(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4831737号
(P4831737)
(45)発行日　平成23年12月7日(2011.12.7)　(24)登録日　平成23年9月30日(2011.9.30)

(51)Int.Cl.　F I
G06F 17/21 (2006.01)　G06F 17/21 550A
G06F 17/21 564P

請求項の数 5　(全 16 頁)

(21)出願番号　特願2006-28325(P2006-28325)
(22)出願日　平成18年2月6日(2006.2.6)
(65)公開番号　特開2007-207161(P2007-207161A)
(43)公開日　平成19年8月16日(2007.8.16)
審査請求日　平成20年12月15日(2008.12.15)

(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　100103827
弁理士　平岡　憲一
(74)代理人　100119161
弁理士　重久　啓子
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　成瀬　博之

最終頁に続く

(54)【発明の名称】キーワード強調装置及びプログラム

(57)【特許請求の範囲】
【請求項１】
質問とその回答の記事のセットを入力する入力手段と、
前記質問の文から疑問詞に後接する名詞又は疑問詞に後接する接尾辞を取り出す疑問詞後接語抽出手段と、
前記回答の記事において取り出した前記疑問詞に後接していた名詞又は接尾辞を強調表示する表示手段とを備えることを特徴としたキーワード強調装置。
【請求項２】
質問とその回答の記事のセットを入力する入力手段と、
前記質問の文から疑問詞に後接する、数字と結合できる所定の名詞、又は、疑問詞に後接する、数字と結合できる所定の接尾辞を取り出す疑問詞後接語抽出手段と、
前記回答の記事において数字と前記取り出した所定の名詞又は所定の接尾辞のうち少なくとも一つを強調表示する表示手段とを備えることを特徴としたキーワード強調装置。
【請求項３】
質問とその回答の記事のセットを入力する入力手段と、
前記質問の文から所定の数量表現を指す疑問詞があることを確認する抽出手段と、
前記抽出手段で確認した前記質問の文に前記所定の数量表現を指す疑問詞があり、前記質問の回答が数字であることにより、前記回答の記事において数字を強調表示する表示手段とを備えることを特徴としたキーワード強調装置。
【請求項４】

質問とその回答の記事のセットを入力する入力手段と、
前記質問の文から、人名を指すか、地名を指すか、時間を指すかの予め指定した疑問詞の種類を特定する抽出手段と、
前記回答の記事において前記疑問詞の種類に対応する、人名、地名、時間の固有表現を抽出して強調表示する表示手段とを備えることを特徴としたキーワード強調装置。
【請求項５】
質問とその回答の記事のセットを入力する入力手段と、
前記質問の文から疑問詞に後接する名詞又は疑問詞に後接する接尾辞を取り出す疑問詞後接語抽出手段と、
前記回答の記事において取り出した前記疑問詞に後接していた名詞又は接尾辞を強調表示する表示手段として、
コンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ユーザが指定した領域に含まれる語の中で、例えば、「疑問詞」＋「数字と結合できる名詞（時、月、年、歳、枚、など）」で示されるキーワードに対応して、本文中において「数値」＋「数字と結合できる名詞」で表される部分を強調表示することで、疑問詞の問いかけに対応する回答部分が容易に判るようにするキーワード強調装置及びプログラムに関する。
【背景技術】
【０００２】
従来のキーワード入力に対する検索結果の強調表示システムは、タイトル中に出てきた単語を本文中において強調表示するものであった（特許文献１参照）。
【特許文献１】特開２００４－２８０１７６号公報
【発明の開示】
【発明が解決しようとする課題】
【０００３】
上記従来の強調表示システムは、タイトルが質問文となり本文が回答文となったものにおいて、回答文の中で質問の疑問詞に対応する部分を強調表示できるものではなかった。
【０００４】
本発明は上記問題点の解決を図り、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分を容易に見つけるようにすることを目的とする。
【課題を解決するための手段】
【０００５】
図１は本発明のキーワード強調装置の説明図である。図１中、１は表示装置（表示手段）、２は入力装置（入力手段）、３は抽出手段（抽出装置）、４は疑問詞後接語抽出装置（疑問詞後接語抽出手段）、５は主要語抽出装置（主要語抽出手段）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）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から疑問詞に後接する名詞又は疑問詞に後接する接尾辞を取り出す疑問詞後接語抽出手段４と、前記回答の記事において取り出した前記疑問詞に後接していた名詞又は接尾辞を強調表示する表示手段１とを備える。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分を容易に見つけることができる。
【０００８】
（２）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から疑問詞に後接する数字と結合できる所定の名詞又は疑問詞に後接する数字と結合できる所定の接尾辞を取り出す疑問詞後接語抽出手段４と、前記回答の記事において数字と前記取り

出した所定の名詞又は所定の接尾辞のうち少なくとも一つを強調表示する表示手段１とを備える。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する回答（数字）の表示部分を容易に見つけることができる。

【０００９】

（３）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から所定の数量表現を指す疑問詞があることを確認する抽出手段３と、前記回答の記事において数字を強調表示する表示手段１とを備える。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する回答（数字）の表示部分を容易に見つけることができる。

【００１０】

（４）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から予め指定した疑問詞の種類を特定する抽出手段３と、前記回答の記事において前記疑問詞の種類に対応する固有表現を抽出して強調表示する表示手段１とを備える。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する固有表現の表示部分を容易に見つけることができる。

【００１１】

（５）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から予め指定した理由を指す疑問詞を特定する抽出手段３と、前記回答の記事において前記理由を示す所定の単語を強調表示する表示手段１とを備える。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する理由を示す表示部分を容易に見つけることができる。

【００１２】

（６）：前記（１）～（５）のキーワード強調装置において、前記質問の文から主要語を取り出す主要語抽出手段５を備え、前記表示手段１は、前記回答の記事において前記取り出した主要語を強調表示する。このため、強調表示される主要語の周辺の回答文書の中で、本当に知りたい疑問詞に対応する表示部分（回答）を容易に見つけることができる。

【００１３】

（７）：前記（６）のキーワード強調装置において、前記表示手段１で強調表示する主要語と他の強調表示では、異なる強調表示を行う。このため、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分をより簡単に見つけることができる。

【００１４】

（８）：質問とその回答の記事のセットを入力する入力手段２と、前記質問の文から疑問詞に後接する名詞又は疑問詞に後接する接尾辞を取り出す疑問詞後接語抽出手段４と、前記回答の記事において取り出した前記疑問詞に後接していた名詞又は接尾辞を強調表示する表示手段１として、コンピュータを機能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、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分を容易に見つけることができるキーワード強調装置を容易に提供することができる。

【発明の効果】

【００１５】

本発明によれば次のような効果がある。

（１）：表示手段で、取り出した疑問詞に後接していた名詞又は接尾辞を回答の記事において強調表示するため、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分を容易に見つけることができる。

【００１６】

（２）：表示手段で、数字と取り出した所定の名詞又は所定の接尾辞のうち少なくとも一つを回答の記事において強調表示するため、表示された回答文書の中で本当に知りたい疑問詞に対応する回答（数字）の表示部分を容易に見つけることができる。

【００１７】

（３）：表示手段で、数字を回答の記事において強調表示するため、表示された回答文書の中で本当に知りたい疑問詞に対応する回答（数字）の表示部分を容易に見つけること

ができる。

【００１８】

（４）：表示手段で、疑問詞の種類に対応する固有表現を抽出して、回答の記事において強調表示するため、表示された回答文書の中で本当に知りたい疑問詞に対応する固有表現の表示部分を容易に見つけることができる。

【００１９】

（５）：表示手段で、理由を示す所定の単語を回答の記事において強調表示するため、表示された回答文書の中で本当に知りたい疑問詞に対応する理由を示す表示部分を容易に見つけることができる。

【００２０】

（６）：表示手段で、取り出した主要語を回答の記事において強調表示するため、回答文書の中で本当に知りたい疑問詞に対応する表示部分を容易に見つけることができる。

【００２１】

（７）：表示手段で強調表示する主要語と他の強調表示では、異なる強調表示を行うため、表示された回答文書の中で本当に知りたい疑問詞に対応する表示部分をより簡単に見つけることができる。

【発明を実施するための最良の形態】

【００２２】

本発明のキーワード強調装置は、ユーザが指定した領域に含まれる語の中で、例えば、「疑問詞」＋「数字と結合できる名詞（時、月、年、歳、枚、など）」で示されるキーワードに対応して、本文中において「数値」＋「数字と結合できる名詞」で表される部分を強調表示することで、疑問詞の問いかけに対応する回答部分が容易に判るようにするものである。

【００２３】

Ｗｅｂ（ウエブ）サイトでの質問とその回答やＦＡＱ（よくある質問とその回答）のように、質問と回答の記事を人手で作成し蓄えておき、ユーザに提示するということが多くなってきている。そのときに、本発明のような強調表示を使用すると、質問に対する回答が容易に判るようになる。

【００２４】

（１）：キーワード強調装置の説明

図１はキーワード強調装置の説明図である。図１において、キーワード強調装置（システム）には、表示装置１、入力装置２、抽出装置３が設けてある。抽出装置３には、疑問詞後接語抽出装置４、主要語抽出装置５が設けてある。

【００２５】

表示装置１は、情報を表示するＣＲＴ、液晶等の表示画面を備えた表示手段である。入力装置２は、情報を入力する入力手段である。抽出手段３は、単語の抽出処理等を行う抽出装置（処理手段）である。疑問詞後接語抽出装置４は、疑問詞の後ろにくる名詞や接尾辞を抽出する疑問詞後接語抽出手段である。主要語抽出装置５は、あまり意味のない単語（「もの」「こと」等の予め指定した単語）を除いた名詞や動詞等を抽出する主要語抽出手段である。

【００２６】

（２）：疑問詞の後ろに付く単語を強調表示する説明(1)

図２は疑問詞の後ろに付く単語を強調表示するフローチャートである。以下、図２の処理Ｓ１～Ｓ４に従って説明する。

【００２７】

Ｓ１：入力装置２により質問とその回答の記事のセットが与えられ、処理Ｓ２に移る。

Ｓ２：疑問詞後接語抽出装置４は、質問の文から疑問詞＋「名詞or接尾辞」を取り出し、処理Ｓ３に移る。

【００２８】

Ｓ３：主要語抽出装置５は、質問の文から主要語を取り出し、処理Ｓ４に移る。

ここで主要語は、名詞や動詞などである。ただし、あらかじめ指定した所定の単語は除く（例えば、「もの」「こと」などのあまり意味をなさない単語）。

【００２９】

Ｓ４：表示装置１は、回答の記事において取り出した主要語、疑問詞に後接していた「名詞or接尾辞」を強調表示（常にバックに黄色を出すなど）する。

【００３０】

例：・・・・何大学・・・・・の質問の場合、回答本文で、大学を黄色で強調表示する。これにより、強調表示部分を見ることで、質問に対する回答を容易に見つけることができる。

【００３１】

なお、ここで強調表示とは、文字の色を変えて表示する、文字の背景の色を変える又は網かけを行う、文字の字体を変える（太文字、斜体文字等）、下線付けや括弧で囲む、文字の上に記号等を設ける等で行うことができる。

【００３２】

（ＦＡＱの具体例による説明）

（質問）東京で偏差値の高いのは何大学ですか.

（回答）受験する学部により偏差値の値は異なりますが、一般的に東京大学の偏差値が各学部とも高いようです.

キーワード強調装置では、以下のように強調表示する（ここでは「＜」、「＞」で強調表示）。

（質問）東京で偏差値の高いのは何＜大学＞ですか.

（回答）受験する学部により偏差値の値は異なりますが、一般的には東京＜大学＞の偏差値が各学部とも高いようです.

【００３３】

（３）：単語の切り出し品詞の特定の説明

疑問詞、名詞、接尾辞、動詞の単語の抽出は、形態素解析を使用して行うことができる。

【００３４】

（形態素解析システムの説明）

ここでは ChaSen （日本語）について説明する。奈良先端大で開発されている形態素解析システム茶筌　http://chasen.aist-nara.ac.jp/index.html.jaで公開されている。

これは、日本語文を分割し、さらに、各単語の品詞も推定してくれる。

【００３５】

例えば、「学校へ行く」を入力すると以下の結果をえる。

```
学校　ガッコウ　学校　名詞- 一般
へ　へ　へ　助詞- 格助詞- 一般
行く　イク　行く　動詞- 自立　五段・カ行促音便　基本形
EOS
```

このように、各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。

【００３６】

英語の品詞のタグ付けの説明

英語の品詞タグつけシステムとしては、次の Brillのものが有名である。

Eric Brill,
Transformation-Based Error-Driven Learning and
Natural Language Processing: A Case Study in Part-of-Speech Tagging,
Computational Linguistics, Vol. 21, No. 4, p.543-565, 1995.

これは、英語文の各単語の品詞を推定してくれるものである。

【００３７】

（４）：疑問詞の後ろに付く単語を利用して強調表示する説明(2)

図３は疑問詞の後ろに付く単語を強調表示するフローチャートである。以下、図３の処理Ｓ１１～Ｓ１４に従って説明する。

【００３８】

Ｓ１１：入力装置２により質問とその回答の記事のセットが与えられ、処理Ｓ１２に移る。

【００３９】

Ｓ１２：疑問詞後接語抽出装置４は、質問の文から疑問詞＋「数字と結合できる所定の名詞or接尾辞」を取り出し、処理Ｓ１３に移る。

【００４０】

Ｓ１３：主要語抽出装置５は、質問の文から主要語を取り出し、処理Ｓ１４に移る。

ここで主要語は、名詞や動詞などである。ただし、あらかじめ指定した所定の単語（例えば、「もの」「こと」などのあまり意味をなさない単語）は除く。

【００４１】

Ｓ１４：表示装置１は、回答の記事において取り出した主要語、数字＋「取り出した名詞or接尾辞」を強調表示する。数字＋「取り出した名詞or接尾辞」は、それ専用の強調表示（例えば、主要語とは異なる色（常にバックに黄色を出すなど））する。

例：・・・・何個・・・・・の質問の場合、回答本文で、「３個」を黄色で強調表示する。これにより、強調表示部分を見ることで、質問に対する回答を容易に見つけることができる。

【００４２】

（ＦＡＱの具体例による説明）

（質問）睡眠時間は何時間くらいがいいですか.

（回答）諸説別れますが，７時間から８時間がよいという説が一般的です．でもいつ寝るかも重要に思います．昼間長時間寝ても，夜寝るのに比べて効果が低いと思います．

キーワード強調装置では、以下のように強調表示する（ここでは「＜」、「＞」で強調表示）。

【００４３】

（質問）睡眠時間は何＜時間＞くらいがいいですか.

（回答）諸説別れますが，＜７時間＞から＜８時間＞がよいという説が一般的です．でもいつ寝るかも重要に思います．昼間長時間寝ても，夜寝るのに比べて効果が低いと思います．

また、以下のように強調表示することもできる（ここでは「＜」、「＞」で強調表示）。

（質問）睡眠時間は何＜時間＞くらいがいいですか.

（回答）諸説別れますが，７＜時間＞から８＜時間＞がよいという説が一般的です．でもいつ寝るかも重要に思います．昼間長＜時間＞寝ても，夜寝るのに比べて効果が低いと思います．

このように、すぐに７時間、８時間の表現に目がいき便利となる。

【００４４】

（５）：数量表現を指す疑問詞を利用して強調表示する説明

図４は数量表現を指す疑問詞を利用して強調表示するフローチャートである。以下、図４の処理Ｓ２１～Ｓ２４に従って説明する。

【００４５】

Ｓ２１：入力装置２により質問とその回答の記事のセットが与えられ、処理Ｓ２２に移る。

【００４６】

Ｓ２２：抽出手段３は、質問の文から所定の数量表現を指す疑問詞があることを確認し、処理Ｓ２３に移る。

【００４７】

Ｓ２３：主要語抽出装置５は、質問の文から主要語を取り出し、処理Ｓ２４に移る。

ここで主要語は、名詞や動詞などである。ただし、あらかじめ指定した所定の単語（例えば、「もの」「こと」などのあまり意味をなさない単語）は除く。

【００４８】

Ｓ２４：表示装置１は、回答の記事において取り出した主要語、数字を強調表示する。ここで数字はそれ専用の強調表示（例えば、主要語（例えば赤）とは異なる色（常にバックに黄色を出すなど））する。

例：・・・・いくつ・・・・の質問の場合、回答本文で、３個の「３」を黄色で強調表示する。これにより、強調表示部分を見ることで、質問に対する回答を容易に見つけることができる。

【００４９】

この場合、回答が数字となる疑問詞は、予めキーワード強調装置の格納手段（図示せず）に記憶して置くものである。回答が数字となる疑問詞として、「いかほど」、「どのくらい」等がある。

【００５０】

（ＦＡＱの具体例による説明）

（質問）睡眠時間はどのくらいがいいですか.

（回答）諸説別れますが，７時間から８時間がよいという説が一般的です．でもいつ寝るかも重要に思います．昼間長時間寝ても，夜寝るのに比べて効果が低いと思います.

キーワード強調装置では、以下のように強調表示する（ここでは「＜」、「＞」で強調表示）。

（質問）睡眠時間はどのくらいがいいですか.

（回答）諸説別れますが，＜７＞時間から＜８＞時間がよいという説が一般的です．でもいつ寝るかも重要に思います．昼間長時間寝ても，夜寝るのに比べて効果が低いと思います.

【００５１】

（６）：疑問詞の意味を利用して強調表示する説明(1)

図５は疑問詞の意味を利用して強調表示するフローチャートである。以下、図５の処理Ｓ３１～Ｓ３４に従って説明する。

【００５２】

Ｓ３１：入力装置２により質問とその回答の記事のセットが与えられ、処理Ｓ３２に移る。

【００５３】

Ｓ３２：抽出装置３は、質問の文から疑問詞の種類を特定し、処理Ｓ３３に移る。人名をさすか、地名をさすか、時間をさすか、など。どの疑問詞なら何の種類であるかといった所定の規則みたいなものは予め用意しておく。

【００５４】

Ｓ３３：主要語抽出装置５は、質問の文から主要語を取り出し、処理Ｓ３４に移る。

ここで主要語は、名詞や動詞などである。ただし、あらかじめ指定した所定の単語（例えば、「もの」「こと」などのあまり意味をなさない単語）は除く。

【００５５】

Ｓ３４：表示装置１は、回答の記事において取り出した主要語、
人名をさす疑問詞（例、「誰」）の場合は人名を
地名をさす疑問詞（例、「どこ」）の場合は地名を
時間をさす疑問詞（例、「いつ」）の場合は時間（春、夏等の季節も含む）を
それ専用の強調表示（常にバックに黄色を出すなど）する。

【００５６】

なお、ここで各単語が人名、地名、時間を指すかを判断するには、固有表現抽出の技術を利用する。

【００５７】

（ＦＡＱの具体例による説明）

（質問）今年もっとも世間を騒がせた人物は誰でしょうか．
（回答）今年もいろいろとありましたが，総選挙，買収劇と，多方面に目立った人は，堀江氏でしょう．来年はどういった人物が出てくるか楽しみですね．
キーワード強調装置では、以下のように強調表示する（ここでは「＜」、「＞」で強調表示）。
（質問）今年もっとも世間を騒がせた人物は＜誰＞でしょうか．
（回答）今年もいろいろとありましたが，総選挙，買収劇と，多方面に目立った人は，＜堀江氏＞でしょう．来年はどういった人物が出てくるか楽しみですね．

【００５８】

（７）：疑問詞の意味を利用して強調表示する説明(2)

図６は疑問詞の意味を利用して強調表示するフローチャートである。以下、図６の処理Ｓ４１～Ｓ４４に従って説明する。

【００５９】

Ｓ４１：入力装置２により質問とその回答の記事のセットが与えられ、処理Ｓ４２に移る。

【００６０】

Ｓ４２：抽出手段３は、質問の文から疑問詞の種類を特定し、処理Ｓ４３に移る。ここでは疑問詞が理由を指すもの（例えば、「なぜ」「どうして」）であるとする。どの疑問詞なら何の種類であるかといった所定の規則みたいなものは予め用意しておく。

【００６１】

Ｓ４３：主要語抽出装置５は、質問の文から主要語を取り出し、処理Ｓ４４に移る。
ここで主要語は、名詞や動詞などである。ただし、あらかじめ指定した所定の単語（例えば、「もの」「こと」などのあまり意味をなさない単語）は除く。

【００６２】

Ｓ４４：表示装置１は、回答の記事において取り出した主要語、理由を示す所定の単語「ので」「ため」「から」「だから」「理由」「原因」「このため」などを、それ専用の強調表示（常にバックに黄色を出すなど）を行う。

【００６３】

（ＦＡＱの具体例による説明）
（質問）なぜコンピュータは便利なのでしょうか．
（回答）コンピュータは計算機とも呼ばれるもので，人間に代わって様々な計算をしてくれる便利な機械です．コンピュータは，一般に演算装置と記憶装置からなります．コンピュータは，プログラムを与えると演算装置と記憶装置でそれを実行し様々な計算をします．与えるプログラムを変えると，コンピュータはそれに応じた異なった処理を実行することができます．このため，コンピュータは様々な処理をできて便利なのです．

【００６４】

キーワード強調装置では、以下のように強調表示する（ここでは「＜」、「＞」で強調表示）。
（質問）なぜコンピュータは便利なのでしょうか．
（回答）コンピュータは計算機とも呼ばれるもので，人間に代わって様々な計算をしてくれる便利な機械です．コンピュータは，一般に演算装置と記憶装置からなります．コンピュータは，プログラムを与えると演算装置と記憶装置でそれを実行し様々な計算をします．与えるプログラムを変えると，コンピュータはそれに応じた異なった処理を実行することができます．＜このため＞，コンピュータは様々な処理をできて便利なのです．
このように、強調表示した「このため」の前方に理由が書いてあることがすぐにわかり，便利である。

【００６５】

なお、キーワード強調装置を使用するユーザにおいて、前記の強調表示は行わない設定も可能である。

【００６６】

また、ここで各単語が人名、地名、時間を指すかを判断するには、固有表現抽出の技術を利用する。

【００６７】

（８）：固有表現抽出の説明

固有表現とは、人名、地名、組織名などの固有名詞、金額などの数値表現といった、特定の事物・数量を意味する言語表現のことである。固有表現抽出とは、そういった固有表現を文章中から計算機で自動で抽出する技術である。例えば、「日本の首相は小泉純一郎である」という文に対して固有表現抽出を行なうと、固有表現の「日本」と「小泉純一郎」が地名、人名として、抽出されるものである。

【００６８】

ａ、形態素解析を用いる場合の説明

固有表現を抽出するには、前に説明した形態素解析システム ChaSen を用いることができる。例えば、「日本の首都は東京です」を形態素解析システム ChaSen に入力すると、出力として、次のものが得られる。

【００６９】

出力
日本　ニッポン　日本　名詞－固有名詞－地域－国
の　ノ　の　助詞－連体化
首都　シュト　首都　名詞－一般
は　ハ　は　助詞－係助詞
東京　トウキョウ　東京　名詞－固有名詞－地域－一般
です　デス　です　助動詞特殊・デス基本形
EOS

これだと名詞－固有名詞－地域という品詞が出力されるので、このシステムを使って地名の固有表現を取り出すことができる。

【００７０】

また、例えば、前記システムに「村山首相が言った」を入力すると、出力として、次のものが得られる。

【００７１】

出力
村山　ムラヤマ　村山　名詞－固有名詞－人名－姓
首相　シュショウ　首相　名詞－一般
が　ガ　が　助詞－格助詞－一般
言っ　イッ　言う　動詞－自立五段・ワ行促音便連用タ接続
た　タ　た　助動詞特殊・タ基本形
EOS

これだと名詞－固有名詞－人名という品詞が出力される。このシステムを使って人名の固有表現を取り出すことができる。

【００７２】

ｂ、人手でルールを作る場合の説明

形態素解析を用いる場合の他に、人手でルールを作って固有表現を取り出すという方法もある。

【００７３】

例えば、人手でルールを作っておくことで、抽出手段（装置）では、次のルールで固有表現（人名、地名等）を取り出すことができる。

名詞＋「さん」だと人名とする
名詞＋「首相」だと人名とする
名詞＋「町」だと地名とする
名詞＋「市」だと地名とする

【００７４】

ｃ、機械学習を用いる場合の説明

（ユーザ依存型固有表現抽出表示システムの説明）

一部のコーパス（言語資源、例えば、新聞の電子データ）で固有表現をユーザがタグづけし、他のデータでそれら固有表現を自動抽出する技術である。

【００７５】

固有表現の抽出には、学習結果を利用して、入力データの所定の単位のデータについてその素性の場合になりやすい分類先を推定するものである。

【００７６】

例えば、固有表現の抽出に、サポートベクトルマシン法を用いる場合には、機械学習手段では、教師データから解となりうる分類先を特定し、その分類先を正例と負例に分割し、所定のカーネル関数を用いたサポートベクトルマシン法を実行する関数にしたがって素性の集合を次元とする空間上で正例と負例の間隔を最大にして正例と負例を超平面で分割する超平面を求め、その超平面を学習結果とし、その超平面を学習結果記憶手段に記憶する。そして、この学習結果記憶手段に記憶されている学習結果の超平面を利用して、入力データの素性の集合がこの超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果に基づいて定まる分類先を、入力データの素性の集合の場合になりやすい分類先と推定する。

【００７７】

固有表現抽出処理とは、テキストデータから地名、人名、組織名、数値表現などの固有な表現を抽出する処理をいう。固有表現抽出処理において解析結果となる分類先は、例えば地名、人名、組織名、日付表現、時間表現、金額表現、割合表現などである。教師データには、これらの分類先それぞれに対応する分類ラベルが付与される。

【００７８】

教師データ作成のためのタグ登録手段は、ユーザが、入力装置を介して、以下のような固有表現抽出処理の分類先とそれに対応する分類タグを指定すると、ユーザが指定した分類先およびその分類タグ（開始タグと終了タグ）を入力してタグ記憶手段に記憶する。

【００７９】

＜PERSON＞＜/PERSON ＞：分類先＝人名、
＜LOCATION＞＜/LOCATION ＞：分類先＝地名、
＜ORGANIZATION＞＜/ORGANIZATION ＞：分類先＝組織名、
＜ARTIFACT＞＜/ARTIFACT ＞：分類先＝固有物名、
＜DATE＞＜/DATE ＞：分類先＝日付表現、
＜TIME＞＜/TIME ＞：分類先＝時間表現、
＜MONEY ＞＜/MONEY＞：分類先＝金額表現、
＜PERCENT ＞＜/PERCENT＞：分類先＝割合表現、…。

【００８０】

本例では、付与する分類ラベルを文字単位に付与した教師データを作成する。例えば、＜PERSON＞＜/PERSON ＞分類タグが対応する分類先「人名」の分類ラベルは、先頭文字を示す「B-」または先頭以外の文字を示す「I-」を付けて、「B-PERSON」、「I-PERSON」とする。また、分類先に該当しない文字に付与するラベルとして、「OTHER 」を登録する。

【００８１】

また、固有表現抽出処理の分類先として字種を用いる場合には、以下のような分類先および分類タグをタグ記憶手段に格納する。

【００８２】

＜KANJI ＞＜/KANJI＞：分類先＝漢字、
＜KATAKANA＞＜/KATAKANA ＞：分類先＝カタカナ、
＜ALPHABETIC＞＜/ALPHABETIC ＞：分類先＝英字、
＜NUMERIC ＞＜/NUMERIC＞：分類先＝数字。

【００８３】

そして、コーパス入力手段が、固有表現抽出処理の分類先が付与されていないテキスト

データで構成されるコーパスを入力すると、タグ付与手段は、コーパスのテキストデータを表示しユーザにタグ付与操作を促すタグ付与画面を表示装置に表示する。

【００８４】

ユーザによって、分類先を付与したい箇所および付与する分類先が指定されたら、タグ付与手段は、タグ付与画面で指定された箇所に対応する文字列の前後に選択された分類タグを挿入する。

【００８５】

例えば、入力されたコーパスに、テキストデータ「…日本の首相は小泉さんです。小泉さんはいつも思いきったことをしています。…」が含まれていたとする。ユーザが、タグ付与画面の指定項目に表示されたテキストデータ上で、マウスドラッグ操作などにより、分類先を付与する単語「日本」を指定する。さらにマウスの右ボタンクリック操作を行って表示させた選択項目から、マウス左ボタンクリック操作などにより分類先「地名」を選択する。同様に、指定項目で単語「小泉」を指定し、選択項目から分類先「人名」を選択する。

【００８６】

タグ付与手段は、タグ付与画面で指定された箇所に対応するテキストデータ中の文字列の前後に、選択された分類タグを挿入する。分類タグが付与されたテキストデータは以下のようになる。

「…＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。小泉さんはいつも思いきったことをしています。…」

さらに、ユーザによって、指定項目で分類先を付与する作業を行い教師データとして使用する範囲が指定されると、タグ付与手段は、タグ付与画面で指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグの開始タグおよび終了タグを付加する。例えば、ユーザが、マウスドラッグにより文「日本の首相は小泉さんです。」を範囲として指定したとする。タグ付与手段は、指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグを挿入する。範囲指定タグが付与されたテキストデータは以下のようになる。

「…＜UC＞＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。＜/UC ＞小泉さんはいつも思いきったことをしています。…」

一方、ユーザが、分類先を付与した後、教師データとして使用する範囲を指定しなかった場合には、タグ付与手段は、指定項目で分類先が付与された箇所を含む所定の箇所をユーザが選択した範囲とみなし、その範囲の前後に範囲指定タグを付加する。例えば、タグ付与手段は、テキストデータ中の分類タグが付与された文字列に単語の前後に連なる所定の文字数や単語数などの範囲を、ユーザが選択した範囲とみなし、みなした範囲の前後に範囲指定タグを付加する。

【００８７】

そして、タグ付与手段は、テキストデータに分類タグおよび範囲指定タグを付加したテキストデータ（タグ付きコーパス）をコーパス記憶手段に記憶する。

【００８８】

その後、ユーザ範囲抽出手段は、コーパス記憶手段のタグ付きコーパスから、範囲指定タグの開始タグ＜UC＞と終了タグ＜/UC ＞とに囲まれた範囲のテキストデータ（ユーザ範囲データ）を抽出する。なお、ここではユーザがUCのタグを付ける説明をしたが、システム作成者がこのタグを付与することもでき、また、UCのタグを付けずに全データを教師データとして使用することも可能である。

【００８９】

そして、教師データ変換手段は、抽出されたテキストデータを所定の単位（ここでは文字単位とする）に分割し、抽出されたテキストデータから分類タグに囲まれた文字列を検出し、各単位（文字）のうち分類タグが付与されている文字に分類タグに対応する分類ラベルを付与し、分類タグが付与されていない文字に分類先がないことを示す分類ラベルを付与して、教師データとする。

【００９０】
例えば、教師データとして、範囲指定タグに囲まれたテキストデータ「＜UC＞＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。＜/UC ＞」が抽出されたとする。教師データ変換手段は、例えば、テキストデータの分類タグ＜PERSON＞と＜/PERSON ＞に囲まれた文字列「小、泉」の先頭文字「小」に、分類先「人名」の先頭を示す分類ラベル「B-PERSON」を、同じく次の文字「泉」に分類先「人名」の先頭以外を示す分類ラベル「I-PERSON」を付与する。また、テキストデータのうち分類タグに囲まれていない部分「の、首、相、は、さ、ん、で、す、。」について、各文字にユーザが指定した分類先に該当しない旨を示す分類ラベル「０」を付与する。
【００９１】
そして、素性抽出手段により、教師データに対して形態素解析処理を行い、所定の単位（例えば文字）ごとの素性を抽出し、素性の集合と分類ラベルとの組を生成する。
【００９２】
素性として、例えば、品詞情報（名詞、固有名詞、人名、姓、などの分類）、形態素における文字の位置情報（先頭、それ以外などの分類）、字種情報（漢字、カタカナ、英字、数字などの分類）、分類先などが抽出される。
【００９３】
言語解析処理は、機械学習手段では、素性の集合と分類ラベルの組を利用して、各単位（文字）について、その素性の集合の場合にどのような分類先になりやすいかを学習し、学習結果を学習結果記憶手段に記憶する。
【００９４】
機械学習手段は、例えば、各文字の素性と分類ラベルとの組において、文字「小」についての学習には、素性の集合を用いて行う。
【００９５】
ここで、機械学習法としては、多分類に対応できる拡張したサポートベクトルマシン法を用いる。
【００９６】
サポートベクトルマシン法は、空間を超平面で分割することにより２つの分類からなるデータを分類する手法である。このとき、２つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン）が大きいものほど、オープンデータで誤った分類をする可能性が低いと考えられ、このマージンを最大にする超平面を求め、求めた超平面を用いて分類を行う。
【００９７】
サポートベクトルマシン法の最大マージンは、ある空間で求める分離超平面と、分類超平面に平行かつ等距離にある超平面の距離（マージン）が最大になるような分離超平面を求める。
【００９８】
サポートベクトルマシン法では、通常、学習データにおいて、マージンの内部領域に小量の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線形にする拡張（カーネル関数の導入）がなされたものが用いられる。このような拡張された方法は、識別関数を用いて分類することと等価であり、その識別関数の出力値が正か負かによって、２つの分類を判別することができる。
【００９９】
なお、サポートベクトルマシンは、正例・負例の二値分類であるため、ワン・バーサス・レスト（One v.s. Rest ）法、ペア・ワイズ(Pair Wise)法などの手法を用いて二値分類を多値分類に拡張する。
【０１００】
ワン・バーサス・レスト（One v.s. Rest ）法では、例えば３つの分類先ａ、ｂ、ｃがある場合に、「ａとその他」、「ｂとその他」、「ｃとその他」という３つの組の二値分類器（ある分類先か、それ以外の分類先か）を用意し、それぞれをサポートベクトルマシ

ンで学習する。そして、解である分類先を推定する場合には、３つのサポートベクトルマシンの学習結果を利用する。推定するべき入力データが、これらの３つのサポートベクトルマシンでは、どのように推定されるかをみて、３つのサポートベクトルマシンのうち、その他でない側（正例）に分類されかつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を、求める解とする。

【０１０１】

ペア・ワイズ(Pair Wise)法では、ｋ個の分類先から任意の２つの分類先についての二値分類器を$_{k}C_{2}$個用意して、分類先同士の総当たり戦を行い、このうち最も分類先として選ばれた回数が多い分類先を求める解とする。

【０１０２】

機械学習の学習終了後、データ入力手段では、言語解析の対象のテキストデータを入力する。素性抽出手段では、教師データ作成処理と同様に、入力されたテキストデータ（入力データ）に対して形態素解析を行い、所定の単位（例えば文字）ごとの素性を抽出する。

【０１０３】

そして、解推定手段では、学習結果記憶手段に記憶された学習結果を利用して、入力データの所定の単位（文字）について、その素性の場合になりやすい分類ラベルを推定する。

【０１０４】

そして、タグ付与手段は、解と推定された分類ラベルに対応する分類タグを、入力データの該当する文字または文字列の前後に挿入する。

【０１０５】

解析結果表示処理手段では、分類タグが付加された入力データを、所定の表示規則に従った表示態様で表示装置に表示する。ここで、分類タグ＜PERSON＞＜/PERSON ＞で囲まれた文字列及び＜LOCATION＞＜/LOCATION ＞で囲まれた文字列を、特定の固有表現として抽出することができる。

【０１０６】

（９）：プログラムインストールの説明

表示装置（表示手段）１、入力装置（入力手段）２、抽出手段（抽出装置）３、疑問詞後接語抽出装置（疑問詞後接語抽出手段）４、主要語抽出装置（主要語抽出手段）５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。

【０１０７】

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】

【０１０８】

【図１】本発明のキーワード強調装置の説明図である。

【図２】本発明の疑問詞の後ろに付く単語を強調表示するフローチャートである。

【図３】本発明の疑問詞の後ろに付く単語を強調表示するフローチャートである。

【図４】本発明の数量表現を指す疑問詞を利用して強調表示するフローチャートである。

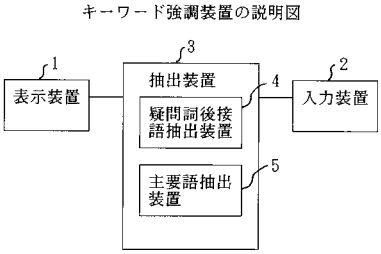
【図５】本発明の疑問詞の意味を利用して強調表示するフローチャートである。

【図６】本発明の疑問詞の意味を利用して強調表示するフローチャートである。

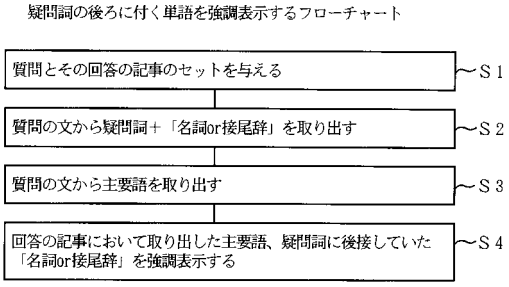
【符号の説明】

【０１０９】
１　表示装置（表示手段）
２　入力装置（入力手段）
３　抽出手段（抽出装置）
４　疑問詞後接語抽出装置（疑問詞後接語抽出手段）
５　主要語抽出装置（主要語抽出手段）

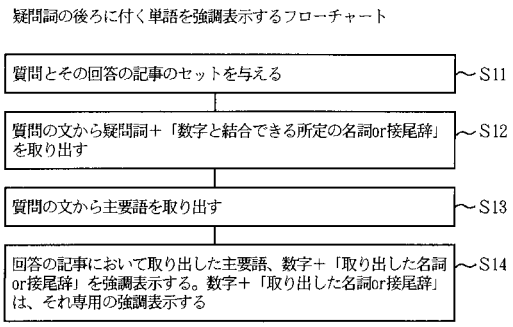
【図１】



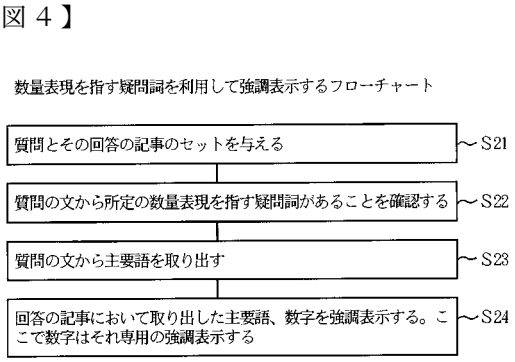
【図２】



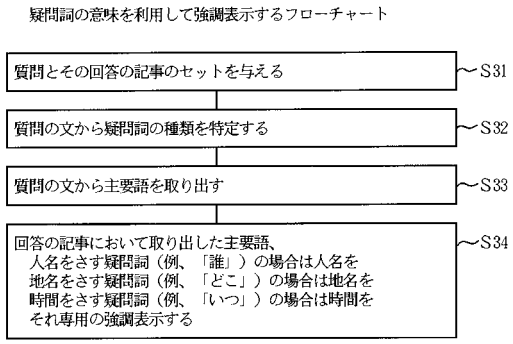
【図３】



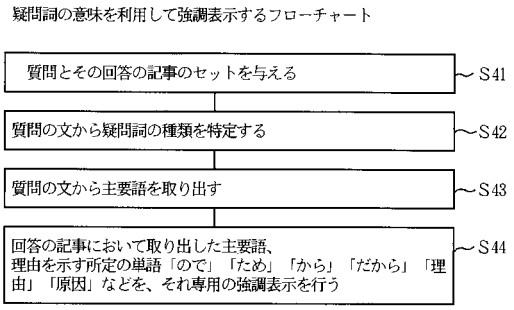
【図４】



【図５】


疑問詞の意味を利用して強調表示するフローチャート
質問とその回答の記事のセットを与える
S31
質問の文から疑問詞の種類を特定する
S32
質問の文から主要語を取り出す
S33
回答の記事において取り出した主要語、
人名をさす疑問詞（例、「誰」）の場合は人名を
地名をさす疑問詞（例、「どこ」）の場合は地名を
時間をさす疑問詞（例、「いつ」）の場合は時間を
それ専用の強調表示する
S34

【図６】


疑問詞の意味を利用して強調表示するフローチャート
質問とその回答の記事のセットを与える
S41
質問の文から疑問詞の種類を特定する
S42
質問の文から主要語を取り出す
S43
回答の記事において取り出した主要語、
理由を示す所定の単語「ので」「ため」「から」「だから」「理由」「原因」などを、それ専用の強調表示を行う
S44

フロントページの続き

(56)参考文献　特開平０６－３３２９４５（ＪＰ，Ａ）
特開２００４－２８０１７６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０－１７／２６