の結び付きより、動詞と格の結び付きの方が強く、後者に着目して学習しているＫＭ１やＫＭ３、ＫＭ５のモデルが潜在的に自然な文となる候補を上位に順序付けていたからである可能性が高いと考えられる。
【００４２】
以上の結果から、本発明の評価部（７）では上述したキーワード生成モデル１，３，５と、形態素モデル、係り受けモデルを組み合わせて評価部を構成するのが望ましく、とりわけＫＭ３を用いると特に好適である。
これらの組み合わせでは、表２に見るように９割前後の割合で正しいテキストを生成す

ることに成功した。

【００４３】

【発明の効果】

本発明は、以上の構成を備えるので、次の効果を奏する。

すなわち、請求項１又は２に記載のテキスト生成装置によれば、従来のテキスト生成方法では困難であった入力するキーワードが十分でない場合にも、有意なテキストを生成することができる。

【００４４】

特に、請求項２に記載のテキスト生成装置では、キーワードと係り受け関係を持つ単語を抽出し、キーワードを追加することができるため、より広義のテキスト生成が実現可能となる。

【００４５】

本テキスト生成装置では、抽出された文・語句から文節等候補の生成規則を自動獲得することができるため、効率的に文節等候補を生成することができ、処理の高速化、低コスト化に寄与する。

本発明は、上記のように優れたテキスト生成方法を提供するテキスト生成装置を創出し、自然言語処理技術の向上に寄与することが出来る。

【図面の簡単な説明】

【図１】　本発明によるテキスト生成装置の説明図である。

【図２】　同、構成図である。

【図３】　キーワードからのテキスト生成の例を示す説明図である。

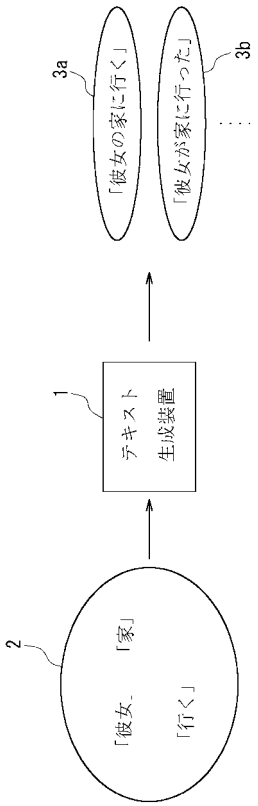
【図４】　キーワードと単語の関係を示す説明図である。

【図５】　本発明に係る係り受け関係語抽出部の構成図である。

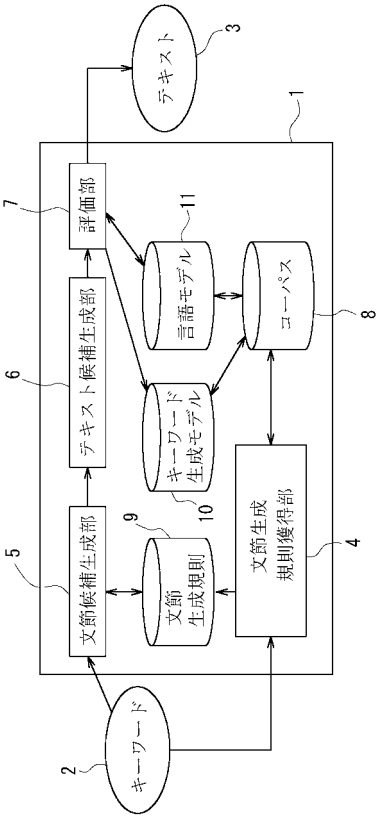
【符号の説明】

１　テキスト生成装置
２　キーワード
３　テキスト
４　文節生成規則獲得部
５　文節候補生成部
６　テキスト候補生成部
７　評価部
８　コーパス
９　文節生成規則
１０　キーワード生成モデル
１１　言語モデル

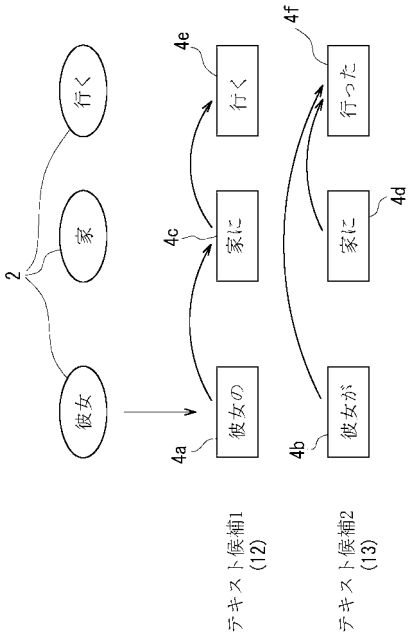
【図１】


「彼女の家に行く」
「彼女が家に行った」
3a
3b
テキスト
生成装置
1
2
「彼女」
「家」
「行く」

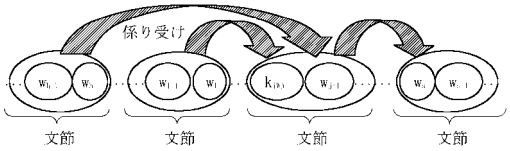
【図２】


テキスト
3
1
評価部
7
テキスト候補生成部
6
文節候補生成部
5
言語モデル
11
コーパス
8
キーワード
生成モデル
10
文節
生成規則
9
文節生成
規則獲得部
4
キーワード
2

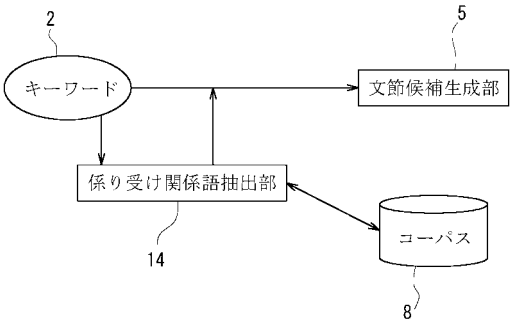
【図３】


彼女
家
行く
2
彼女の
家に
行く
4a
4c
4e
彼女が
家に
行った
4b
4d
4f
テキスト候補1
(12)
テキスト候補2
(13)

【図４】


係り受け
文節
文節
文節
文節

【図５】


2
キーワード
5
文節候補生成部
係り受け関係語抽出部
14
コーパス
8

フロントページの続き

(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

合議体
審判長　長島　孝志
審判官　野崎　大進
審判官　菅原　浩二

(56)参考文献　特開平０５－２５０４０７（ＪＰ，Ａ）
内元清貴・井佐原均「最大エントロピーモデルを用いた日本語テキストの一貫処理」人工知能学会研究資料ＳＩＧ－ＣＩＩ－２０００－ＮＯＶ－０９（２０００．１１．１４）
内元清貴・村田真樹・馬清・内山将夫・関根聡・井佐原均「コーパスからの語順の学習」情報処理学会研究報告２０００－ＮＬ－１３５－８，Ｖｏｌ．２０００，Ｎｏ．１１，ｐ．５５－ｐ．６２（２０００．０１．２８）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F17/28
G06F17/30