(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2014-500547
(P2014-500547A)
(43)公表日　平成26年1月9日(2014.1.9)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/27	(2006.01)	G06F 17/27	Z	5B091
G06F 17/30	(2006.01)	G06F 17/30	170A	

審査請求 有　予備審査請求 未請求　（全 33 頁）

(21)出願番号	特願2013-539361 (P2013-539361)
(86)(22)出願日	平成23年11月18日 (2011.11.18)
(85)翻訳文提出日	平成25年5月21日 (2013.5.21)
(86)国際出願番号	PCT/IB2011/003364
(87)国際公開番号	WO2012/095696
(87)国際公開日	平成24年7月19日 (2012.7.19)
(31)優先権主張番号	201010555763.4
(32)優先日	平成22年11月22日 (2010.11.22)
(33)優先権主張国	中国 (CN)
(31)優先権主張番号	13/298,941
(32)優先日	平成23年11月17日 (2011.11.17)
(33)優先権主張国	米国 (US)

(71)出願人　510330264
アリババ・グループ・ホールディング・リミテッド
ＡＬＩＢＡＢＡ　ＧＲＯＵＰ　ＨＯＬＤＩＮＧ　ＬＩＭＩＴＥＤ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７
(74)代理人　110000028
特許業務法人明成国際特許事務所
(74)代理人　100102989
弁理士　井上　佳知

最終頁に続く

(54)【発明の名称】複数の粒度でのテキスト分割

(57)【要約】
【解決手段】テキスト処理は、最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、中間粒度の分割結果を結合して、中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を最小意味単位の辞書内で検索し、それぞれの検索要素に基づいて、中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成することを含む。
【選択図】図３

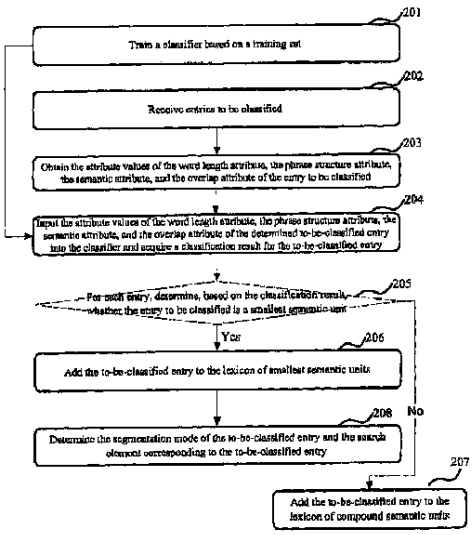

FIG. 3

【特許請求の範囲】
【請求項１】
テキスト処理の方法であって、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成すること、
を備える、方法。
【請求項２】
請求項１に記載の方法であって、さらに、
テキストを分類するための分類子を訓練し、
前記訓練は、複数の訓練サンプルエントリに基づいて行われ、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含も、
前記最小意味単位の辞書を構築し、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含むこと、
を備える、方法。
【請求項３】
請求項１に記載の方法であって、前記受信したテキストは、単語区切り記号のない言語である、方法。
【請求項４】
請求項２に記載の方法であって、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加することを備える、方法。
【請求項５】
請求項２に記載の方法であって、前記訓練された分類子を用いた前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、方法。
【請求項６】
請求項２に記載の方法であって、さらに、
前記エントリに対応する検索要素を決定し、
前記最小意味単位の辞書に前記検索要素を保存すること、

を備える、方法。
【請求項７】
請求項２に記載の方法であって、前記エントリに対応する検索要素の決定は、
前記エントリが分割可能であるか否かを判定し、
前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、
前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、を含む、方法。
【請求項８】
請求項１に記載の方法であって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、方法。
【請求項９】
テキスト処理のためのシステムであって、
１または複数のプロセッサであって、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するよう構成されている１または複数のプロセッサと、
前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、
を備える、システム。
【請求項１０】
請求項９に記載のシステムであって、前記１または複数のプロセッサは、さらに、
複数の訓練サンプルエントリに基づいて、テキストを分類するための分類子を訓練し、
前記最小意味単位の辞書を構築するよう構成され、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含む、
システム。
【請求項１１】
請求項９に記載のシステムであって、前記テキストは、単語区切り記号のない言語である、システム。

【請求項１２】
請求項１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加するよう構成されている、システム。
【請求項１３】
請求項１０に記載のシステムであって、前記訓練された分類子を用いた、前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、システム。
【請求項１４】
請求項１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、
前記エントリに対応する検索要素を決定しと、
前記最小意味単位の辞書に前記検索要素を保存するよう構成されている、システム。
【請求項１５】
請求項１０に記載のシステムであって、前記エントリに対応する検索要素の決定は、
前記エントリが分割可能であるか否かを判定し、
前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、
前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、を含む、システム。
【請求項１６】
請求項９に記載のシステムであって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、システム。
【請求項１７】
テキスト処理ためのコンピュータプログラム製品であって、前記コンピュータプログラム製品は、コンピュータ読み取り可能な記憶媒体内に具現化され、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得するためのコンピュータ命令と、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得するためのコンピュータ命令と、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索するためのコンピュータ命令と、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するためのコンピュータ命令と、
を備える、コンピュータプログラム製品。
【請求項１８】
請求項１７に記載のコンピュータプログラム製品であって、さらに、
複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練するためのコンピュータ命令と、
前記最小意味単位の辞書を構築するためのコンピュータ命令とを備え、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリと

の重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加すること、を含む、コンピュータプログラム製品。
【請求項１９】
テキスト処理のためのシステムであって、
１または複数のプロセッサであって、
複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練し、
最小意味単位の辞書を構築するよう構成されている、１または複数のプロセッサと
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含み、
前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、
を備える、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
他の出願の相互参照
本願は、すべての目的のために参照により本明細書に組み込まれる、発明の名称を「ＡＭＥＴＨＯＤ　ＯＦ　ＰＲＯＶＩＤＩＮＧ　ＭＵＬＴＩ－ＧＲＡＮＵＬＡＲＩＴＹ　ＳＥＧＭＥＮＴＡＴＩＯＮ　ＲＥＳＵＬＴＳ　ＡＮＤ　Ａ　ＤＥＶＩＣＥ　ＦＯＲ　ＳＡＭＥ（複数粒度の分割結果を提供するための方法および装置）」とする、２０１０年１１月２２日出願の中国特許出願第２０１０１０５５５７６３．４号に基づく優先権を主張する。
【０００２】
本願は、単語情報処理技術の分野に関し、特に、単語分割辞書の構築に関する。
【背景技術】
【０００３】
言語は、単語区切り記号を有するか否かに従って、２つのタイプに分類できる。一方のタイプは、英語、ドイツ語、および、多くの他のヨーロッパ言語などであり、単語区切り

記号を有する。一般に、単語間のスペースが、単語区切り記号として機能する。もう一方のタイプは、文中の単語に印をつけるための単語区切り記号を持たない。中国語、日本語、および、韓国語など、多くの東アジアの言語は、単語区切り記号のない言語である。

【０００４】

検索エンジン、機械翻訳、および、音声合成アプリケーションは、単語区切り記号のない言語のテキストを分割し、文に由来するセグメント（文節）を含むセグメント列を形成することをしばしば必要とする言語テキスト処理の問題を伴う。分割処理は、しばしば、単語分割辞書（語彙）を含んでおり、単語分割辞書は、かなりの数の予め格納されたエントリを含むデータベース／辞書を備える。単語分割の際、所与のテキストは、特定の戦略に従って単語分割辞書のエントリとマッチングされる（例えば、左から右への順方向最大マッチング法（ｆｏｒｗａｒｄ　ｍａｘｉｍｕｍ　ｍａｔｃｈｉｎｇ　ｍｅｔｈｏｄ）、右から左への逆方向最大マッチング法（ｂａｃｋｗａｒｄ　ｍａｘｉｍｕｍ　ｍａｔｃｈｉｎｇ　ｍｅｔｈｏｄ）、最小分割法など）。例えば、最大マッチング法では、入力テキストに一致しうる最長のエントリが辞書内で見つかると、それが単語として特定され、特定された単語がセグメントと見なされる。これを繰り返すと、所与のテキストを、セグメントからなるセグメント列に分割することができる。セグメントは、うまく一致した単語および文字、もしくは、動的に特定された単語を含みうる。

【０００５】

所与のテキストについて、結果として得られた単語セグメント列内のセグメントが長いほど（すなわち、セグメント列に含まれるセグメントの数が少ないほど）、単語分割粒度が大きくなる。逆に、結果として得られた単語セグメント列内のセグメントの数が多いほど、単語分割粒度が小さくなる。なお、以下では中文をＣ～Ｃの符号で記し、中文と符号との対応関係は末尾の表４に示す。例えば、所与のテキスト「Ｃ１人民共和国成立了」（中華人民共和国が建国された）について、細粒度の単語分割結果は、「Ｃ１－人民－共和国－成立－了」（中華－人民－共和国－建国－された）であり、粗粒度の単語分割結果は、「Ｃ１人民共和国－成立－了」（中華人民共和国－建国－された）である。

【０００６】

異なるアプリケーションでは、分割結果の粒度に関する要件が異なる。例えば、機械翻訳においては、粒度はいくぶん大きい方が好ましく、例えば、「Ｃ２管理」（企業管理）は、単一のセグメントであることが好ましい。しかし、検索エンジンの索引システムでは、「Ｃ２管理」は、一般に２つのセグメントに分割される（「Ｃ２」（企業）および「管理」（管理））。

【０００７】

分割結果に関する粒度の要件は、同じタイプのアプリケーションでも異なりうる。以下では、説明の目的で、検索エンジンアプリケーションの例を用いる。検索エンジンアプリケーションにおいて、検索エンジンは、異なる分野に対して異なる単語分割粒度を必要とする。例えば、電子商取引の分野で（例えば、商品検索を行うために）用いられる検索エンジンでは、販売者および購入者の両者が、検索において高い再現率を求める。これを達成するために、検索システムは、より小さい索引の粒度を有する必要があるため、より細粒度の分割結果を必要とする。一般的なウェブページ検索に用いられる検索エンジンでは、莫大な数のインターネットウェブページがあることから、検索の精度が、ユーザにとって特に重要になる。これを達成するために、検索システムは、より粗粒度の分割結果を必要とする。このように、検索の再現率および検索の精度が、検索の質を評価するための重要な尺度になる。検索の再現率は、システムが関連情報をどれだけうまく見つけるかを測るものであり、見いだされた関連文書の、関連文書の総数に対する比である。検索の精度は、システムが関連情報を見つける際にどれだけうまく実行するかを測るものであり、見いだされた関連文書の、見出された全文書に対する比である。単語分割粒度は、検索再現率および検索精度に関係する。一般的に言うと、単語分割粒度が小さいほど、検索再現率が高くなり、単語分割粒度が大きいほど、検索精度が高くなる。

【０００８】

分割結果に関する粒度の要件は、同じタイプのアプリケーションの同じ分野内でも様々な使用段階に応じて異なっている。再び、説明のためにウェブ検索エンジンアプリケーションを一例として用いる。検索再現率および検索精度の両方に関するユーザの要求を満たすために、粒度要件は、検索の索引段階および順序付け段階の間で異なる。索引段階では、十分な数のウェブページが検索されうるように、より細粒度の分割結果が必要とされる。順序付け段階では、検索精度への要求を満たし、関連のないウェブページをユーザに提供することを避けるために、より粗粒度の分割結果が必要とされる。

【０００９】

上述の問題を解決するために、従来技術は、主に、２つのスキームを用いて、複数の粒度の分割結果を提供する。

【００１０】

図１Ａは、複数の粒度で分割結果を提供するための典型的なスキームを示している。まず、最小粒度の単語分割が実行される。次いで、下から上への動的な結合が行われる。具体的には、より細粒度の単語分割辞書Ａを用いて、所与のテキストに単語分割を実行する。異なるセグメント列が、単語分割処理で生成されうる。例えば、テキスト$S_1S_2S_3S_4S_5S_6S_7$（ここで、S_nは文字を表す）は、$S_1S_2－S_3S_4－S_5－S_6S_7$または$S_1S_2S_3－S_4S_5－S_6S_7$に分割されうる。次いで、セグメント列の１つ（ここでは、$S_1S_2－S_3S_4－S_5－S_6S_7$とする）が、所定の選択アルゴリズムに従って、最適セグメント列として選択されうる。所定のアルゴリズムは、統計モデルに基づいたアルゴリズムであってよい。

【００１１】

より粗粒度の分割結果を提供するために、列$S_1S_2－S_3S_4－S_5－S_6S_7$に結合が実行される。具体的な結合処理は、列$S_1S_2－S_3S_4－S_5－S_6S_7$内の２つのセグメントの組み合わせが、より長いエントリを含む単語分割辞書Ｂ内のエントリと一致するか否かを評価することを必要とする。これら２つのセグメントが結合されると、より粗粒度の結合済みセグメント列が生じる。ここで、S_1S_2およびS_3S_4を結合することができ、S_5およびS_6S_7を結合することができると仮定すると、より粗粒度の結合済みセグメント列は、$S_1S_2S_3S_4－S_5S_6S_7$となる。

【００１２】

この方法を用いる場合、いくつかの意味項目が、単語分割中に失われる。例えば、意味要素$S_1S_2S_3$およびS_4S_5が失われる。説明のために、ここで実際の例を用いる。テキストを「本Ｃ３管用Ｃ５Ｃ６Ｃ７」（このステンレス鋼管は、１級鋼を用いて鋳造されている）とする。ここで、「Ｃ３管」（ステンレス鋼管）は、実際は、２つの意味項目を含む。すなわち、「Ｃ３」（ステンレス鋼）および「「ム「オ管」（鋼管）である。最小粒度の「Ｃ３管」（ステンレス鋼管）を「Ｃ３－管」（ステンレス鋼－管）（ここで「－」は２つの隣接するセグメントを分離する記号）に分割した後に、再びこれらを結合して「Ｃ３管」（ステンレス鋼管）を形成した場合、意味項目「Ｃ６管」（鋼管）が失われる。その結果として、「「ム「オ管」（鋼管）という用語は、このテキストの検索中に見つからなくなる。最小粒度の「Ｃ３管」（ステンレス鋼管）を「Ｃ４管」（ステン－レス－鋼管）に分割した後に、再びこれらを結合して「Ｃ３管」（ステンレス鋼管）を形成した場合、意味項目「Ｃ３」（ステンレス鋼）が失われる。したがって、「Ｃ３」（ステンレス鋼）は、このテキストの検索中に見つからなくなる。

【００１３】

さらに、結合の精度を保証するのは困難である。所与のテキストの最小粒度単語分割から得られたセグメント列が、「本－Ｃ３－管－用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼－管－１級－鋼－鋳造）であるとすると、結合の際に曖昧性が生じる。結合された結果は、「Ｃ３管」（ステンレス鋼管）または「管用」（有用）でありうる。所与のテキストの最小粒度単語分割から得られたセグメント列が、「本－Ｃ３－管用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼－有用－１級－鋼－鋳造）であった場合、再び結合しても、意味項目「Ｃ３管」（ステンレス鋼管）を得ることはできない。

【００１４】
図１Ｂは、複数の粒度で分割結果を提供するための別の典型的なスキームを示している。まず、最大粒度の単語分割が実行される。次いで、上から下への分割が実行される。具体的には、より粗粒度の単語分割辞書Ｃが用いられ、モデルおよびアルゴリズムが、所与のテキスト$S_1S_2S_3S_4S_5S_6S_7$の動的な単語分割を実行（最適なセグメント列を選択）してセグメント列$S_1S_2S_3S_4－S_5S_6S_7$を得るために用いられる。
【００１５】
より細粒度の単語分割結果を得るために、$S_1S_2S_3S_4－S_5S_6S_7$内の各意味要素が、再び分割される。具体的な分割処理は、列$S_1S_2S_3S_4－S_5S_6S_7$内の各セグメントを評価して、単語分割辞書Ｃ内の２以上のその他のより細粒度のエントリを含むか否かを判定する。含む場合、このセグメントは、２以上のその他のエントリに分割される。$S_1S_2S_3S_4$がS_1S_2およびS_3S_4に分割され、$S_5S_6S_7$がS_5およびS_6S_7に分割されうると仮定すると、分割後に得られる細粒度の単語分割結果は、$S_1S_2－S_3S_4－S_5－S_6S_7$となる。
【００１６】
この方法を用いる場合、最大粒度の単語分割中に生じる曖昧性の問題を解決するために、より多くの粗粒度のエントリが辞書に記録される必要がある。例えば、「Ｃ２管理科学Ｃ８」（企業管理科学技術）というテキストがあるとすると、より粗粒度のエントリ「Ｃ２管理」（企業管理）および「管理科学」（管理科学）が辞書に記録されている場合、「Ｃ２管理科学」（企業管理科学）は、「Ｃ２管理－科学」（企業管理－科学）または「Ｃ２－管理科学」（企業－管理科学）に分割されうる。この曖昧性の解決方法は、さらに長いエントリ「Ｃ２管理科学」（企業管理科学）も辞書に記録することである。しかしながら、「Ｃ２管理科学」（企業管理科学）は、「科学Ｃ８」（科学技術）に関する分割の曖昧性も生じる。したがって、かかる粗粒度のエントリで構成された集合は、閉集合ではない。辞書を拡大すると、辞書の維持が困難になる。
【００１７】
以上のように、単語分割辞書内のエントリの粒度が大きくなるほど、単語分割中に生成される異なるセグメント列の数が多くなる。すなわち、より多い単語分割経路があるため、曖昧性の問題も多くなる。最大粒度分割の精度を保証することが困難になる。
【００１８】
最大粒度の分割結果がある時、辞書をチェックすることによって、これらのセグメントの細粒度の単語を取得できる。しかしながら、辞書が拡大すると、エントリの質を維持しつつ、これらのエントリと、これらのエントリの細粒度の単語とを手作業で維持するのは、高コストになりうる。
【００１９】
要約すると、複数の粒度で分割結果を提供するための従来技術には、通例、再現率が低いことにより意味項目が失われるという問題、または、単語分割辞書が非常に膨大であり単語分割処理の精度が低いという問題がある。
【図面の簡単な説明】
【００２０】
以下の詳細な説明と添付の図面において、本発明の様々な実施形態を開示する。
【００２１】
【図１Ａ】複数の粒度で分割結果を提供するための典型的なスキームを示す図。
【００２２】
【図１Ｂ】複数の粒度で分割結果を提供するための別の典型的なスキームを示す図。
【００２３】
【図２】テキストを分割し、複数の粒度の分割結果を提供するためのシステムの一実施形態を示す図。
【００２４】
【図３】単語分割辞書、特に、最小意味単位の辞書を構築するための処理の一実施形態を

示すフローチャート。
【００２５】
【図４】最小意味単位の辞書および複合意味単位の辞書に基づいて、複数の粒度の分割結果を取得する処理の一実施形態を示すフローチャート。
【００２６】
【図５】単語分割ツリー構造の一例を示す図。
【００２７】
【図６】単語分割辞書を構築するためのシステムの一実施形態を示すブロック図。
【００２８】
【図７】複数の粒度の分割結果を提供するよう構成されたシステムの一実施形態を示すブロック図。
【００２９】
【図８】単語分割処理モジュールの一実施形態を示すブロック図。
【００３０】
【図９】決定モジュールの一実施形態を示すブロック図。
【発明を実施するための形態】
【００３１】
本発明は、処理、装置、システム、物質の組成、コンピュータ読み取り可能な格納媒体上に具現化されたコンピュータプログラム製品、および／または、プロセッサ（プロセッサに接続されたメモリに格納および／またはそのメモリによって提供される命令を実行するよう構成されたプロセッサ）を含め、様々な形態で実装されうる。本明細書では、これらの実装または本発明が取りうる任意の他の形態を、技術と呼ぶ。一般に、開示された処理の工程の順序は、本発明の範囲内で変更されてもよい。特に言及しない限り、タスクを実行するよう構成されるものとして記載されたプロセッサまたはメモリなどの構成要素は、ある時間にタスクを実行するよう一時的に構成された一般的な構成要素として、または、タスクを実行するよう製造された特定の構成要素として実装されてよい。本明細書では、「プロセッサ」という用語は、１または複数のデバイス、回路、および／または、コンピュータプログラム命令などのデータを処理するよう構成された処理コアを指すものとする。
【００３２】
以下では、本発明の原理を示す図面を参照しつつ、本発明の１または複数の実施形態の詳細な説明を行う。本発明は、かかる実施形態に関連して説明されているが、どの実施形態にも限定されない。本発明の範囲は、特許請求の範囲によってのみ限定されるものであり、多くの代替物、変形物、および、等価物を含む。以下の説明では、本発明の完全な理解を提供するために、多くの具体的な詳細事項が記載されている。これらの詳細事項は、例示を目的としたものであり、本発明は、これらの具体的な詳細事項の一部または全てがなくとも特許請求の範囲に従って実施可能である。簡単のために、本発明に関連する技術分野で周知の技術要素については、本発明が必要以上にわかりにくくならないように、詳細には説明していない。
【００３３】
複数の粒度のセグメントにテキストを分割することが開示されている。いくつかの実施形態において、単語分割辞書（例えば、最小意味単位の辞書）が構築される。最小意味単位の辞書内のエントリは、合理的な長さを有し、意味的な完全性も有する。さらに、それらのエントリに対応する検索要素が辞書に格納される。所与のエントリが単語分割を受ける時、中間粒度の分割結果を得るために、構築された最小意味単位の辞書に基づいて単語分割を受ける。中間粒度の分割結果は、より粗粒度のエントリを含む単語分割辞書を用いて結合され、それによって、より粗粒度の分割結果が得られる。最小意味単位の辞書に格納されたエントリに対応する検索要素を用いて、より細粒度の分割結果が、中間粒度の分割結果に基づいて取得される。いくつかの実施形態において、テキストは、中国語など、単語区切り記号のない言語である。

【００３４】
図２は、テキストを分割し、複数の粒度の分割結果を提供するためのシステムの一実施形態を示す。明らかに、フォームデザインのためのコンテキスト依存のスクリプト編集を実行するために、他のコンピュータシステムアーキテクチャおよび構成が用いられてもよい。以下に述べるような様々なサブシステムを備えるコンピュータシステム１００は、少なくとも１つのマイクロプロセッササブシステム（プロセッサまたは中央処理装置（ＣＰＵ）とも呼ばれる）１０２を備える。例えば、プロセッサ１０２は、シングルチッププロセッサまたはマルチプロセッサによって実装できる。いくつかの実施形態において、プロセッサ１０２は、コンピュータシステム１００の動作を制御する汎用デジタルプロセッサである。メモリ１１０から読み出された命令を用いて、プロセッサ１０２は、入力データの受信および操作、ならびに、出力デバイス（例えば、ディスプレイ１１８）上でのデータの出力および表示を制御する。いくつかの実施形態において、プロセッサ１０２は、テキストを分割し、複数の粒度の分割結果を提供することを含む、および／または、そのために用いられる。

【００３５】
プロセッサ１０２は、メモリ１１０と双方向的に接続されており、メモリ１１０は、第１のプライマリストレージ（通例は、ランダムアクセスメモリ（ＲＡＭ））および第２のプライマリストレージ領域（通例は、読み出し専用メモリ（ＲＯＭ））を含みうる。当業者に周知のように、プライマリストレージは、一般的な記憶領域として、および、スクラッチパッドメモリとして利用可能であり、また、入力データおよび処理済みデータを格納するために利用可能である。プライマリストレージは、さらに、プロセッサ１０２上で実行される処理のための他のデータおよび命令に加えて、データオブジェクトおよびテキストオブジェクトの形態で、プログラミング命令およびデータを格納できる。また、当業者に周知のように、プライマリストレージは、通例、機能（例えば、プログラムされた命令）を実行するためにプロセッサ１０２によって用いられる基本的な動作命令、プログラムコード、データ、および、オブジェクトを備える。例えば、メモリ１１０は、例えば、データアクセスが双方向である必要があるか、単方向である必要があるかに応じて、後述する任意の適切なコンピュータ読み取り可能な記憶媒体を含みうる。例えば、プロセッサ１０２は、頻繁に必要になるデータをキャッシュメモリ（図示せず）に直接的かつ非常に迅速に格納し取り出すことができる。

【００３６】
着脱可能なマスストレージデバイス１１２が、コンピュータシステム１００にさらなるデータ記憶容量を提供しており、プロセッサ１０２に対して双方向（読み出し／書き込み）または単方向（読み出しのみ）に接続されている。例えば、ストレージ１１２は、磁気テープ、フラッシュメモリ、ＰＣカード、携帯型マスストレージデバイス、ホログラフィックストレージデバイス、および、その他のストレージデバイスなどのコンピュータ読み取り可能な媒体も含みうる。固定マスストレージ１２０も、例えば、さらなるデータ記憶容量を提供しうる。マスストレージ１２０の最も一般的な例は、ハードディスクドライブである。マスストレージ１１２および１２０は、一般に、プロセッサ１０２によって通例はあまり利用されないさらなるプログラミング命令、データなどを格納する。当然のことながら、マスストレージ１１２および１２０に保持された情報は、必要であれば、仮想メモリとしてのメモリ１１０（例えば、ＲＡＭ）の一部に標準的な方式で組み込まれてよい。

【００３７】
プロセッサ１０２がストレージサブシステムにアクセスできるようにすることに加えて、バス１１４は、その他のサブシステムおよびデバイスへのアクセスを可能にするために用いられてもよい。図に示すように、これらは、ディスプレイモニタ１１８、ネットワークインターフェース１１６、キーボード１０４、および、ポインティングデバイス１０６、ならびに、必要に応じて、補助入力／出力デバイスインターフェース、サウンドカード、スピーカ、および、その他のサブシステムを含みうる。例えば、ポインティングデバイ

ス１０６は、マウス、スタイラス、トラックボール、または、タブレットであってよく、グラフィカルユーザインターフェースと相互作用するのに有用である。

【００３８】

ネットワークインターフェース１１６は、図に示すように、ネットワーク接続を用いて、別のコンピュータ、コンピュータネットワーク、または、遠隔通信ネットワークにプロセッサ１０２を接続することを可能にする。例えば、ネットワークインターフェース１１６を通して、プロセッサ１０２は、方法／処理ステップを実行する過程で、別のネットワークから情報（例えば、データオブジェクトまたはプログラム命令）を受信したり、別のネットワークに情報を出力したりすることができる。情報は、プロセッサ上で実行される一連の命令として表されることが多く、別のネットワークから受信されたり、別のネットワークへ出力されたりしうる。インターフェースカード（または同様のデバイス）と、プロセッサ１０２によって実装（例えば、実行／実施）される適切なソフトウェアとを用いて、コンピュータシステム１００を外部ネットワークに接続し、標準プロトコルに従ってデータを転送することができる。例えば、本明細書に開示された様々な処理の実施形態は、プロセッサ１０２上で実行されてもよいし、処理の一部を共有するリモートプロセッサと共に、ネットワーク（インターネット、イントラネットワーク、または、ローカルエリアネットワークなど）上で実行されてもよい。さらなるマスストレージデバイス（図示せず）が、ネットワークインターフェース１１６を通してプロセッサ１０２に接続されてもよい。

【００３９】

補助Ｉ／Ｏデバイスインターフェース（図示せず）が、コンピュータシステム１００と共に用いられてよい。補助Ｉ／Ｏデバイスインターフェースは、プロセッサ１０２がデータを送信すること、ならびに、より典型的には、他のデバイス（マイクロホン、タッチセンサ方式ディスプレイ、トランスデューサカードリーダ、テープリーダ、音声または手書き認識装置、バイオメトリクスリーダ、カメラ、携帯型マスストレージデバイス、および、他のコンピュータなど）からデータを受信することを可能にする汎用インターフェースおよびカスタマイズされたインターフェースを含みうる。

【００４０】

さらに、本明細書に開示された様々な実施形態は、さらに、様々なコンピュータ実装された動作を実行するためのプログラムコードを備えたコンピュータ読み取り可能な媒体を含むコンピュータストレージ製品に関する。コンピュータ読み取り可能な媒体は、データを格納できる任意のデータストレージデバイスであり、そのデータは、後にコンピュータシステムによって読み出されうる。コンピュータ読み取り可能な媒体の例は、以下の媒体すべてを含むがそれらに限定されない。ハードディスク、フロッピーディスク、および、磁気テープなどの磁気媒体、ＣＤ－ＲＯＭディスクなどの光学媒体、光学ディスクなどの磁気光学媒体、ならびに、特定用途向け集積回路（ＡＳＩＣ）、プログラム可能論理デバイス（ＰＬＤ）、および、ＲＯＭ／ＲＡＭデバイスなど、特別に構成されたハードウェアデバイス。プログラムコードの例としては、例えば、コンパイラによって生成されるマシンコード、または、インタープリタを用いて実行できる高水準コード（例えば、スクリプト）を含むファイルが挙げられる。

【００４１】

図２に示したコンピュータシステムは、本明細書に開示された様々な実施形態と共に利用するのに適切なコンピュータシステムの一例にすぎない。かかる利用に適した他のコンピュータシステムは、より多いまたは少ないサブシステムを含みうる。さらに、バス１１４は、サブシステムをリンクさせるよう機能する任意の相互接続スキームの例である。異なる構成のサブシステムを有する他のコンピュータアーキテクチャが利用されてもよい。

【００４２】

図３は、単語分割辞書、特に、最小意味単位の辞書を構築するための処理の一実施形態を示すフローチャートである。処理２００は、システム（１００など）上で実行されてよい。

【００４３】
工程２０１では、訓練セットに基づいて訓練される分類子が取得される。いくつかの実施形態において、訓練セットは、多くのサンプルエントリを含んでおり、訓練セット内の各訓練サンプルエントリは、単語長属性、単独利用率、句構造規則属性、意味属性、重複属性、および、分類結果を含む。
【００４４】
単語長属性の属性値は、訓練サンプルエントリ内のテキストの文字数を含む。
【００４５】
句構造属性の値は、訓練サンプルエントリの細粒度の単語の単独利用率値と、訓練サンプルエントリが句構造規則に従っているか否かを示すインジケータとを含む。
【００４６】
句の単独利用率の値は、（例えば、ログエントリ、アンカーテキストなどから取得された単独句など）単独句のセット内での出現頻度または出現回数を単位として測られてよい。
【００４７】
いくつかの実施形態において、システムは、様々なカテゴリの列挙エントリの所定のセット（例えば、ＴＶ番組、本のタイトル、商品ブランドなどの列挙エントリのセット）を提供する。意味属性の値は、サンプルエントリが列挙エントリのセットに含まれるか否かに依存する。言い換えると、意味属性値は、列挙エントリのセットにおける訓練サンプルエントリの包含状態を示す。訓練サンプルエントリが列挙エントリの所定のセットに含まれる場合、意味属性の値は、列挙エントリの対応する所定のセットのための識別子である。訓練サンプルエントリが、列挙エントリの所定のセットのいずれにも見いだされない場合、意味属性の値には、列挙エントリの任意の所定のセットの識別子とは異なる識別子が割り当てられる。
【００４８】
重複属性の値は、訓練サンプルエントリが様々なカテゴリの列挙エントリのいずれか内の別のエントリと重複する確率と、重複部分が細粒度の単語であるか否かを示すインジケータとを含む。
【００４９】
分類結果は、予め格付けされた訓練サンプルエントリが、複合意味単位であるか最小意味単位であるかを示すインジケータを含む。本明細書で用いられているように、複合意味単位とは、意味論的に意味を持つ（例えば、人間に理解可能な）部分にさらに分割できるテキストの一部のことであり、最小意味単位とは、意味を持つ部分にさらに分割できないテキストのことである。
【００５０】
例えば、訓練サンプルエントリ「Ｃ２管理」（企業管理）は、４文字である。したがって、この訓練サンプルエントリの単語長の値（すなわち、文字数）は４である。訓練サンプルエントリ「Ｃ２管理」（企業管理）は、細粒度の単語「Ｃ２」（企業）および「管理」（管理）を含む。これら２つの細粒度の単語が単独利用エントリのセットに現れる率が決定され、最も高い率が、訓練サンプルエントリ「Ｃ２管理」（企業管理）の句構造属性の値に対する細粒度の単語の単独利用率として機能するよう選択される。本明細書で用いられているように、単独利用エントリセットは、インターネットクエリログ、アンカーテキスト、または、任意の他の適切な技術によって取得されてよい。例えば、単独利用エントリセットを構築するための収集段階中に、ユーザが検索キーワード「Ｃ２」（企業）をインターネット検索エンジンに入力した場合、「Ｃ２」（企業）はクエリログに記録され、「Ｃ２」（企業）が単独で利用されうることが示される。さらに、カンマまたはスペースなどの区切りマーカによって分離されたユーザによって入力された各単語は、単独で利用された単語と見なすことができる。１，０００，０００件の事例で、細粒度の単語「管理」（管理）が単独利用エントリセット内で最も頻繁に出現すると仮定する。一方で、エントリが独立した細粒度の単語を欠いている（例えば、その単語が他の単語から独立して

検索エンジンに入力されたことがない）場合、率は０である。

【００５１】

句構造規則は、所与の言語の構文を記述する方法である。規則は、一般的に、自然言語に関する幅広い研究を通して得られる。中国語の場合、句は、一般に、「形容詞＋名詞」、「名詞＋名詞」、または、「動詞＋名詞」で構成される。句構造規則は、正規表現の形態で格納されうる。訓練サンプルエントリ「Ｃ２管理」（企業管理）は、２つの細粒度の名詞で構成されている。「Ｃ２」（企業）および「管理」（管理）である。したがって、訓練サンプルエントリ「Ｃ２管理」（企業管理）は、句構造規則に従っている。句構造規則に従っていることを示すインジケータが１に設定され、句構造規則に従っていないことを示すインジケータが０であると仮定する。したがって、訓練サンプルエントリ「信息Ｃ９工程」（情報システム工学）の句構造属性の値は、（１，０００，０００，１）である。

【００５２】

いくつかの実施形態において、システムは、様々なカテゴリの列挙エントリの所定のセット（例えば、映画のタイトル、本のタイトル、商品ブランドなどの列挙エントリのセット）を提供する。意味属性の値は、サンプルエントリが列挙エントリのセットに含まれるか否かに依存する。例えば、ＴＶ／映画のタイトルのセットのための識別子をＳ２１とする。映画のタイトルのカテゴリに含まれるエントリは、Ｓ２１＝｛ゴッドファーザー，シュレック，甲方乙方．．．｝である。小説のタイトルのセットのための識別子はＳ２２である。小説のタイトルのセットに含まれるエントリは、Ｓ２２＝｛ホビット，二都物語，紅楼夢，．．．｝である。教科書の題材のセットのための識別子はＳ２３である。教科書の題材のセットに含まれるエントリは、Ｓ２３＝｛情報工学，心理学，哲学，企業管理，産業，および，商業管理，．．．｝である。都市名のセットのための識別子はＳ２４であり、エントリは、Ｓ２４＝｛北京，上海，ニューヨーク，フフホト，．．．｝を含む。訓練サンプルエントリ「Ｃ２管理」（企業管理）は、教科書の題材のセットに含まれる。したがって、訓練サンプルエントリ「Ｃ２管理」（企業管理）に対応する識別子は、Ｓ２３である。訓練サンプルエントリがいずれのタイプの列挙エントリセットにも含まれない場合、この訓練サンプルエントリの句構造値は、どのタイプの列挙エントリセットのための識別子とも異なる識別子、例えば、どの列挙エントリセットにも対応しないＳ２０になる。

【００５３】

重複属性の値を決定するために、訓練サンプル単語が、辞書に含まれる別のエントリと、訓練テキスト内で重複する確率が計算される。本明細書で用いられているように、重複とは、訓練サンプルエントリ内のいくつかの文字が、訓練サンプルエントリを含む訓練テキスト内で前または後ろに位置するいくつかの文字と組み合わされた時に、辞書内の別のエントリを形成する状況を指す。例えば、訓練サンプルエントリが「Ｃ２管理」（企業管理）であり、訓練テキストが「・・・Ｃ１０，Ｃ２管理科学Ｃ１１・・・」（周知の通り、企業管理科学は新たな主題である・・・）を含むと仮定する。ここで、「Ｃ２管理」（企業管理）および「管理科学」（管理科学）は、重複するテキスト「管理」（管理）を有する。２つの単語が重複する時、重複部分は、この例における「管理」（管理）のように、意味論的に意味を持つ粒度の細かい単語でありうる。一部の例では、重複は、単一の文字であってもよい。例えば、「甲方乙方」（Ｆｉｒｓｔ Ｐａｒｔｙ Ｓｅｃｏｎｄ Ｐａｒｔｙ（中国映画））および「方才」（たった今）は、文脈を拡張された訓練テキスト「・・・甲方乙方才上映・・・」（・・・Ｆｉｒｓｔ Ｐａｒｔｙ Ｓｅｃｏｎｄ Ｐａｒｔｙ、たった今劇場公開・・・）内で重複する。「甲方乙方」（Ｆｉｒｓｔ Ｐａｒｔｙ Ｓｅｃｏｎｄ Ｐａｒｔｙ）における細粒度の単語は、「甲方／乙方」（Ｆｉｒｓｔ Ｐａｒｔｙ／Ｓｅｃｏｎｄ Ｐａｒｔｙ）であり、重複部分は、文字「方」（「当事者」または「ちょうど」、文脈による）であり、「甲方乙方」（Ｆｉｒｓｔ Ｐａｒｔｙ Ｓｅｃｏｎｄ Ｐａｒｔｙ）の意味論的に意味を持つ細粒度の単語ではない。したがって、訓練サンプルエントリが訓練テキストに出現した時にその訓練サンプルエントリが辞書

内の別のエントリと重複する確率が計算される。重複部分が細粒度の単語である場合、対応するインジケータは１に設定されてよく、そうでない場合、インジケータは０に設定されてよい。この実施形態において、訓練サンプルエントリ「Ｃ２管理」（企業管理）が他のエントリと重複する確率が２％であり、単語「管理」（管理）と重複する部分が、粒度の細かい単語であると仮定する。この場合、訓練サンプルエントリ「Ｃ２管理」（企業管理）の重複値は、（２％，１）である。

【００５４】

この例において、訓練サンプルエントリ「Ｃ２管理」（企業管理）は、複合意味要素に分類される。ここで、訓練サンプルエントリ「Ｃ２管理」（企業管理）、「フフホト」などの値および分類結果を、表１に示す。

【００５５】

【表１】

表１　訓練セット内の訓練サンプルエントリの値および格付け結果の例

訓練サンプルエントリ	単語長属性値	句構造属性値	意味属性値	重複属性値	分類結果
「企业管理」（企業管理）	4	(1,000,000, 1)	S23	(2%, 1)	複合意味単位
「呼和浩特」（フフホト）	4	(-1, 0)	S24	(0.001%, 0)	最小意味単位
・・・・	・・・	・・・	・・・	・・・	・・・

【００５６】

表１に示された訓練セット内のすべての訓練サンプルエントリの属性値および格付け分類結果に基づいて、ＧＢＤＴ（勾配ブースト決定木）、最大エントロピ、サポートベクターマシン（ＳＶＭ）、または、分類子を訓練するための任意のその他の最適な技術などの機械学習技術が用いられてよい。本実施形態における分類子は、訓練セット内のエントリの単語長属性値、句構造属性値、意味属性値、および、重複属性値に少なくとも部分的に基づいて確立される。分類子は、分類されるエントリが、複合意味単位であるか最小意味単位であるかを判定するために用いられる。一般に、より大きい単語長属性値と、第１の要素の値が比較的高く、従来の句構造規則に従うエントリに適合する句構造属性とを有するエントリは、分類子によって複合意味単位であると判定される可能性が高い。小さい単語長属性値と、第１の要素の値が比較的低く、従来の句構造規則に従うエントリに適合しない句構造属性とを有するエントリは、分類子によって最小意味単位であると判定される可能性が高い。

【００５７】

分類子の訓練に用いられる機械学習技術は、当業者に周知である。例えば、決定木学習技術は、ソースセットを属性値テストに基づいてサブセットに分割し、再帰的に各派生サブセットに処理を繰り返すことによって、入力変数（例えば、文字数、単独利用率、句構造規則値、意味属性値、および、重複属性値）に基づいて目標変数（例えば、分類結果）の値を予測するためのモデルを構築する。サポートベクターマシン技術は、Ｎ個のクラスの１つに属するものとして訓練セットエントリをマークし、所与の各入力値について、その入力値を含む可能性のあるクラスを予測するモデルを構築する。

【００５８】

工程２０２では、分類対象のエントリが受信される。エントリは、既存の単語分割辞書

、データベース、インターネットなど、様々なソースから受信されうる。

【００５９】

この例において、分類対象のエントリは、「五大Ｃ１２」（五大連池、中国の観光地）、「菊花茶」（菊花茶）、および、「Ｃ３管」（ステンレス鋼管）である。

【００６０】

工程２０３では、分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値が取得される。

【００６１】

分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値を決定する処理は、訓練セット内の訓練サンプルエントリについて上述の４つの属性の属性値を決定するために工程２０１において用いられたアプローチと同様である。分類対象のエントリの例の属性値情報を表２に示す。

【００６２】

【表２】

表２　分類対象のエントリの属性値の例

分割対象のエントリ	単語長属性値	句構造属性値	意味属性値	重複属性値
「五大连池」 （五大連池）	4	(9, 1)	0	(0.01%, 1)
「菊花茶」 （菊花茶）	3	(21, 1)	0	(2%, 1)
「不锈钢管」 （ステンレス鋼管）	4	(11, 1)	0	(1%, 1)
「笔记本电脑包」 （ノートブックコンピュータバッグ）	6	(35, 1)	0	(4%, 1)
「迷你轿车」 （ミニセダン）	4	(66, 1)	0	(5%, 1)
・・・・	・・・	・・・	・・・	・・・

【００６３】

工程２０４では、分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値が、分類対象のエントリの分類結果を取得するために、分類子に入力される。

【００６４】

この例における「五大Ｃ１２」（五大連池）、「菊花茶」（菊花茶）、および、「Ｃ３管」（ステンレス鋼管）の分類結果は、最小意味単位に対応する。「Ｃ１３」（ノートブックコンピュータバッグ）および「Ｃ１４」（ミニセダン）の分類結果は、複合意味単位に対応する。複合意味単位の粒度は、最小意味単位の粒度よりも大きい。

【００６５】

工程２０５では、分類対象の各エントリの分類結果に基づいて、エントリが最小意味単

位か否かが判定される。エントリが最小意味単位でない場合、処理は工程２０７に進み、ここで、分類対象のエントリは複合意味単位の辞書に追加される。しかしながら、エントリが最小意味単位である場合、処理は工程２０６に進み、ここで、分類対象のエントリは最小意味単位の辞書に追加される。

【００６６】

この例では、「五大Ｃ１２」（五大連池）、「菊花茶」（菊花茶）、および、「Ｃ３管」（ステンレス鋼管）が最小意味単位の辞書に追加される。「Ｃ１３」（ノートブックコンピュータバッグ）および「Ｃ１４」（ミニセダン）は、複合意味単位の辞書に追加される。

【００６７】

最小意味単位の辞書および複合意味単位の辞書は、いくつかの実施形態において、細粒度単語のセグメント列（例えば、最小意味単位を含む列）を、より粗粒度の単語のセグメント列に変換するために用いられる。例えば、テキストは、最初に、最小意味単位の辞書に従って、最小意味単位を含むセグメント列に分割される（換言すると、列内のセグメントは、最小意味単位の辞書に見いだされる）。複合意味単位の辞書に基づいて、この最初の列内のセグメントは、複合意味単位の辞書に見いだされる粗粒度のセグメントを形成するように結合される。

【００６８】

工程２０８では、分類対象のエントリの分割モードと、エントリに対応する検索要素が、決定されて格納される。分割モードは２つのタイプを含む。分割可能および分割不可能である。エントリが分割可能であるか分割不可能であるかは、以下の２つの基準に基づいて判定される。（１）固有名詞であるか？これは、固有名詞データベースでエントリを検索することによって判定できる。固有名詞である場合、さらなる分割は不可能であり、分割不可能となる。「五大Ｃ１２」（五大連池）が一例である。（２）意味論的にさらに分割できるか？「黄金周」（ゴールデンウィーク）または「大哥大」（ダゲダ、携帯電話の中国語の俗語）のように、エントリが定型表現（例えば、全体として具体的な意味を持つ表現）である場合、分割不可能である。エントリが分割可能であるか否かは、固有名詞および定型表現の所定のデータベース内でエントリを検索することによって判定できる。固有名詞でも定型表現でもないエントリは、分割可能である。例えば、「Ｃ１５」（保湿化粧水）および「菊花茶」（菊花茶）は、固有名詞でも定型表現でもないので、分割可能である。

【００６９】

分類対象のエントリの分割モードが分割可能である場合、分類対象のエントリに対応する検索要素は、分類対象のエントリ内に含まれる細粒度の単語である。分類対象のエントリの分割モードが分割不可能である場合、分類対象のエントリに対応する検索要素は、分類対象のエントリ自体である。

【００７０】

最小意味単位の辞書内のエントリのデータ構造例を表３に示す。

【表３】

エントリ	分割モード	検索要素
「五大连池」 （五大連池）	分割不可能	「五大连池」 （五大連池）
「菊花茶」 （菊花茶）	分割可能	菊花（菊）、花茶（花茶）、 茶（茶）
「不锈钢管」 （ステンレス鋼管）	分割可能	不锈钢（ステンレス鋼）、 鋼管（鋼管）
大哥大（ダゲダ）	分割不可能	大哥大（ダゲダ）
黄金周 （ゴールデンウィーク）	分割不可能	黄金周 （ゴールデンウィーク）
润肤乳（保湿化粧水）	分割可能	润肤（保湿）、乳（化粧水）
・・・・	・・・	・・・

表３　最小意味単位の辞書内のエントリのデータ構造例

【００７１】
処理２００は、既存の単語分割辞書内のエントリ（または、他の手段によって取得されたエントリ）を取得して、最小意味単位の辞書または複合意味単位の辞書のために分類するものとしても理解されうる。

【００７２】
上述のスキームによって決定された最小意味単位の辞書が含むエントリは、一般に、既存の粗粒度の単語分割辞書よりも短くて数が少ないため、単語分割辞書に基づく分割に必要な時間と、単語分割の曖昧性の可能性が低減される。したがって、単語分割処理の精度が向上し、辞書の維持の困難さが減少する。

【００７３】
図４は、最小意味単位の辞書および複合意味単位の辞書に基づいて、複数の粒度の分割結果を取得する処理の一実施形態を示すフローチャートである。

【００７４】
工程３０１では、受信されたテキストに対して分割が実行される。分割は、所与のテキストについて最小意味単位の辞書に基づいて実行される。単語分割を通して取得されたセグメント列は、分割結果の中間セットと見なされる（中間粒度の分割結果ともいう）。分割された列の中のセグメントは、最小意味単位を含む。次いで、処理は、同時に工程３０２および工程３０３に進む。

【００７５】
いくつかの実施形態において、所与のテキストは、最小意味単位の辞書内のエントリとマッチングされ、既存の曖昧性除去モデルを用いて、生じうる任意の分割の曖昧性問題が解決される。例えば、所与のテキストが「本Ｃ３管用Ｃ５Ｃ６Ｃ７」（このステンレス鋼管は１級鋼を用いて鋳造されている）であり、辞書クエリが、最小意味単位の辞書に基づいて左から右へと実行されるとする。セグメント内の文字の最大数が６であると仮定すると、所与のテキスト「本Ｃ３管用」（このステンレス鋼管用いる）の最も左にある文字か

ら始まる最初の６文字が、最小意味単位の辞書に見いだされうるか否かを判定するために評価される。言い換えると、６文字の処理窓が、処理対象の文字列を抽出するために用いられる。それらの文字が最小意味単位の辞書に見いだされた場合、これらの６文字からなるセグメントは、第１の単語分割列に記録される。見いだされなかった場合、最も右側の文字が削除され、残りの５文字「本Ｃ３管」（このステンレス鋼管）を含むセグメントが再び比較される。この処理は、すべての文字が処理されるか、最小意味単位が見いだされるまで、残りの文字に対して繰り返される。この例では、文字列「本Ｃ３管用」（このステンレス鋼管用いる）について、マッチする最小意味単位が見いだされない。

【００７６】

すべての６文字が処理された後、窓は１文字分移動し、処理は次の６文字「Ｃ３管用一」（ステンレス鋼管用いる１）について繰り返される。それらの文字が最小意味単位の辞書に見いだされるか否かを判定するために、評価が行われる。見いだされた場合、これら６文字のセグメントは、第１の単語分割列に記録される。見いだされなかった場合、最も右側の文字が削除され、残りの５文字「Ｃ３管用」（ステンレス鋼管用いる）が再び比較され、この処理が残りの文字に対して繰り返される。この反復において、最小意味単位「Ｃ３」（ステンレス鋼）が特定される。

【００７７】

窓の移動および処理の反復を繰り返すことによって、所与のテキストに含まれるすべての最小意味単位が取得される。特定された最小意味単位は、複数のセグメント列（複数の単語分割経路）を構成して、曖昧性を生じうる。いくつかの実施形態において、単語分割の曖昧性が生じた場合、複数の列から１つのセグメント列が、条件付き確率場（ＣＲＦ）モデル、隠れマルコフモデル（ＨＭＭ）、最大エントロピ（ＭＥ）モデルなどの曖昧性除去モデルに基づいて選択される。当業者に周知のように、これらの曖昧性除去モデルは、統計的情報学習を用いたコーパス分析に基づいており、ここで、単語分割は様々な文脈素性に従って実行される。セグメント列「本－Ｃ３管－用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼管－用いる－１級－鋼－鋳造）が得られるまで、このように処理が続けられる。

【００７８】

最小粒度の単語分割から取得されたセグメント列は、「本－Ｃ３－管－用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼－管－用いる－１級－鋼－鋳造）である。既存の下から上への動的結合スキームは、意味項目「Ｃ６管」（鋼管）を含まない「本－Ｃ３－管用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼－有用－１級－鋼－鋳造）を生成しうる。したがって、分割の曖昧性があり、後の検索の際に精度が低くなる。さらに、失われた用語があるため、「Ｃ６管」（鋼管）に関連する文書が見つからず、検索の再現率も減少する。対照的に、本願は、最小の単語粒度ではなく最小意味単位に基づいて単語分割を実行することにより、分割の曖昧性の可能性を低減し、上述の問題をより効果的に解決する。

【００７９】

別の例として、所与のテキストが、「Ｃ２管理科学Ｃ８」（企業管理科学技術）であるとする。最小意味単位の辞書に従って単語分割から取得されるセグメント列は、「Ｃ２－管理－科学－Ｃ８」（企業－管理－科学－技術）である。既存の下から上への動的結合スキーム（例えば、既存の最大粒度単語分割スキーム）に従った場合、最大粒度単語分割を行った際に、「Ｃ２管理」（企業管理）および「管理科学」（管理科学）の間、「管理科学」（管理科学）および「科学Ｃ８」（科学技術）の間、ならびに、「管理」（管理）および「理科」（科学）の間に、分割の曖昧性の問題が生じる。最大粒度に基づいた単語分割は、大量のエントリを必要とし、その結果、大量の不必要な分割の曖昧性が生じ、分割の精度が下がる。最小意味単位の中のエントリの方が、連結修正語（ｃｏｍｂｉｎａｔｉｏｎ　ｍｏｄｉｆｉｅｒ）を有する可能性が低い。したがって、最小意味単位に基づいた単語分割は、分割の精度を高めうる。

【００８０】

工程３０２では、最小意味単位よりも大きい粒度を有する単語分割辞書（例えば、処理

２００を用いて得られた複合意味単位の辞書）に基づいて、曖昧性除去モデルを用いて、中間粒度の分割結果内のセグメントが結合され、第１の粒度の分割結果が取得される。第１の粒度の分割結果は、中間粒度の分割結果よりも粗い粒度（すなわち、大きい粒度）であることから、粗粒度の分割結果とも呼ばれる。

【００８１】

「Ｃ２－管理－科学－Ｃ８」（企業－管理－科学－技術）という中間粒度の分割結果を例として、より大きい粒度を有する単語分割辞書がエントリ「Ｃ２管理」（管理科学）および「科学Ｃ８」（科学技術）を含むと仮定する。したがって、列「Ｃ２－管理－科学－Ｃ８」（企業－管理－科学－技術）内のセグメントは、より粗粒度のセグメントに結合されて、「Ｃ２管理－科学Ｃ８」（企業管理－科学技術）という結合後のより粗粒度の分割結果を形成しうる。

【００８２】

工程３０３では、単語分割ツリー構造が任意選択的に形成される。ここで、所与のテキストが、ルートノードを形成するために用いられ、工程３０１において取得された中間粒度の分割結果内の各セグメントが、ルートノードのサブノードを形成するために用いられる。左から右へ順番に、セグメントに対応する各ノードがルートノードに追加される。図５は、単語分割ツリー構造の一例を示す図である。この例では、ノード５５２がルートノード（所与のテキスト）に対応し、ノード５５４がサブノード（中間の粒度の分割結果）に対応する。

【００８３】

図４に戻ると、工程３０４では、中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素が、最小意味単位の辞書内で検索される。処理２００に関連して上述したように、セグメントおよびそれらに対応する検索要素は、最小意味単位の辞書に格納される。一例として表３を参照すると、所与のテキスト「本Ｃ３管用Ｃ５Ｃ６Ｃ７」（このステンレス鋼管は１級鋼を用いて鋳造されている）について、中間粒度の分割結果は「本－Ｃ３管－用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼管－用いる－１級－鋼－鋳造）である。例えば、セグメント「Ｃ３管」（ステンレス鋼管）に対応する検索要素は、「Ｃ３」（ステンレス鋼）および「Ｃ６管」（鋼管）である。

【００８４】

工程３０５では、エントリに対応する検索要素は、単語分割結果ツリー内のリーフノードを形成するために用いられる。図５のツリーの例に示すように、リーフノードはノード５５６である。

【００８５】

工程３０６では、第２の粒度の分割結果が、検索要素に基づいて取得される。第２の粒度の分割結果は、中間粒度の分割結果よりも細かい粒度であることから、細粒度の分割結果とも呼ばれる。いくつかの実施形態において、単語分割結果ツリー内のリーフノードは、第２の粒度の分割結果と見なされる。図５のツリーの例を参照すると、所与のテキスト「本Ｃ３Ｃ６管用Ｃ５Ｃ６Ｃ７」（このステンレス鋼管は１級鋼を用いて鋳造されている）について取得されるより細粒度の分割結果は、「本－Ｃ３－Ｃ６管－用－Ｃ５－Ｃ６－Ｃ７」（この－ステンレス鋼－鋼管－用いる－１級－鋼－鋳造）である。

【００８６】

処理３００は、処理２００で構築された最小意味単位の辞書を用いて、所与のテキストに単語分割を実行し、中間粒度の分割結果を取得する。次いで、中間粒度の分割結果よりも大きい粒度を有する第１の粒度の分割結果を取得するために、最小意味単位の辞書よりも大きい粒度を有する辞書に従って、中間粒度の分割結果に対して結合が行われる。また、中間粒度の分割結果よりも粒度の細かい第２の粒度の分割結果は、最小意味単位に格納された各エントリに対応する検索要素と、中間粒度の分割結果とに基づいて取得される。このように、所与のテキストに対応する少なくとも３つの粒度の分割結果を提供することが可能であり、単語分割の粒度に関して様々なタイプのアプリケーションが要求する異なる要件を満たすことができる。したがって、従来技術の問題、すなわち、従来技術が複数

の粒度の分割結果を提供した時に（意味項目が失われた結果として）再現率が低くなる問題および単語分割の精度が低くなる問題を回避することができる。

【００８７】

図６は、単語分割辞書を構築するためのシステムの一実施形態を示すブロック図である。システムは、処理２００を実行するよう構成される。この例において、システム５００は、分類子取得モジュール５０１、インターフェースモジュール５０２、属性値決定モジュール５０３、分類結果決定モジュール５０４、および、第１のエントリ追加モジュール５０５を備える。

【００８８】

分類子取得モジュール５０１は、訓練セットに基づいて分類子を訓練するよう構成されており、訓練セット内の各訓練サンプルエントリは、上述のように、単語長属性、句構造属性、意味属性、重複属性、および、分類結果を有する。

【００８９】

インターフェースモジュール５０２は、分類対象のエントリを受信するよう構成されている。インターフェースモジュールの例としては、ポート、ケーブル、有線または無線ネットワークインターフェースカードなどの外部接続、および、通信バスなどの内部接続が挙げられるが、これらに限定されない。

【００９０】

属性値決定モジュール５０３は、分類対象エントリ取得モジュール５０２によって取得された分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値を決定するよう構成されている。

【００９１】

分類結果決定モジュール５０４は、分類子取得モジュール５０１によって取得された分類子と、属性値決定モジュール５０３によって決定された分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値とに基づいて、分類対象のエントリを決定し、分類対象のエントリが最小意味単位であるか否かを判定するよう構成されている。

【００９２】

第１のエントリ追加モジュール５０５は、分類対象のエントリが分類結果決定モジュール５０４によって最小意味単位であると判定された場合に、最小意味単位の辞書に分類対象のエントリを追加するよう構成されている。

【００９３】

図５のデバイスは、さらに、分類対象のエントリが最小意味単位でないと第２の決定モジュール５０４によって判定された場合に、複合意味単位の辞書に分類対象のエントリを追加するよう構成された第２のエントリ追加モジュール５０６を備えることが好ましい。

【００９４】

デバイスは、さらに、第１のエントリ追加モジュール５０５が分類対象のエントリを最小意味単位の辞書に追加した後に、分類対象のエントリの分割モードと、分類対象のエントリに対応する検索要素とを、最小意味単位の辞書に格納するための検索要素格納モジュール５０７を備えることが好ましい。

【００９５】

図７は、複数の粒度の分割結果を提供するよう構成されたシステムの一実施形態を示すブロック図である。システム６００は、単語分割辞書構築モジュール６０１、単語分割処理モジュール６０２、結合モジュール６０３、検索モジュール６０４、および、決定モジュール６０５を備える。システムは、処理３００を実行するよう構成されている。

【００９６】

単語分割辞書構築モジュール６０１は、最小意味単位の辞書を構築するよう構成されている。

【００９７】

単語分割処理モジュール６０２は、辞書構築モジュール６０１によって構築された最小

意味単位の辞書に従って所与のテキストに単語分割を実行し、中間粒度の分割結果を取得するよう構成されている。

【００９８】

結合モジュール６０３は、最小意味単位の辞書よりも大きい粒度の辞書と、曖昧性除去モデルとに基づいて、単語分割処理モジュール６０２によって取得された中間粒度の分割結果を結合し、より粗粒度の分割結果を取得するよう構成されている。

【００９９】

検索モジュール６０４は、単語分割処理モジュール６０２によって取得された中間粒度の分割結果内の列に含まれる各セグメントに対応する検索要素を、（単語分割辞書構築モジュール６０１によって構築された）最小意味単位の辞書内で検索するよう構成されている。

【０１００】

決定モジュール６０５は、より細粒度の分割結果を決定するよう構成されている。

【０１０１】

図８は、単語分割処理モジュールの一実施形態を示すブロック図である。この例において、単語分割処理モジュール７００（例えば、図７の６０２）は、単語分割サブモジュール７０１、第１の決定サブモジュール７０２、および、第２の決定サブモジュール７０３を備える。

【０１０２】

単語分割サブモジュール７０１は、単語分割辞書構築モジュール６０１によって構築された最小意味単位の辞書に基づいて所与のテキストに単語分割を実行するよう構成されている。

【０１０３】

第１の決定サブモジュール７０２は、単語分割サブモジュール７０１による単語分割によって取得されたセグメント列が１つだけであった場合に、中間粒度の単語分割結果としてセグメント列を形成するよう構成されている。

【０１０４】

第２の決定サブモジュール７０３は、単語分割サブモジュール７０１が複数のセグメント列を生成した場合に、曖昧性除去モデルに基づいて、中間の粒度の単語分割結果として１つのセグメント列を選択するよう構成されている。

【０１０５】

図９は、決定モジュールの一実施形態を示すブロック図である。この例において、決定モジュール８００（例えば、図７の６０５）は、分割結果ツリー構築サブモジュール８０１および決定サブモジュール８０２を備える。

【０１０６】

分割結果ツリー構築サブモジュール８０１は、分割結果のツリーを形成するよう構成されている。いくつかの実施形態において、このモジュールは、所与のテキストを用いてルートノードを形成し、中間粒度の分割結果内の各セグメントを用いてルートノードのサブノードを形成し、セグメントに対応するノードのリーフノードとしてセグメントに対応する検索要素を用いる。

【０１０７】

決定サブモジュール８０２は、分割結果ツリー構築サブモジュール８０１によって構築された分割結果ツリーにおける各リーフノードを順番に取得し、順番に取得されたリーフノードを、より細粒度の単語分割結果と見なすよう構成されている。

【０１０８】

上述のモジュールは、１または複数の汎用プロセッサ上で実行されるソフトウェアコンポーネントとして、特定の機能を実行するよう設計されたプログラム可能論理デバイスおよび／または特定用途向け集積回路などのハードウェアとして、もしくは、それらの組み合わせとして実装することができる。いくつかの実施形態において、モジュールは、コンピュータデバイス（パーソナルコンピュータ、サーバ、ネットワーク装置など）に本発明

の実施形態に記載された方法を実行させるための複数の命令など、不揮発性記憶媒体（光学ディスク、フラッシュ記憶装置、携帯用ハードディスクなど）に格納することができるソフトウェア製品の形態で具現化されてよい。モジュールは、単一のデバイス上に実装されてもよいし、複数のデバイスにわたって分散されてもよい。モジュールの機能は、互いに統合されてもよいし、複数のサブモジュールにさらに分割されてもよい。

【０１０９】

当業者であれば、適切なハードウェアにプログラムから命令させることによって、上述の実施形態の実現に関与する工程の全部または一部を実現できることを理解できる。このプログラムは、ＲＯＭ／ＲＡＭ、磁気ディスク、光学ディスクなどの読み取り可能な記憶媒体に格納できる。

【０１１０】

明らかに、当業者は、本発明の精神および範囲から逸脱することなく、本願を変形および変更することができる。したがって、本願のこれらの変形例および変更例が、特許請求の範囲および等価の技術の範囲内にある場合、本願は、これらの変形例および変更例をも網羅するものである。

【０１１１】

上述の実施形態は、理解しやすいようにいくぶん詳しく説明されているが、本発明は、提供された詳細事項に限定されるものではない。本発明を実施する多くの代替方法が存在する。開示された実施形態は、例示であり、限定を意図するものではない。

【０１１２】

【表４】

Ｃ１：中华

Ｃ２：企业

Ｃ３：不锈钢

Ｃ４：不一锈一钢

Ｃ５：一级

Ｃ６：钢

Ｃ７：铸造

Ｃ８：技术

Ｃ９：系统

Ｃ１０：众所周知

Ｃ１１：是一门新兴的学科

Ｃ１２：连池

Ｃ１３：笔记本电脑包

Ｃ１４：迷你轿车

Ｃ１５：润肤乳

【図１Ａ】

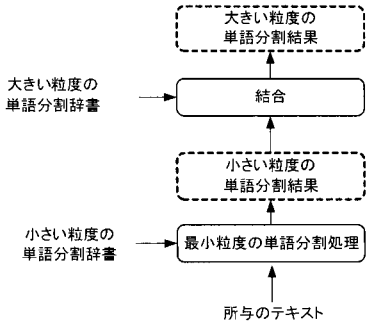

FIG. 1A

【図１Ｂ】

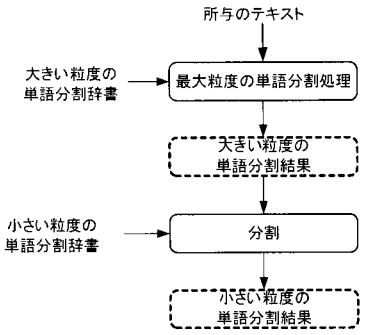

FIG. 1B

【図２】

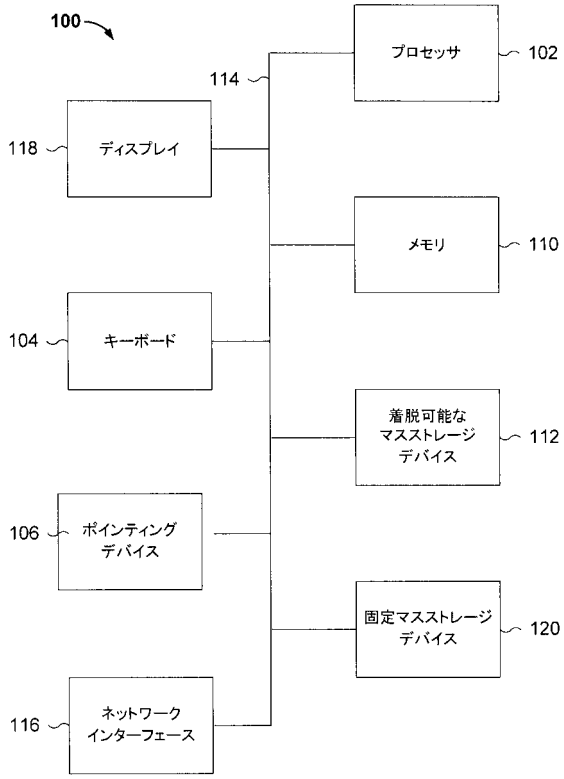

FIG. 2

【図３】

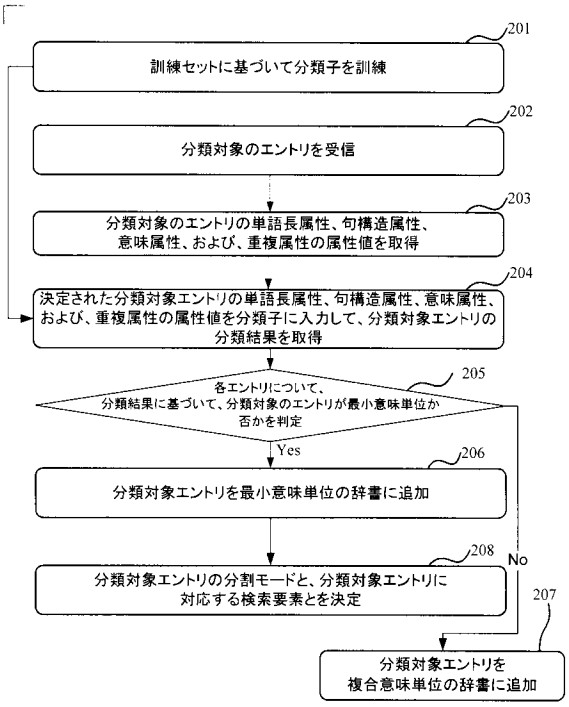

FIG. 3

【図４】

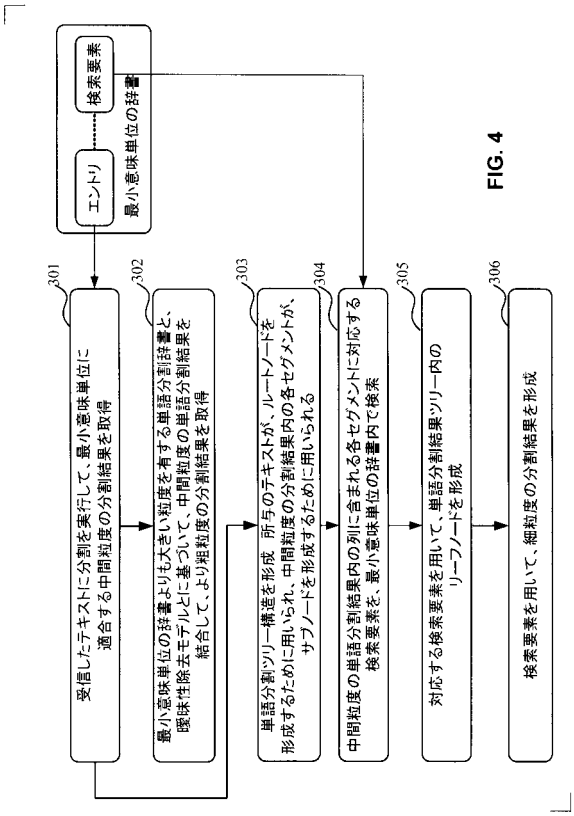

FIG. 4

【図５】

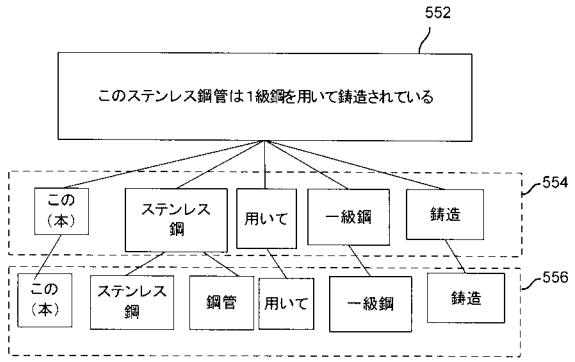

FIG. 5

【図６】

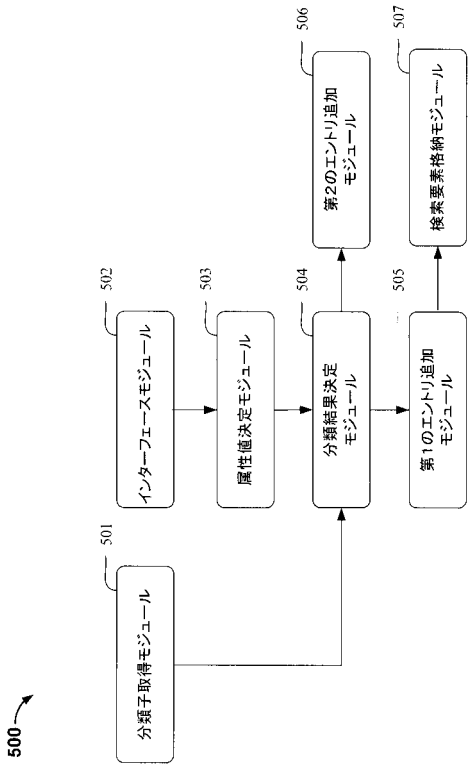

FIG. 6

【図７】

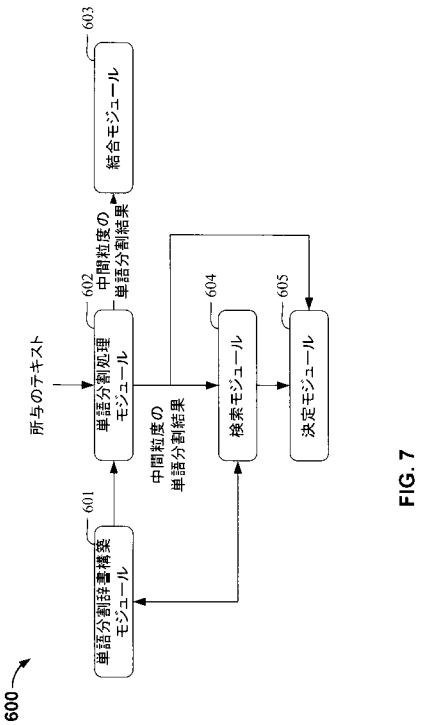

FIG. 7

【図８】

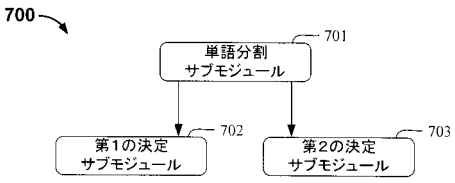

FIG. 8

【図９】

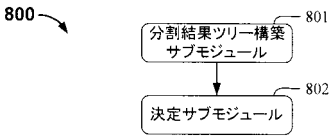

FIG. 9

【手続補正書】
【提出日】平成25年9月27日(2013.9.27)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
テキスト処理の方法であって、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成すること、
を備える、方法。
【請求項２】
請求項１に記載の方法であって、さらに、
テキストを分類するための分類子を訓練し、
前記訓練は、複数の訓練サンプルエントリに基づいて行われ、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
前記最小意味単位の辞書を構築し、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含むこと、
を備える、方法。
【請求項３】
請求項１に記載の方法であって、前記受信したテキストは、単語区切り記号のない言語である、方法。
【請求項４】
請求項２に記載の方法であって、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加することを備える、方法。
【請求項５】
請求項２に記載の方法であって、前記訓練された分類子を用いた前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの

単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、方法。
【請求項６】
請求項２に記載の方法であって、さらに、
前記エントリに対応する検索要素を決定し、
前記最小意味単位の辞書に前記検索要素を保存すること、
を備える、方法。
【請求項７】
請求項６に記載の方法であって、前記エントリに対応する検索要素の決定は、
前記エントリが分割可能であるか否かを判定し、
前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、
前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、
を含む、方法。
【請求項８】
請求項１に記載の方法であって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、方法。
【請求項９】
テキスト処理のためのシステムであって、
１または複数のプロセッサであって、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するよう構成されている１または複数のプロセッサと、
前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、
を備える、システム。
【請求項１０】
請求項９に記載のシステムであって、前記１または複数のプロセッサは、さらに、
複数の訓練サンプルエントリに基づいて、テキストを分類するための分類子を訓練し、
前記最小意味単位の辞書を構築するよう構成され、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複

合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含む、
システム。
【請求項１１】
請求項９に記載のシステムであって、前記テキストは、単語区切り記号のない言語である、システム。
【請求項１２】
請求項１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加するよう構成されている、システム。
【請求項１３】
請求項１０に記載のシステムであって、前記訓練された分類子を用いた、前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、システム。
【請求項１４】
請求項１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、
前記エントリに対応する検索要素を決定し、
前記最小意味単位の辞書に前記検索要素を保存するよう構成されている、システム。
【請求項１５】
請求項１４に記載のシステムであって、前記エントリに対応する検索要素の決定は、
前記エントリが分割可能であるか否かを判定し、
前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、
前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、を含む、システム。
【請求項１６】
請求項９に記載のシステムであって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、システム。
【請求項１７】
テキスト処理ためのコンピュータプログラムであって、
最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得するための機能と、
前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得するための機能と、
前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索するための機能と、
前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するための機能と、
をコンピュータによって実現させるコンピュータプログラム。
【請求項１８】
請求項１７に記載のコンピュータプログラムであって、さらに、
複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練するための機能と、
前記最小意味単位の辞書を構築するための機能とをコンピュータによって実現させ、
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、

文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
前記最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加すること、を含む、コンピュータプログラム。
【請求項１９】
テキスト処理のためのシステムであって、
１または複数のプロセッサであって、
複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練し、
最小意味単位の辞書を構築するよう構成されている、１または複数のプロセッサと
前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、
文字数と、
単独利用率と、
前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、
列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、
前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、
前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、
最小意味単位の辞書の構築は、
分類対象のエントリを受信し、
前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、
前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含み、
前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、
を備える、システム。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００８１
【補正方法】変更
【補正の内容】
【００８１】
「Ｃ２－管理－科学－Ｃ８」（企業－管理－科学－技術）という中間粒度の分割結果を例として、より大きい粒度を有する単語分割辞書がエントリ「Ｃ２管理」（企業管理）および「科学Ｃ８」（科学技術）を含むと仮定する。したがって、列「Ｃ２－管理－科学－Ｃ８」（企業－管理－科学－技術）内のセグメントは、より粗粒度のセグメントに結合さ

れて、「Ｃ２管理－科学Ｃ８」（企業管理－科学技術）という結合後のより粗粒度の分割結果を形成しうる。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００９０
【補正方法】変更
【補正の内容】
【００９０】
属性値決定モジュール５０３は、インターフェースモジュール５０２によって取得された分類対象のエントリの単語長属性、句構造属性、意味属性、および、重複属性の属性値を決定するよう構成されている。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】００９３
【補正方法】変更
【補正の内容】
【００９３】
図６のデバイスは、さらに、分類対象のエントリが最小意味単位でないと分類結果決定モジュール５０４によって判定された場合に、複合意味単位の辞書に分類対象のエントリを追加するよう構成された第２のエントリ追加モジュール５０６を備えることが好ましい。
【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】０１１１
【補正方法】変更
【補正の内容】
【０１１１】
上述の実施形態は、理解しやすいようにいくぶん詳しく説明されているが、本発明は、提供された詳細事項に限定されるものではない。本発明を実施する多くの代替方法が存在する。開示された実施形態は、例示であり、限定を意図するものではない。
適用例１：テキスト処理の方法であって、最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成すること、を備える、方法。
適用例２：適用例１に記載の方法であって、さらに、テキストを分類するための分類子を訓練し、前記訓練は、複数の訓練サンプルエントリに基づいて行われ、前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、文字数と、単独利用率と、前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、前記最小意味単位の辞書を構築し、前記最小意味単位の辞書の構築は、分類対象のエントリを受信し、前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含むこと、を備える、方法。
適用例３：適用例１に記載の方法であって、前記受信したテキストは、単語区切り記号のない言語である、方法。

適用例４：適用例２に記載の方法であって、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加することを備える、方法。

適用例５：適用例２に記載の方法であって、前記訓練された分類子を用いた前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、方法。

適用例６：適用例２に記載の方法であって、さらに、前記エントリに対応する検索要素を決定し、前記最小意味単位の辞書に前記検索要素を保存すること、を備える、方法。

適用例７：適用例２に記載の方法であって、前記エントリに対応する検索要素の決定は、前記エントリが分割可能であるか否かを判定し、前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、を含む、方法。

適用例８：適用例１に記載の方法であって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、方法。

適用例９：　テキスト処理のためのシステムであって、１または複数のプロセッサであって、最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得し、前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得し、前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索し、前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するよう構成されている１または複数のプロセッサと、前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、を備える、システム。

適用例１０：適用例９に記載のシステムであって、前記１または複数のプロセッサは、さらに、複数の訓練サンプルエントリに基づいて、テキストを分類するための分類子を訓練し、前記最小意味単位の辞書を構築するよう構成され、前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、文字数と、単独利用率と、前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、前記最小意味単位の辞書の構築は、分類対象のエントリを受信し、前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含む、システム。

適用例１１：適用例９に記載のシステムであって、前記テキストは、単語区切り記号のない言語である、システム。

適用例１２：適用例１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、前記エントリが複合意味単位であると判定された場合に、複合意味単位の辞書に前記エントリを追加するよう構成されている、システム。

適用例１３：適用例１０に記載のシステムであって、前記訓練された分類子を用いた、前記エントリが最小意味単位であるか複合意味単位であるかの判定は、前記エントリの文字数、前記エントリの単独利用率、前記エントリが句構造規則に従うか否かを示す句構造規則インジケータ、前記列挙エントリの所定のセットにおける前記エントリの包含状態を示す意味属性、および、前記エントリの重複属性を、前記訓練された分類子に入力することを含む、システム。

適用例１４：適用例１０に記載のシステムであって、前記１または複数のプロセッサは、さらに、前記エントリに対応する検索要素を決定しと、前記最小意味単位の辞書に前記検索要素を保存するよう構成されている、システム。

適用例１５：適用例１０に記載のシステムであって、前記エントリに対応する検索要素の決定は、前記エントリが分割可能であるか否かを判定し、前記エントリが分割可能である場合、前記エントリに含まれる細粒度の単語に前記検索要素を設定し、前記エントリが分割不可能である場合、前記エントリに前記検索要素を設定すること、を含む、システム。

適用例１６：適用例９に記載のシステムであって、前記最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得することは、前記中間粒度の分割結果の曖昧性を解決することを含む、システム。

適用例１７：　テキスト処理ためのコンピュータプログラム製品であって、前記コンピュータプログラム製品は、コンピュータ読み取り可能な記憶媒体内に具現化され、最小意味単位の辞書に基づいて、受信したテキストを分割して、中間粒度の分割結果を取得するためのコンピュータ命令と、前記中間粒度の分割結果を結合して、前記中間粒度の分割結果よりも粗い粒度を有する粗粒度の分割結果を取得するためのコンピュータ命令と、前記中間粒度の分割結果内のセグメントに対応するそれぞれの検索要素を前記最小意味単位の辞書内で検索するためのコンピュータ命令と、前記それぞれの検索要素に基づいて、前記中間粒度の分割結果よりも細かい粒度を有する細粒度の分割結果を形成するためのコンピュータ命令と、を備える、コンピュータプログラム製品。

適用例１８：適用例１７に記載のコンピュータプログラム製品であって、さらに、複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練するためのコンピュータ命令と、前記最小意味単位の辞書を構築するためのコンピュータ命令とを備え、前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、文字数と、単独利用率と、前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、前記最小意味単位の辞書の構築は、分類対象のエントリを受信し、前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加すること、を含む、コンピュータプログラム製品。

適用例１９：　テキスト処理のためのシステムであって、１または複数のプロセッサであって、複数の訓練サンプルエントリに基づいて行われる、テキストを分類するための分類子を訓練し、最小意味単位の辞書を構築するよう構成されている、１または複数のプロセッサと、前記複数の訓練サンプルエントリ内の訓練サンプルエントリは、文字数と、単独利用率と、前記訓練サンプルエントリが句構造規則に従うか否かを示す句構造規則値と、列挙エントリの所定のセットにおける前記訓練サンプルエントリの包含状態を示す意味属性値と、前記列挙エントリの所定のセット内の別のエントリと前記訓練サンプルエントリとの重複を示す重複属性値と、前記訓練サンプルエントリが複合意味単位であるか最小意味単位であるかを示す分類結果と、を含み、最小意味単位の辞書の構築は、分類対象のエントリを受信し、前記訓練された分類子を用いて、分類対象の前記エントリが最小意味単位であるか複合意味単位であるかを判定し、前記エントリが最小意味単位であると判定された場合に、前記最小意味単位の辞書に前記エントリを追加することを含み、前記１または複数のプロセッサに接続され、前記１または複数のプロセッサに命令を提供するよう構成されている１または複数のメモリと、を備える、システム。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB11/03364

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27, 17/21 (2012.01)
USPC - 704/8, 9, 10
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8): G06F 17/28, 17/20, 17/21, 17/27, 13/00 (2012.01)
USPC: 704/8, 1, 9, 10; 707/100

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-C,B, DE-A, DE-T, DE-U, GB-A, FR-A); DialogPRO; IEEE/IEEEXplore; Google/Google Scholar; IP.com; seach, query, phrase, clause, sentence, class, category, train, learn, segment, partition, lexical, dictionary, Chinese, Japanese, Far Eastern*, Korea, agglutinative, ideogram, ideographic

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X – Y	US 5,029,084 A (MOROHASI, M.) figures 1, 5, 9, column 2, lines 61-68, column 3, lines 47-62, column 4 lines 42-49, column 4 lines 57-59, column 7, lines 33-40, column 8 line 66 through column 9, line 15, column 9, lines 23-50, column 12, lines 7-13, column 12, lines 29-51	1, 3, 8, 9, 11, 16, 17 2, 4-7, 10, 12-15, 18, 19
Y	US 2005/0197829 A1 (OKUMURA, K.) figures 2, 3, 5, paragraphs [0006], [0007], [0032]-[0035], [0039], [0040], [0044], [0045]	2, 5-7, 10, 13-15, 18, 19
Y	US 2009/0327313 A1 (KUNG, Y.) et al, abstract, figure 1, paragraphs [0014], [0015]	4, 12

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 August 2012 (20.08.2012)	11 SEP 2012
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN

(72)発明者 スン・ジエン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 ホウ・レイ
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 ターン・ジーン ミーン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 チュウ・ミン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 リヤオ・シャオ リーン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 シュイ・ビーン ジーン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 プオン・レン ゴーン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内
(72)発明者 ヤーン・ヤーン
中華人民共和国 ハンチョウ, ワーナー・ロード, ザ・ウエスト・レイク・インターナショナル・プラザ・オブ・エス アンド ティー, ビルディング エー, １０階, ナンバー３９１, アリババ・グループ・リーガル・デパートメント内

Ｆターム(参考) 5B091 AA15 AB11 CA02 CA05 CA12 EA01