(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-96245
(P2011-96245A)
(43)公開日　平成23年5月12日(2011.5.12)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/22	(2006.01)	G06F 17/22	514T	5B091
G06F 17/27	(2006.01)	G06F 17/27	E	5B109

審査請求 未請求　請求項の数 16　ＯＬ　（全 31 頁）

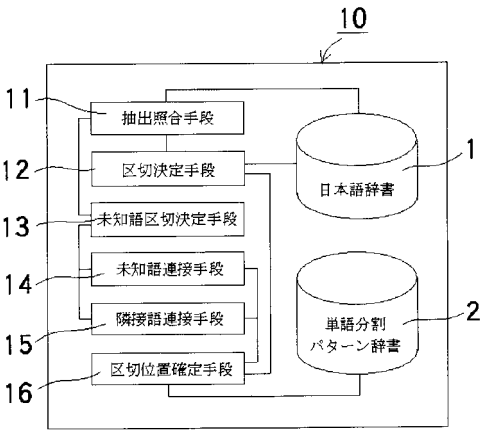
(21)出願番号　特願2010-222057 (P2010-222057)
(22)出願日　平成22年9月30日 (2010.9.30)
(31)優先権主張番号　特願2009-228800 (P2009-228800)
(32)優先日　平成21年9月30日 (2009.9.30)
(33)優先権主張国　日本国 (JP)

(71)出願人　592218300
学校法人神奈川大学
神奈川県横浜市神奈川区六角橋３丁目２７番１号
(74)代理人　100131679
弁理士　▲高▼橋　幸夫
(72)発明者　後藤　智範
神奈川県横浜市栄区庄戸５－２０－１３
(72)発明者　梅木　定博
神奈川県鎌倉市笛田４－８－５

Ｆターム(参考)　5B091 AA15　AB11　CA02
5B109 MB16　QA02

(54)【発明の名称】漢字複合語分割方法及び漢字複合語分割装置

(57)【要約】
【課題】　日本語文書に含まれる連続する漢字列で構成された漢字複合語を超高精度で正しく分割することができ、分割した各漢字列の信頼性が実用化することができる程度まで高められた、漢字複合語分割方法及び漢字複合語分割装置を提供する。
【解決手段】　連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と基本単語に該当する品詞を関連付けて記録した日本語辞書と、漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、漢字複合語の字数毎に分類して記録した単語分割パターン辞書とを参照して、分割対象の漢字複合語を分割する漢字複合語分割方法などにより課題を解決した。
【選択図】　図１



【特許請求の範囲】
【請求項１】
連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、該漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと該漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、該漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書とを参照して、該漢字複合語を分割することを特徴とする漢字複合語分割方法。
【請求項２】
連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語と該品詞の両者を記録した日本語辞書と、該漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと該漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、該漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書とを参照して、該漢字複合語を分割することを特徴とする漢字複合語分割方法。
【請求項３】
前記漢字複合語分割方法は、前記漢字複合語の語頭の漢字又は前記漢字複合語の直前に決定した区切位置の直後にある漢字から、予め設定した抽出字数の順番に従って、抽出字数分の漢字列を順次抽出し、前記日本語辞書を参照して、抽出した漢字列を基本単語と照合する第一のステップと、
第一のステップで抽出した漢字列と一致する基本単語が見つかった場合には、該基本単語と一致する漢字複合語から抽出した漢字列の後方に漢字があるか確認し、該基本単語と一致する漢字複合語から抽出した漢字列の後方に漢字があるときは、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定し、第一のステップに戻る第二のステップと、
第一のステップで予め設定した全ての抽出字数から抽出した漢字列の全部と一致する基本単語が見つからなかった場合には、抽出した漢字１字を前記日本語辞書に存在しない１字未知語と定め、該抽出した漢字１字の後方に漢字があるときは、該抽出した漢字１字とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定し、第一のステップに戻る第三のステップと、
を含むことを特徴とする請求項１又は２に記載の漢字複合語分割方法。
【請求項４】
予め設定した抽出字数の順番は、前記日本語辞書に記録された前記基本単語の字数の大きい順とすることを特徴とする請求項３に記載の漢字複合語分割方法。
【請求項５】
前記漢字複合語分割方法は、二以上の前記１字未知語を連接する第四のステップをさらに含むことを特徴とする請求項３又は４に記載の漢字複合語分割方法。
【請求項６】
前記漢字複合語分割方法は、前記１字未知語を含む隣接する漢字列を連接する第五のステップをさらに含むことを特徴とする請求項３～５のいずれか１項に記載の漢字複合語分割方法。
【請求項７】
前記漢字複合語分割方法は、前記単語分割パターン辞書を参照して、前記分割パターンの出現頻度の高い順に、前記漢字複合語を複数の漢字列に順次仮分割した後、前記日本語辞書を参照して、該仮分割した全ての漢字列を基本単語と照合する第六のステップと、
第六のステップで仮分割した全ての漢字列について一致する基本単語が見つかった場合には、仮分割した全ての漢字列と一致する基本単語が見つかった分割パターンに従い、前記

漢字複合語を分割する区切位置を決定する第七のステップと、
第六のステップで仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかった場合には、前記日本語辞書に存在しない漢字列を未知語と定めると共に、全ての分割パターンについて仮分割したか確認して、全ての分割パターンについて仮分割していないときは、第六のステップに戻り、全ての分割パターンについて仮分割したときは、該未知語の個数が最小であり、かつ分割パターンの出現頻度の最も高い分割パターンに従い、前記漢字複合語を分割する区切位置を決定する第八のステップと、
を含むことを特徴とする請求項１又は２に記載の漢字複合語分割方法。
【請求項８】
前記漢字複合語分割方法は、前記漢字複合語から抽出する漢字列の先頭の文字位置としての抽出先頭位置を前記漢字複合語の語頭又は前記漢字複合語の語頭から設定変更した最新の抽出先頭文字の位置とし、前記漢字複合語の中から、該抽出先頭位置から設定した抽出字数分の漢字列を抽出する第九のステップと、
第九のステップで抽出した漢字列のいずれかの漢字に変更したフラグが付与されているか判定し、第九のステップで抽出した漢字列のいずれかの漢字に変更したフラグが付与されている場合には、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、第九のステップに戻る第十のステップと、
前記日本語辞書を参照して、第九のステップで抽出した漢字列を基本単語と照合する第十一のステップと、
第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つかった場合には、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定すると共に、該基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更した後、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるか確認し、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数未満の文字数の漢字しかないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、第九のステップに戻り、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から抽出字数分後方のものに設定変更して、第九のステップに戻る第十二のステップと、
第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つからなかった場合には、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるか確認し、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字がないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、第九のステップに戻り、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、第九のステップに戻る第十三のステップと、
前記漢字複合語を構成するすべての漢字に変更されたフラグが付与されている場合又は設定した抽出字数が０になった場合には、第十二のステップで決定した区切位置を、前記漢字複合語を分割する区切位置として確定する第十四ステップと、
を含むことを特徴とする請求項１又は２に記載の漢字複合語分割方法。
【請求項９】
第十二ステップは、さらに、第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つかった場合には、前記漢字複合語を分割する区切位置として決定すると共に、前記基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更する前に、前記日本語辞書に従い、前記基本単語と一致する抽出した漢字列に品詞を付与し、第十四ステップは、さらに、変更したフラグが付与されていない漢

字については１字未知語と定めることを特徴とする請求項８に記載の漢字複合語分割方法。
【請求項１０】
連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、
前記漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと前記漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、前記漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書と、
前記漢字複合語の語頭の漢字又は前記漢字複合語の直前に決定した区切位置の直後にある漢字から、予め設定した抽出字数の順番に従って、抽出字数分の漢字列を順次抽出し、前記日本語辞書を参照して、抽出した漢字列を基本単語と照合する抽出照合手段と、
前記抽出照合手段で抽出した漢字列と一致する基本単語が見つかった場合には、前記日本語辞書に従い、基本単語と一致する抽出した漢字列に品詞を付与し、該基本単語と一致する抽出した漢字列の後方に漢字があるときは、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、該漢字複合語を分割する区切位置として決定する区切決定手段と、
前記抽出照合手段で予め設定した全ての抽出字数から抽出した漢字列の全部と一致する基本単語が見つからなかった場合には、抽出した漢字１字を前記日本語辞書に存在しない１字未知語と定め、該抽出した漢字１字の後方に漢字があるときは、該抽出した漢字１字とその直後の漢字の間を、該漢字複合語を分割する区切位置として決定する未知語区切決定手段と
を含むことを特徴とする漢字複合語分割装置。
【請求項１１】
前記漢字複合語分割装置は、二以上の前記１字未知語を連接する未知語連接手段をさらに含むことを特徴とする請求項１０に記載の漢字複合語分割装置。
【請求項１２】
前記漢字複合語分割装置は、前記１字未知語を含む隣接する漢字列を連接する隣接語連接手段をさらに含むことを特徴とする請求項１０又は１１に記載の漢字複合語分割装置。
【請求項１３】
前記漢字複合語分割装置は、決定した区切位置を、前記単語分割パターン辞書を参照して、前記漢字複合語を分割する区切位置として確定する区切位置確定手段をさらに含むことを特徴とする請求項１０～１２のいずれか１項に記載の漢字複合語分割装置。
【請求項１４】
連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、
前記漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと前記漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、前記漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書と、
前記単語分割パターン辞書を参照して、前記分割パターンの出現頻度の高い順に、前記漢字複合語を複数の漢字列に順次仮分割した後、前記日本語辞書を参照して、該仮分割した全ての漢字列を基本単語と照合する仮分割照合手段と、
仮分割照合手段で仮分割した全ての漢字列について一致する基本単語が見つかった場合には、前記日本語辞書に従い、基本単語と一致する全ての漢字列に品詞を付与して、仮分割した全ての漢字列と一致する基本単語が見つかった分割パターンに従い、前記漢字複合語

を分割する分割位置を決定する分割決定手段と、
仮分割照合手段で仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかった場合には、前記日本語辞書に存在しない漢字列を未知語と定め、全ての分割パターンについて仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかったときは、該未知語の個数が最小であり、かつ分割パターンの出現頻度の最も高い分割パターンに従い、前記漢字複合語を分割する分割位置を決定する未知語分割決定手段と
を含むことを特徴とする漢字複合語分割装置。
【請求項１５】
前記漢字複合語分割装置は、決定した分割位置を、前記単語分割パターン辞書を参照して、前記漢字複合語を分割する分割位置として確定する分割位置確定手段をさらに含むことを特徴とする請求項１４に記載の漢字複合語分割装置。
【請求項１６】
連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、
前記漢字複合語から抽出する漢字列の先頭の文字位置としての抽出先頭位置を前記漢字複合語の語頭又は前記漢字複合語の語頭から設定変更した最新の抽出先頭文字の位置とし、前記漢字複合語の中から、該抽出先頭位置から設定した抽出字数分の漢字列を抽出する漢字列抽出処理手段と、
前記漢字列抽出処理手段で抽出した漢字列のいずれかの漢字に変更したフラグが付与されているか判定し、前記漢字列抽出処理手段で抽出した漢字列のいずれかの漢字に変更したフラグが付与されている場合には、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、前記漢字列抽出処理手段に戻るフラグ付与判定処理手段と、
前記日本語辞書を参照して、前記漢字列抽出処理手段で抽出した漢字列を基本単語と照合する基本単語照合処理手段と、
前記基本単語照合処理手段において、前記漢字列抽出処理手段で抽出した漢字列と一致する基本単語が見つかった場合には、前記日本語辞書に従い、該基本単語と一致する抽出した漢字列に品詞を付与してから、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定すると共に、該基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更した後、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるか確認し、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数未満の文字数の漢字しかないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、前記漢字列抽出処理手段に戻り、
前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から抽出字数分後方のものに設定変更して、前記漢字列抽出処理手段に戻る第一の照合結果処理手段と、
前記基本単語照合処理手段において、前記漢字列抽出処理手段で抽出した漢字列と一致する基本単語が見つからなかった場合には、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるか確認し、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字がないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、前記漢字列抽出処理手段に戻り、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、前記漢字列抽出処理手段に戻る第二の照合結果処理手段と、
前記漢字複合語を構成するすべての漢字に変更されたフラグが付与されている場合又は設定した抽出字数が０になった場合には、第一の照合結果処理手段で決定した区切位置を、

前記漢字複合語を分割する区切位置として確定し、変更したフラグが付与されていない漢字については１字未知語と定める区切位置確定処理手段と、
を含むことを特徴とする漢字複合語分割装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、日本語文書に含まれる連続する漢字列で構成された漢字複合語を超高精度で分割することができる漢字複合語分割方法及び漢字複合語分割装置に関するものである。
【背景技術】
【０００２】
日本語文書において、主要な概念・テーマは、漢字熟語又は漢字熟語を含む名詞句に表現されることが多い。
【０００３】
漢字複合語は、専門性、特殊性が高く、情報の価値が高いため、漢字複合語を適切に分割する必要性が高まっている。ところが、数文字（例えば、５文字）以上の連続する漢字列で構成された漢字複合語は、非常に複雑な構造を有するため、漢字複合語を高精度で分割することは容易でない。
【０００４】
漢字複合語を分割する手法として、例えば、特許文献１には、単語分割処理として入力した単語の漢字列部分の文字数を設定し、頻度情報配列、単語分割指標配列、分割識別子配列をクリアした後、漢字２文字組の文字列の単語頭及び単語末に出現する頻度情報を備えた辞書に基づいて設定された文字境界の単語末頻度と単語頭頻度から、文字境界に基本単語分割指標（相乗平均・相加平均）及び接辞分割指標（頻度差・頻度和）を設定して、設定した指標により、２文字の漢字語基と１文字の接辞（接頭辞又は接尾辞）に分割する複合語分割装置及び複合語分割方法が開示されている。
【０００５】
特許文献１では、分割は、頻度情報配列、単語分割指標配列、分割識別子配列の３つのデータに基づいてなされる。最初に、対象漢字熟語の長さを設定する。先頭からの個々の文字の位置を示す文字位置と先頭から文字間の境界を示す文字境界位置の２つの指標を用いる。先頭の文字境界位置は０に設定される。文字境界位置に対して、前の２文字漢字列の単語末頻度、後ろの２文字漢字列の単語頭頻度を、頻度格納配列ｆ［Ｉ，ｎ］（Ｉ＝１，２，ｎ＝０，・・・，Ｎ）に設定する。文字位置ｐ（＝１）から１字ずつずらしながら、対象漢字熟語中の２文字漢字列（ｐ＝１，・・・，Ｎ－１）を辞書と照合し、対応する２種類の頻度を設定する。これら２つの頻度に基づき、基本単語分割指標（ｗ［１，ｉ］と接辞分割指標（ｗ［２，ｉ］）を設定し、単語分割指標配列に格納される。
【０００６】
特許文献１では、これらの指標について、複数の計算式を提案している。
（ａ） 和と差
ｗ［１，ｉ］＝ｆ［１，ｉ］＋ｆ［２，ｉ］
ｗ［２，ｉ］＝ｆ［１，ｉ］－ｆ［２，ｉ］
（ｂ）相乗平均と頻度差を頻度和で正規化された値
ｗ［１，ｉ］＝（ｆ［２，ｉ］・ｆ［１，ｉ］）／２
ｗ［２，ｉ］＝（ｆ［１，ｉ］＋ｆ［２，ｉ］））／（ｆ［１，ｉ］－ｆ［２，ｉ］）
【０００７】
特許文献１では、これらの指標以外に、基本単語分割指標として擬似的な確率指標や確率の積、また接辞分割指標としてこれらの正規化差を提案している。
【０００８】
特許文献１では、分割境界の決定は、上述の２つの指標、基本単語分割指標（Ｃｕｔ－Ｗ）と接辞分割指標（Ｃｕｔ－Ｐ）の値の大きさ基づいてなされる。最初に、基本単語分割指標の最大の大きさもつｉ番目の境界で、対象漢字列を２つに部分漢字列に分割する。

それぞれの部分漢字列をさらに２分割し、部分漢字列の長さが４文字以下になるまで、再帰的に繰り返す。次に、長さが３文字以上の部分漢字列を対象に、接辞分割指標に基づいて、接頭辞と基本単語に分割する。接辞分割指標の値が正の場合には、接頭辞と基本単語に分割され、接辞分割指標の値が負の場合には基本単語と接尾辞に分割される。

【０００９】

特許文献１では、実例として「対共産圏輸出統制委員会」を挙げて、分割の過程が説明されている。新聞記事１年分（１２０ＭＢ）を対象に、２文字漢字列の２種類の出現頻度情報を算出している。当該熟語を構成する２文字漢字列と、単語頭頻度、単語末頻度は、「委員」（１９３０，２９７２）、「員会」（３，７５９４）、「共産」（１７３５，２１７）、「産圏」（０，１５）、「制委」（０，１）、「対共」（２４，０）、「統制」（９９，１４５）、「輸出」（１５２９，９００）とし、これらの頻度から、基本単語分割指標（ｗ［１，ｉ］）として上述の（ｂ）を使用すると、「対／共産圏／輸出／統制／委／員会」（１７３５，１５１．４，２８．５，５２９．０，１．７）となる。ここで、“／”は分割境界を示し、カッコ内の数値はその単語分割指標を示している。また、接辞の分割境界とその値は、「対／共産／圏／輸出／統制／委／員会」（＋１，－１，－０．９８，－０．８０，＋０．８６，＋０．５，－１）となる。最初に最大値５２９．０をもつ８文字目の境界で分割し、「対共産圏輸出統制」、「委員会」の２つの部分漢字列に分割される。前者は４文字以上で、さらに、「対共産圏輸出」と「統制」に分割されるが、後者は３文字なのでこれ以上分割されない。対共産圏輸出」は、「対共産圏」と「輸出」に分割される。次に、「対共産圏」と「委員会」に対して、接辞分割指標に基づいて、分割がなされ、正の値をとる「対」が接頭辞に、負の値をとる「圏」、「会」が接尾辞として識別される。

【００１０】

漢字複合語の分割に関する特許文献以外の先行研究としては、例えば、係り受けに着目した手法（非特許文献１）、語基間の接続確率に基づく手法（非特許文献２）、名詞間の意味の共起確率を利用した手法（非特許文献３）、文脈情報を利用した手法（非特許文献４）が挙げられる。

【００１１】

係り受け解析を用いた手法（非特許文献１）

非特許文献では、漢字複合語を構成する語基間の係り受けに着目した自動分割手法が提案されている。「前方の単語から後方の単語に係る」、「単語の係り先は一つに限る」、「複数の単語を一つの単語が受けてもいい」、「係り受けの非交差性を守る」を原則として、数詞、接辞、一般語の３種類に品詞分類し、品詞毎に係り受け規則を定めている。

【００１２】

非特許文献１では、分割は、形態素解析を行い、全分割パターンを作成し、基本単語数をそれぞれ算出するステップ１と、各分割パターンの係り受けの個数を求めるステップ２と、係り受け解析を行いステップ２で求めた語基数の差を求めるステップ３と、差が最小となる分割パターンを自動分割の解とするステップ４の４つのステップにより構成され、ステップ４で解が一意に判断できない場合には、単語の使用頻度による選択を行っている。

【００１３】

非特許文献１において、例えば、「畜産物価格安定法」は次の過程を経て分割される。分割パターン１を「畜産　物価　格安　定法」、分割パターン２を「畜産　物　価格　安　定法」、分割パターン３を「畜産　物価　格　安　定法」、分割パターン４を「畜産　物　価格　安定　法」、分割パターン５を「畜産　物価　格　安定　法」とする。分割パターン１の基本単語数は４、分割パターン２の基本単語数は５、分割パターン３の基本単語数は５、分割パターン４の基本単語数は５、分割パターン５の基本単語数は５となる（ステップ１）。分割パターン１の係り受けの個数は１、分割パターン２の係り受けの個数は２、分割パターン３の係り受けの個数は１、分割パターン４の係り受けの個数は３、分割パターン５の係り受けの個数は２となる（ステップ２）。分割パターン１の語基数の差

は４－１＝３、分割パターン２の語基数の差は５－２＝３、分割パターン３の語基数の差は５－０＝５、分割パターン４の語基数の差は５－３＝２、分割パターン５の語基数の差は５－２＝３となり、ステップ３の最小値は３で、結果として分割解「畜産　物　価格　安定　法」を得る。

【００１４】

語基間の接続確率に基づく手法（非特許文献２）

非特許文献２では、漢字複合語をマルコフモデルの出力と考え、状態遷移モデルで表現し、基本単語からなる語の各遷移確率を用いた自動分割手法の提案を行っている。非特許文献２は、漢字熟語を（接頭辞）基本単語（接尾辞）の形で表現し、初期状態から終了状態までの遷移確率を求め、それが最大となるパターンを解とする。遷移確率は、ベイズの事後確率推定法を利用し、初期確率と繰り返し時の確率を求めるという方法で、レーニングデータを対象に「状態遷移確率推定アルゴリズム」を用いて、トレーニングデータ中の基本単語間の遷移確率を算出している。

【００１５】

非特許文献２において、熟語分割は、:漢字複合語の短単位モデルの遷移図を生成し（ステップ１）、各状態遷移確率を求め（ステップ２）、状態遷移確率が最大のものを解とする（ステップ３）という手順で行われる。

【００１６】

非特許文献２において、例えば、「太陽熱発電」は以下のように分割される。分割解１「太　陽熱　発電」の遷移起確率は０．０１７５、分割解２「太陽　熱　発電」の遷移起確率は０．０５６、分割解３「太陽　熱　発電」の遷移確率は０．０３６、分割解４「太陽　熱発　電」の遷移確率は０．０１２となる。ここで、分割解２と分割解３は分割位置が同じであるが、分割解２では「熱」が接尾辞として扱われ、分割解３では「熱」が接頭辞として扱われるため、同じ分割位置となる２通りの分割パターンが存在する。非特許文献２では、長さ３～１０文字の２５００語の漢字熟語に対して、上述の手法を用いた評価実験を行っている。

【００１７】

名詞間の意味的共起情報による手法（非特許文献３）

非特許文献３では、漢字複合語を構成する基本単語を意味カテゴリーに分類し、カテゴリー間の共起頻度を用いた分割手法の提案し、分割実験を行っている。

【００１８】

非特許文献３では、分割は次の手順で行われる。まず、トレーニングデータの漢字複合語を手動で基本単語に分割し、個々の基本単語に対してあらかじめ体系化されているクラスを付与する。その後、対象漢字複合語を基本単語と照合して、分割する（ステップ１）。ステップ１では全ての分割パターンを求める。次に、基本単語を意味分類辞書と照合してクラス番号を付与し、可能なクラス列を求め（ステップ２）、次いで、クラス間の係り受け規則に基づき、全係り受けクラス列を求める（ステップ３）。そして、提案されている優先度算出方法に基づき、係り受けパターン毎に優先度を算出し、最大の優先度をもつ係り受けパターンを解とする（ステップ４）。

【００１９】

非特許文献３において、例えば、「歩行者通路」は以下のように分割される。まず、ステップ１で対象漢字複合語を基本単語と照合し、「歩行　者　通路」と「歩　行者　通路」に分割される。次に、ステップ２で、基本単語を意味分類辞書と照合して、クラス番号を付与し、可能なクラス列を求めると、「歩行［１３３］者［１１０：１２０］通路［１４７］」と「歩［１１９：１３３：１４５］行者［１２４］通路［１４７］」となる。“：”は、複数のクラスが存在する場合を示している。クラス間の係り受け規則に基づき、［［１３３：１１０］，１４７］、［１３３］、［１１０：１４７］、・・・、［［１１９：１２４］，１４７］、・・・、［１４５，［１２４：１４７］］の合計１０種類の係り受けクラス列が得られ（ステップ３）、個々のクラス列に対する優先度を計算すると、最大の優先度１．３６となる［［１３３：１１０］，１４７］が解となるクラス列で、分

割解は「歩行　者　通路」となる（ステップ４）。特許文献３では、４文字以上の３００８語の漢字熟語に対して、上述の手法を用いた評価実験を行っている。

【００２０】

文脈情報を利用した手法

非特許文献４では、基本単語間の共起情報に基づく、（ａ）共起割合とよんでいる熟語内の基本単語間の修飾比率、（ｂ）相互情報量とよんでいる共起する比率に基づく計算指標、（ｃ）優先度と呼んでいる（ｂ）の相互情報量とテキスト中の名詞の頻度を考慮した指標という３種類の手法－計算式を提案し、評価実験を行っている。

【００２１】

非特許文献４では、分割は次の手順で行われる。まず、対象漢字複合語を基本単語と照合し、分割する（ステップ１）。この段階では全ての分割パターンを求める。次に、各分割パターンに対して上述した指標を算出する（ステップ２）。ここで、各指標における最大の値をもつパターンが分割解となる。

【００２２】

非特許文献４において、例えば、「砂糖類価格安定」は、上述した（ａ）共起割合の指標では、「砂　糖　類　価　格　安　定」は０、「砂　糖　類　価　格　安定」は０、「砂　糖　類　価　格安　定」は０、・・・「砂　糖類　価　格　安定」は０．１０、・・・「砂糖　類　価格　安定」は０．２５となり、最大の値をとる「砂糖　類　価格　安定」が分割解となる。非特許文献４では、５文字、７文字、１０文字の漢字熟語それぞれ１００語に対し、上述した手法を用いた評価実験を行っている。

【先行技術文献】

【特許文献】

【００２３】

【特許文献１】特開２００２－２５９３７０号公報

【非特許文献】

【００２４】

【非特許文献１】宮崎正弘，係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Ｖｏｌ２５，Ｎｏ６，９７０－９７９（１９８４）

【非特許文献２】武田，藤崎，統計的手法による漢字複合語の自動分割，情報処理学会論文誌，Ｖｏｌ２８，Ｎｏ９，９５２－９６１（１９８７）

【非特許文献３】小林義行，徳永健伸，田中穂積，名詞間の意味的共起情報を用いた複合名詞の解析，自然言語処理，Ｖｏｌ３，Ｎｏ１，２９－４３（１９９６）

【非特許文献４】韓東力，加藤浩一，古郡廷治，文脈情報を利用した多文字複合語の分割，電子情報通信学会技術研究報告，Ｖｏｌ１０１，Ｎｏ４０，２９－３４（２００１）

【発明の概要】

【発明が解決しようとする課題】

【００２５】

特許文献１及び非特許文献１～４には、対象熟語の分割に使用される数量的指標はそれぞれ異なるが、いずれも大量の漢字熟語集合から基本単語の出現頻度に基づいて計算され、これらの文献が依拠している熟語の構造、すなわち基本単語の構成パターンについての情報は全く考慮されておらず、実際には長い漢字熟語は構文構造をもっているという共通する特徴がある。

【００２６】

しかしながら、特許文献１及び非特許文献１～４には、漢字複合語の分割に際し、分割候補の生成に慨して多くの計算が必要とされる上、分割対象の熟語が辞書に登録されていない基本単語を含んでいると、数量的指標が算出できず、理論的に分割不能となるという共通する問題点がある。また、非特許文献２～４については、本願発明の発明者らが評価実験を行ったが、性能評価で用いている分割対象熟語の量は３００～３０００語程度であり、熟語が長くなると分割精度は大きく低下するという問題点もある。

【００２７】

以上のことから、学術・特許データベース、あるいはインターネット上のｗｅｂ文書のような大量の文書を対象とする場合には、特許文献１及び非特許文献１～４では、性能評価で得られた分割精度が過度に低下することは容易に推測され、とても実用化することができる程度のものでない。

【００２８】

本発明の目的とするところは、日本語文書に含まれる連続する漢字列で構成された漢字複合語を超高精度で正しく分割することができ、分割した各漢字列の信頼性が実用化することができる程度まで高められた、漢字複合語分割方法及び漢字複合語分割装置を提供することにある。

【課題を解決するための手段】

【００２９】

本発明の発明者は、前記課題を解決するため、鋭意検討を重ねた結果、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と基本単語に該当する品詞を関連付けて記録した日本語辞書と、漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、漢字複合語の字数毎に分類して記録した単語分割パターン辞書とを参照し、分割対象の漢字複合語を分割する漢字複合語分割方法などが上記目的を達成することを見出して、本発明をするに至った。

【００３０】

即ち、本発明の漢字複合語分割方法は、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、該漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと該漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、該漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書とを参照して、該漢字複合語を分割することを特徴とする。

【００３１】

また、本発明の漢字複合語分割方法は、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語と該品詞の両者を記録した日本語辞書と、該漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと該漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、該漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書とを参照して、該漢字複合語を分割することを特徴とする。

【００３２】

本発明の漢字複合語分割方法においては、前記漢字複合語の語頭の漢字又は前記漢字複合語の直前に決定した区切位置の直後にある漢字から、予め設定した抽出字数の順番に従って、抽出字数分の漢字列を順次抽出し、前記日本語辞書を参照して、抽出した漢字列を基本単語と照合する第一のステップと、第一のステップで抽出した漢字列と一致する基本単語が見つかった場合には、該基本単語と一致する漢字複合語から抽出した漢字列の後方に漢字があるか確認し、該基本単語と一致する漢字複合語から抽出した漢字列の後方に漢字があるときは、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定して、第一のステップに戻る第二のステップと、第一のステップで予め設定した全ての抽出字数から抽出した漢字列の全部と一致する基本単語が見つからなかった場合には、抽出した漢字１字を前記日本語辞書に存在しない１字未知語と定め、該抽出した漢字１字の後方に漢字があるときは、該抽出した漢字１字とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定して、第一のステップに戻る第三のステップとを含む構成を採用することができる。

【００３３】
また、本発明の漢字複合語分割方法においては、予め設定した抽出字数の順番は、前記日本語辞書に記録された前記基本単語の字数の大きい順とする構成も採用することができる（以下、「手法１」ということがある。）。
【００３４】
さらに、本発明の漢字複合語分割方法においては、二以上の前記１字未知語を連接する第四のステップをさらに含む構成をも採用することができ、前記１字未知語を含む隣接する漢字列を連接する第五のステップをさらに含む構成をも採用することができる。
【００３５】
本発明の漢字複合語分割方法においては、前記単語分割パターン辞書を参照して、前記分割パターンの出現頻度の高い順に、前記漢字複合語を複数の漢字列に順次仮分割した後、前記日本語辞書を参照して、該仮分割した全ての漢字列を基本単語と照合する第六のステップと、第六のステップで仮分割した全ての漢字列について一致する基本単語が見つかった場合には、仮分割した全ての漢字列と一致する基本単語が見つかった分割パターンに従い、前記漢字複合語を分割する区切位置を決定する第七のステップと、第六のステップで仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかった場合には、前記日本語辞書に存在しない漢字列を未知語と定めると共に、全ての分割パターンについて仮分割したか確認して、全ての分割パターンについて仮分割していないときは、第六のステップに戻り、全ての分割パターンについて仮分割したときは、該未知語の個数が最小であり、かつ分割パターンの出現頻度の最も高い分割パターンに従い、前記漢字複合語を分割する区切位置を決定する第八のステップとを含む構成を採用することができる（以下、「手法２」ということがある。）。
【００３６】
本発明の漢字複合語分割方法においては、前記漢字複合語から抽出する漢字列の先頭の文字位置としての抽出先頭位置を前記漢字複合語の語頭又は前記漢字複合語の語頭から設定変更した最新の抽出先頭文字の位置とし、前記漢字複合語の中から、該抽出先頭位置から設定した抽出字数分の漢字列を抽出する第九のステップと、第九のステップで抽出した漢字列のいずれかの漢字に変更したフラグが付与されているか判定し、第九のステップで抽出した漢字列のいずれかの漢字に変更したフラグが付与されている場合には、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、第九のステップに戻る第十のステップと、前記日本語辞書を参照して、第九のステップで抽出した漢字列を基本単語と照合する第十一のステップと、第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つかった場合には、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定すると共に、該基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更した後、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるか確認し、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数未満の文字数の漢字しかないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、第九のステップに戻り、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から抽出字数分後方のものに設定変更して、第九のステップに戻る第十二のステップと、第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つからなかった場合には、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるか確認し、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字がないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、第九のステップに戻り、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から一字

分後方のものに設定変更して、第九のステップに戻る第十三のステップと、前記漢字複合語を構成するすべての漢字に変更されたフラグが付与されている場合又は設定した抽出字数が０になった場合には、第十二のステップで決定した区切位置を、前記漢字複合語を分割する区切位置として確定する第十四ステップとを含む構成を採用することができる（以下、「手法３」ということがある。）。

【００３７】

また、本発明の漢字複合語分割方法においては、第十二ステップは、さらに、第十一のステップにおいて、第九のステップで抽出した漢字列と一致する基本単語が見つかった場合には、前記漢字複合語を分割する区切位置として決定すると共に、前記基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更する前に、前記日本語辞書に従い、前記基本単語と一致する抽出した漢字列に品詞を付与し、第十四ステップは、さらに、変更したフラグが付与されていない漢字については１字未知語と定める構成も採用することができる。

【００３８】

また、本発明の第一の漢字複合語分割装置は、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、前記漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと前記漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、前記漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書と、前記漢字複合語の語頭の漢字又は前記漢字複合語の直前に決定した区切位置の直後にある漢字から、予め設定した抽出字数の順番に従って、抽出字数分の漢字列を順次抽出し、前記日本語辞書を参照して、抽出した漢字列を基本単語と照合する抽出照合手段と、前記抽出照合手段で抽出した漢字列と一致する基本単語が見つかった場合には、前記日本語辞書に従い、基本単語と一致する抽出した漢字列に品詞を付与し、該基本単語と一致する抽出した漢字列の後方に漢字があるときは、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、該漢字複合語を分割する区切位置として決定する区切決定手段と、前記抽出照合手段で予め設定した全ての抽出字数から抽出した漢字列の全部と一致する基本単語が見つからなかった場合には、抽出した漢字１字を前記日本語辞書に存在しない１字未知語と定め、該抽出した漢字１字の後方に漢字があるときは、該抽出した漢字１字とその直後の漢字の間を、該漢字複合語を分割する区切位置として決定する未知語区切決定手段とを含むことを特徴とする。

【００３９】

本発明の第一の漢字複合語分割装置については、二以上の前記１字未知語を連接する未知語連接手段をさらに含む構成を採用することができ、前記１字未知語を含む隣接する漢字列を連接する隣接語連接手段をさらに含む構成を採用することができ、決定した区切位置を、前記単語分割パターン辞書を参照して、前記漢字複合語を分割する区切位置として確定する区切位置確定手段をさらに含む構成を採用することができる。

【００４０】

本発明の第二の漢字複合語分割装置は、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、前記漢字複合語を分割した後に構成される各漢字列の字数の配列を示した分割パターンと前記漢字複合語を分割した後に構成される各漢字列に該当する品詞の配列を表した品詞列パターンのうち当該分割パターンに存在するものを関連付け、前記漢字複合語の字数毎に分類して、該分割パターンと該品詞列パターンの両者を記録した単語分割パターン辞書と、前記単語分割パターン辞書を参照して、前記分割パターンの出現頻度の高い順に、前記漢字複合語を複数の漢字列に順次仮分割した後、前記日本語辞書を参照して、該仮分割した全ての漢字列を基本単語と照合する仮分割照合手段と、仮分割照合手段で仮分割した全ての漢字列について

一致する基本単語が見つかった場合には、前記日本語辞書に従い、基本単語と一致する全ての漢字列に品詞を付与して、仮分割した全ての漢字列と一致する基本単語が見つかった分割パターンに従い、前記漢字複合語を分割する分割位置を決定する分割決定手段と、仮分割照合手段で仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかった場合には、前記日本語辞書に存在しない漢字列を未知語と定め、全ての分割パターンについて仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかったときは、該未知語の個数が最小であり、かつ分割パターンの出現頻度の最も高い分割パターンに従い、前記漢字複合語を分割する分割位置を決定する未知語分割決定手段とを含むことを特徴とする。

【００４１】

本発明の第二の漢字複合語分割装置については、決定した分割位置を、前記単語分割パターン辞書を参照して、前記漢字複合語を分割する分割位置として確定する分割位置確定手段をさらに含む構成を採用することができる。

【００４２】

本発明の第三の漢字複合語分割装置は、連続する漢字列で構成された漢字複合語を分割する場合の基となる基本単語と該基本単語に該当する品詞を関連付け、該基本単語の字数毎に分類して、該基本単語と該品詞の両者を記録した日本語辞書と、前記漢字複合語から抽出する漢字列の先頭の文字位置としての抽出先頭位置を前記漢字複合語の語頭又は前記漢字複合語の語頭から設定変更した最新の抽出先頭文字の位置とし、前記漢字複合語の中から、該抽出先頭位置から設定した抽出字数分の漢字列を抽出する漢字列抽出処理手段と、前記漢字列抽出処理手段で抽出した漢字列のいずれかの漢字に変更したフラグが付与されているか判定し、前記漢字列抽出処理手段で抽出した漢字列のいずれかの漢字に変更したフラグが付与されている場合には、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、前記漢字列抽出処理手段に戻るフラグ付与判定処理手段と、前記日本語辞書を参照して、前記漢字列抽出処理手段で抽出した漢字列を基本単語と照合する基本単語照合処理手段と、前記基本単語照合処理手段において、前記漢字列抽出処理手段で抽出した漢字列と一致する基本単語が見つかった場合には、前記日本語辞書に従い、該基本単語と一致する抽出した漢字列に品詞を付与してから、該基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、前記漢字複合語を分割する区切位置として決定すると共に、該基本単語と一致する抽出した漢字列を構成する各々の漢字に付与されたフラグを変更した後、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるか確認し、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数未満の文字数の漢字しかないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、前記漢字列抽出処理手段に戻り、前記漢字複合語において、該基本単語と一致する漢字複合語から抽出した漢字列の後方に前記抽出字数以上の文字数の漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から抽出字数分後方のものに設定変更して、前記漢字列抽出処理手段に戻る第一の照合結果処理手段と、前記基本単語照合処理手段において、前記漢字列抽出処理手段で抽出した漢字列と一致する基本単語が見つからなかった場合には、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるか確認し、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字がないときは、前記抽出字数を一つ減らして設定すると共に、前記抽出先頭文字を前記漢字複合語の語頭に設定変更して、前記漢字列抽出処理手段に戻り、前記漢字複合語において、該基本単語と一致しなかった漢字複合語から抽出した漢字列の後方に漢字があるときは、前記抽出先頭文字を前記抽出先頭位置から一字分後方のものに設定変更して、前記漢字列抽出処理手段に戻る第二の照合結果処理手段と、前記漢字複合語を構成するすべての漢字に変更されたフラグが付与されている場合又は設定した抽出字数が０になった場合には、第一の照合結果処理手段で決定した区切位置を、前記漢字複合語を分割する区切位置として確定し、変更したフラグが付与されていない漢字については１字未知語と定め

る区切位置確定処理手段とを含むことを特徴とする。
【発明の効果】
【００４３】
本発明を用いることによって、日本語文書に含まれる漢字複合語を超高精度で正しく分割することができ、かつ分割した単語の信頼性が非常に高くなり、従来よりも、形態素解析、構文解析は勿論のこと、Ｗｅｂ検索エンジン、音声認識、文字認識、仮名漢字変換などの精度が向上するという利点がある。
【００４４】
本発明は、従来よりも、日本語文書に含まれる漢字複合語の分割処理、形態素解析、構文解析の速度が向上するという利点がある。
【００４５】
それ故、本発明は、従来と異なり、実用化に耐え得るものである。
【図面の簡単な説明】
【００４６】
【図１】本発明の漢字複合語分割装置の基本的な構成の一実施態様を説明する概念図である。
【図２】本発明の漢字複合語分割装置の基本的な構成の他の一実施態様を説明する概念図である。
【図３】本発明の漢字複合語分割装置の基本的な構成の他の一実施態様を説明する概念図である。
【図４】本発明の漢字複合語分割方法を用いて漢字複合語を分割する過程の一例を説明するフロー図である。
【図５】本発明の漢字複合語分割方法を用いて漢字複合語を分割する過程の他の一例を説明するフロー図である。
【図６】本発明の漢字複合語分割方法を用いて漢字複合語を分割する過程の他の一例を説明するフロー図である。
【図７】本発明の漢字複合語分割方法の手法１、手法２及び手法３についての分割精度の評価実験の手順を示す図である。
【図８】本発明の漢字複合語分割方法の手法１、手法２及び手法３を用いて漢字複合語の分割を行った場合における成功の確率を表示したグラフである。
【発明を実施するための形態】
【００４７】
以下、本発明をさらに詳細に説明する。本発明の漢字複合語分割装置は、連続する漢字列で構成された漢字複合語を、日本語辞書と単語分割パターン辞書を参照して、単語に分割する。
【００４８】
本発明の第一の漢字複合語分割装置１０は、日本語辞書１と、単語分割パターン辞書２と、抽出照合手段１１と、区切決定手段１２と、未知語決定手段１３と、未知語連接手段１４と、隣接語連接手段１５と、区切位置確定手段１６とを備える（図１）。
【００４９】
日本語辞書１には、基本単語と基本単語の品詞の両方が関連付けられて記録されている。
【００５０】
基本単語は、漢字複合語を分割する場合に基となる単位であって、語基（ｗｏｒｄ　ｂａｓｅ）と称されることもあり、単独で独立した意味をもつ。例えば、「技術文献」という漢字複合語については、「技術」と「文献」が基本単語となる。基本単語は、多くは文章中に単独で使用されるが、接頭辞（例えば、「本手法」の「本」）や接尾辞（例えば、「数量的の「的」」）など熟語の構成要素としてのみ使用されるものもある。基本単語としては、例えば、広辞苑、三省堂国語辞典、角川類義語辞典、ＥＢ科学技術用語大辞典、電気・電子情報用語辞典、コンピュータ用語辞典などから１～４字の単語を抽出した後、

重複を取り除き、更に、固有名詞、仏教用語、故事成語、化学物質名等を除外したものを使用する。

【００５１】

品詞としては、例えば、名詞、動詞、サ変名詞（以下、「サ変」という。）、形容動詞語幹（以下、「形動」という。）、形容詞語幹（以下、「形容」という。）、接頭辞（以下、「接頭」という。）、接尾辞（以下、「接尾」という。）、副詞、数詞の９種類が挙げられるが、適宜、９種類以外の品詞を追加してもよい。複数品詞の場合には「－」でつなぎ複数記述する（例えば、「下」は「接尾－接頭」）。

【００５２】

日本語辞書１には、例えば、基本単語と基本単語の字数と基本単語の品詞数と基本単語の品詞とが関連付けられて記録されていてもよい。具体的には、日本語辞書１には、「記入」は、記入・２・１・サ変、「材料」は、材料・２・１・名詞、「直交」は、直交・２・１・サ変、「下」は、下・１・２・接尾－接頭と記録される。なお、基本単語と基本単語の字数と基本単語の品詞数と基本単語の品詞の順番は、基本単語、基本単語の字数、基本単語の品詞数、基本単語の品詞の順番で配列してもよく、それ以外の順番で配列してもよい。

【００５３】

単語分割パターン辞書２には、分割パターンとその分割パターンに存在する品詞列パターンの両者が関連付けられ、漢字複合語の字数（例えば、６～１０字）毎に分類して記録されている。

【００５４】

単語分割パターン辞書２は、例えば、広辞苑、三省堂国語辞典、角川類義語辞典、ＥＢ科学技術用語大辞典、電気・電子情報用語辞典、コンピュータ用語辞典などから見出し語を抽出して、連続する漢字列で構成された漢字複合語のみを選び出した後、４字までの短い漢字複合語と重複を取り除き、更に、固有名詞、仏教用語、故事成語、化学物質名等を除外し、漢字複合語の字数（例えば、６～１０字）毎に分類したものを使用する。

【００５５】

分割パターンは漢字複合語を分割した後に構成される各漢字列の字数の配列であり、通常数字で表わされる。分割パターンは、理論上、2^{n-1}（ｎは漢字複合語の字数）通りの組み合わせが考えられるが、実際には、一部の特定の分割パターンに偏り、分割対象となる漢字複合語から2^{n-1}通りのうちの全ての分割パターンが出現するわけではない。

【００５６】

出願人らは、角川類義語辞典（１９８９）の見出し語３６１０７語、広辞苑（１９９６）の見出し語１３６９４９語、ＥＢ科学技術用語大辞典（１９９１）の見出し語１３３３８１語、電気・電子情報用語辞典（１９９７）の見出し語２７９８４語、コンピュータ用語辞典（１９９０）の見出し語７９７９語から漢字複合語のみを選び出した後、４字までの短い漢字複合語と重複を取り除き、更に、固有名詞、仏教用語、故事成語、化学物質名等を除外し、６字～１０字の漢字複合語（６字の漢字複合語１２９５１語、７字の漢字複合語６５２７語、８字の漢字複合語３２１６語、９字の漢字複合語６６６語、１０字の漢字複合語２８６語）について、分割パターンの解析を行った。

【００５７】

漢字複合語が６字の場合における分割パターンとその分割パターンの出現数とその分割パターンに存在する品詞列パターンの数の一例を表１に示す。

【００５８】

【表１】

分割パターン	出現数	品詞列の数	分割パターン	出現数	品詞列の数
3-3	3	1	1-1-2-2	281	58
4-2	18	2	1-2-1-2	455	67
1-2-3	18	6	1-2-2-1	958	72
1-3-2	11	3	2-1-1-2	626	84
1-4-1	3	2	2-1-2-1	2830	96
2-1-3	38	7	2-2-1-1	223	41
2-2-2	7253	41	1-1-1-1-2	19	16
2-3-1	24	10	1-1-1-2-1	76	44
3-1-2	13	7	1-1-2-1-1	11	9
3-2-1	39	6	1-2-1-1-1	17	13
4-1-1	3	2	2-1-1-1-1	24	15
1-1-3-1	1	1	1-1-1-1-1-1	7	5

【００５９】

表１から、漢字複合語が６字の場合、３分割（５７％）と４分割（４２％）で全体の９９％となり、２文字の単語が含まれる漢字複合語が非常に多いことがわかる。また、３分割では、２・２・２という分割パターンが３分割の９８％を占め、４分割では、１文字の単語２個と２文字の単語２個で構成される分割パターン（１・１・２・２、１・２・１・２、１・２・２・１、２・１・２・１、２・２・１・１）が４分割の９９．９％を占めていることがわかる。

【００６０】

漢字複合語が７字の場合における分割パターンとその分割パターンの出現数とその分割パターンに存在する品詞列パターンの数の一例を表２に示す。

【００６１】

【表２】

分割パターン	出現数	品詞列の数	分割パターン	出現数	品詞列の数
1-4-2	6	3	3-1-1-2	1	1
2-2-3	46	6	3-1-2-1	8	4
2-3-2	30	7	3-2-1-1	2	2
2-4-1	1	1	1-1-1-2-2	28	16
3-2-2	32	5	1-1-2-1-2	40	27
3-3-1	1	1	1-1-2-2-1	115	48
4-1-2	4	2	1-2-1-1-2	47	32
1-1-3-2	1	1	1-2-1-2-1	235	71
1-2-1-3	1	1	1-2-2-1-1	16	11
1-2-2-2	501	59	2-1-1-1-2	48	32
1-2-3-1	8	5	2-1-1-2-1	399	80
1-3-1-2	3	3	2-1-2-1-1	77	27
1-3-2-1	3	3	2-2-1-1-1	21	13
1-4-1-1	1	1	1-1-1-1-2-1	5	4
2-1-1-3	8	6	1-1-1-2-1-1	3	3
2-1-2-2	1586	73	1-1-2-1-1-1	1	1
2-2-1-2	886	68	1-2-1-1-1-1	2	2
2-2-2-1	2330	69	2-1-1-1-1-1	2	2
2-3-1-1	3	3	1-1-1-1-1-1-1	3	1

【００６２】
表２から、漢字複合語が７字の場合、４分割（８２％）と５分割（１６％）で全体の９８％となることがわかる。また、４分割では、２・２・２・１という分割パターンが４分割全体の４３％、２・１・２・２という分割パターンが４分割の３０％を占め、他にも２・２・１・２という分割パターンや１・２・２・２という分割パターンのように出現頻度の高い分割パターンは存在し、５分割では、２・１・１・２・１、１・２・１・２・１、１・１・２・２・１の３つの分割パターンで、５分割の７３％を占めることがわかる。
【００６３】
漢字複合語が８字の場合における分割パターンとその分割パターンの出現数とその分割パターンに存在する品詞列パターンの数の一例を表３に示す。
【００６４】
【表３】

分割パターン	出現数	品詞列の数	分割パターン	出現数	品詞列の数
2-4-2	2	1	1-3-1-2-1	2	2
4-2-2	6	3	2-1-1-2-2	111	57
1-2-2-3	1	1	2-1-1-3-1	2	1
1-2-3-2	1	1	2-1-2-1-2	197	58
1-3-2-2	2	2	2-1-2-2-1	589	88
1-4-2-1	3	1	2-2-1-1-2	61	35
2-1-2-3	14	5	2-2-1-2-1	489	81
2-1-3-2	11	4	2-2-2-1-1	18	11
2-2-1-3	3	2	3-1-1-2-1	2	2
2-2-2-2	1140	60	1-1-1-1-2-2	2	2
2-2-3-1	8	7	1-1-1-2-1-2	1	1
2-3-1-2	6	5	1-1-1-2-2-1	7	7
2-3-2-1	26	9	1-1-2-1-1-2	2	2
3-1-2-2	9	7	1-2-2-1-2-1	7	7
3-2-1-2	4	3	1-1-2-2-1-1	1	1
3-2-2-1	16	5	1-2-1-1-1-2	2	2
4-1-2-1	1	1	1-2-1-1-2-1	20	17
1-1-2-2-2	22	17	1-2-1-2-1-1	7	7
1-1-2-3-1	1	1	2-1-1-1-1-2	4	3
1-2-1-2-2	106	49	2-1-1-1-2-1	21	15
1-2-1-3-1	2	2	2-1-1-2-1-1	5	4
1-2-2-1-2	51	34	2-2-1-1-1-1	2	2
1-2-2-2-1	144	54	1-1-1-1-1-2-1	1	1

【００６５】
表３から、漢字複合語が８字の場合、４分割（４０％）と５分割（５７％）で全体のほぼ９７％となることがわかる。また、４分割では、２・２・２・２という分割パターンが４分割の９２％を占め、５分割では、１文字の単語２個と２文字の単語３個で構成される分割パターンが５分割の９９％以上を占めているが、各分割パターンで頻度に大きな違いがあることがわかる。なお、漢字複合語６字が３分割で構成される分割パターンが多かったということに比べ、漢字複合語８字は５分割の比率が高くなっているため、漢字複合語の字数が長くなると、２文字の単語のみで構成される分割パターンより、途中で接辞などの１文字の単語を含む分割パターンの方が出現しやすい傾向にあると考えられる。
【００６６】
漢字複合語が９字の場合における分割パターンとその分割パターンの出現数とその分割パターンに存在する品詞列パターンの数の一例を表４に示す。

【００６７】
【表４】

分割パターン	出現数	品詞列の数	分割パターン	出現数	品詞列の数
2-1-4-2	1	1	1-2-1-1-3-1	1	1
2-3-2-2	5	4	1-2-1-2-1-2	5	5
2-4-2-1	4	2	1-2-1-2-2-1	38	30
1-1-4-2-1	1	1	1-2-2-1-1-2	6	5
1-2-2-2-2	31	18	1-2-2-1-2-1	9	8
1-2-3-2-1	1	1	1-2-2-2-1-1	1	1
1-4-1-2-1	1	1	2-1-1-1-2-2	3	3
2-1-1-2-3	2	2	2-1-1-2-1-2	10	8
2-1-2-1-3	1	1	2-1-1-2-2-1	47	25
2-1-2-2-2	92	38	2-1-2-1-1-2	8	7
2-1-2-3-1	1	1	2-1-2-1-2-1	93	33
2-1-3-1-2	1	1	2-1-2-2-1-1	3	3
2-1-3-2-1	1	1	2-2-1-1-2-1	24	11
2-2-1-2-2	89	34	2-2-1-2-1-1	7	5
2-2-2-1-2	41	25	1-1-1-1-2-2-1	2	2
2-2-2-2-1	122	38	1-1-1-2-1-1-2	1	1
2-3-1-2-1	1	1	1-1-2-1-1-2-1	1	1
1-1-2-1-2-2	1	1	1-2-1-1-1-2-1	1	1
1-1-2-2-2-1	3	3	2-1-1-1-1-2-1	3	2
1-2-1-1-2-2	4	4			

【００６８】
表４から、漢字複合語が９字の場合、５分割（５８％）と６分割（４０％）で全体の９８％となることがわかる。また、５分割では、２・２・２・２・１という分割パターンが５分割の３２％を占めているが、出現回数が１といった分割パターンもある程度存在し、上位４つの分割パターン（２・２・２・２・１、２・１・２・１・２・１、２・１・２・２・２、２・２・１・２・２）で全体の５９％を占め、対象となるデータ数が少ないこともあるが、一部の分割パターンに出現が偏っていることがわかる。
【００６９】
漢字複合語が１０字の場合における分割パターンとその分割パターンの出現数とその分割パターンに存在する品詞列パターンの数の一例を表５に示す。
【００７０】

【表５】

分割パターン	出現数	品詞列の数	分割パターン	出現数	品詞列の数
1-4-2-3	1	1	2-1-3-1-2-1	2	1
1-2-2-2-3	1	1	2-2-1-1-2-2	8	6
2-1-2-1-4	1	1	2-2-1-2-1-2	11	7
2-1-2-2-3	1	1	2-2-1-2-2-1	41	25
2-1-4-2-1	2	1	2-2-2-1-1-2	3	3
2-2-2-2-2	45	23	2-2-2-1-2-1	15	7
2-2-2-3-1	1	1	3-1-2-1-2-1	1	1
2-2-3-2-1	1	1	1-1-2-1-2-2-1	2	2
1-1-2-2-2-2	2	2	1-1-2-2-1-2-1	1	1
1-2-1-2-1-3	1	1	1-2-1-1-2-2-1	1	1
1-2-1-2-2-2	4	4	1-2-1-2-1-1-2	1	1
1-2-1-2-3-1	1	1	1-2-1-2-1-2-1	4	4
1-2-2-1-2-2	4	4	2-1-1-1-2-2-1	1	1
1-2-2-2-1-2	4	3	2-1-1-2-1-1-2	2	2
1-2-2-2-2-1	7	5	2-1-1-2-1-2-1	9	9
2-1-1-2-2-2	8	7	2-1-2-1-1-1-2	1	1
2-1-1-3-2-1	1	1	2-1-2-1-1-2-1	8	7
2-1-2-1-2-2	24	19	2-1-2-1-2-1-1	2	2
2-1-2-2-1-2	12	9	2-1-2-2-1-1-1	2	2
2-1-2-2-2-1	46	26	2-2-2-1-1-1-1	1	1

【００７１】

表５から、漢字複合語が１０字の場合、対象となる漢字複合語が少なかったこともあるが、上位４つの分割パターン（２・１・２・２・２・１、２・２・２・２・２、２・２・１・２・２・１、２・１・２・１・２・２）で全体の５５％となり、１文字の単語２個と２文字の単語４個で構成される分割パターンの上位３つの分割パターンのみでも、６分割の５７％、全体の３９％を占めることがわかる。

【００７２】

なお、全体の傾向として、漢字複合語のほとんど全ての分割数は、漢字複合語の字数／２（四捨五入）又は漢字複合語の字数／２（四捨五入）＋１となることがわかる。また、例えば、２・２・２、２・２・２・２、２・２・２・２・２のように全て２文字の単語で構成される分割パターンの出現頻度が高く、２文字の単語を多く含む、例えば、２・２・２・１のような分割パターンの出現頻度も高いが、漢字複合語の字数が長くなると、分割パターンの比率が少なくなる傾向も出ている（例えば、１０文字の２・２・２・２・２と２・１・２・２・２・１）。出現した分割パターンについては、漢字複合語８字までは、漢字複合語の字数が増える毎に分割パターン数が増加しているが、漢字複合語８字以上は、対象となる漢字複合語が減少するので、分割パターンが莫大になってしまうことはないこともわかる。

【００７３】

品詞列パターンは、分割パターンが２・２・２の場合には、例えば、出現頻度が高い順に、名詞・名詞・名詞、名詞・サ変・名詞、名詞・名詞・サ変、サ変・サ変・名詞、サ変・名詞・名詞、名詞・サ変・サ変、サ変・名詞・サ変、形動・名詞・名詞、形動・サ変・名詞、サ変・サ変・サ変、名詞・形動・名詞、形動・名詞・サ変、名詞・形動・サ変、サ変・形動・名詞、名詞・サ変・形動、形動・サ変・サ変、サ変・形動・サ変、名詞・名詞・動詞、名詞・名詞・形動、形動・形動・名詞、名詞・動詞・名詞、動詞・サ変・名詞、動詞・名詞・名詞、名詞・動詞・サ変、形動・形動・サ変、名詞・数詞・名詞、形動・動詞・名詞、サ変・名詞・動詞、サ変・サ変・形動、名詞・サ変・動詞、名詞・形動・形動、サ変・動詞・名詞、形動・名詞・動詞、形動・名詞・形動、動詞・サ変・サ変、動詞・

名詞・サ変、サ変・名詞・形動、接頭辞・名詞・名詞、形動・動詞・サ変、形動・サ変・形動、サ変・動詞・サ変が存在する。

【００７４】

また、品詞列パターンは、分割パターンが２・１・２・１の場合には、例えば、出現頻度が高い順に、名詞・接尾・名詞・名詞、名詞・名詞・サ変・名詞、名詞・接尾・サ変・名詞、名詞・名詞・名詞・名詞、サ変・接尾・名詞・名詞、サ変・接尾・サ変・名詞、名詞・接尾・サ変・接尾、名詞・接尾・名詞・接尾、サ変・名詞・サ変・名詞、サ変・名詞・名詞・名詞、名詞・名詞・サ変・接尾、サ変・接尾・名詞・接尾、サ変・接尾・サ変・接尾、名詞・接頭辞・名詞・名詞、名詞・動詞・サ変・名詞、名詞・名詞・名詞・接尾、サ変・名詞・サ変・接尾、形動・名詞・名詞・名詞、名詞・動詞・名詞・名詞、名詞・接頭辞・サ変・名詞、形動・接尾・名詞・名詞、名詞・接頭辞・名詞・接尾、形動・名詞・サ変・名詞、サ変・名詞・名詞・接尾、名詞・接頭辞・サ変・接尾、名詞・名詞・動詞・名詞、名詞・形容・名詞・名詞、名詞・接尾・形動・接尾、名詞・接尾・名詞・動詞、名詞・名詞・形動・名詞、名詞・接尾・名詞・形容、名詞・接尾・形動・名詞、形動・接尾・サ変・名詞、サ変・接尾・形動・接尾、形動・名詞・名詞・接尾、形動・名詞・サ変・接尾、名詞・数詞・名詞・名詞、サ変・接尾・名詞・形容、名詞・サ変・サ変・名詞、サ変・接尾・形動・名詞、形動・接尾・サ変・接尾、形動・接頭辞・名詞・名詞、サ変・サ変・サ変・名詞、サ変・接頭辞・サ変・接尾、名詞・形容・サ変・名詞、形動・接頭辞・サ変・名詞、サ変・形容・名詞・名詞、名詞・名詞・名詞・動詞、名詞・動詞・名詞・接尾、サ変・接頭辞・名詞・名詞、サ変・接尾・動詞・名詞、サ変・接尾・名詞・動詞、名詞・接頭辞・形動・接尾、名詞・数詞・サ変・名詞、名詞・接尾・形容・名詞、動詞・接尾・名詞・名詞、名詞・サ変・名詞・名詞、名詞・名詞・名詞・形容、名詞・接続・サ変・接尾、名詞・接尾・動詞・名詞、名詞・形容・サ変・接尾、サ変・名詞・動詞・名詞、形動・形容・名詞・名詞、名詞・サ変・名詞・接尾、サ変・形容・サ変・名詞、サ変・名詞・名詞・サ変、動詞・名詞・名詞・名詞、サ変・動詞・名詞・名詞、サ変・名詞・形動・名詞、名詞・接頭辞・形動・名詞、名詞・名詞・サ変・動詞、形動・動詞・名詞・名詞、形動・接尾・形動・接尾、形動・動詞・サ変・名詞、形動・形容・サ変・名詞、サ変・サ変・名詞・名詞、形動・名詞・形動・名詞、動・サ変・名詞・名詞、形動・数詞・サ変・名詞、サ変・サ変・サ変・接尾、名詞・形容・名詞・動詞、名詞・動詞・サ変・動詞、形動・サ変・名詞・接尾、動詞・名詞・サ変・名詞、サ変・接頭辞・名詞・形容、サ変・接頭辞・名詞・接尾、サ変・形容・名詞・動詞、サ変・動詞・サ変・接尾、サ変・接頭辞・サ変・名詞、名詞・動詞・動詞・名詞、名詞・接尾・サ変・動詞、形動・形容・名詞・接尾、動詞・名詞・動詞・接尾、サ変・接頭辞・形動・接尾、形動・名詞・形動・接尾が存在する。

【００７５】

単語分割パターン辞書２には、漢字複合語の字数（例えば、６～１０字）毎に、分割パターンの漢字複合語における出現頻度（出現数）の多い順番で、例えば、漢字複合語の字数、分割数、分割パターンを含む単語分割パターンと、単語分割パターンの出現順位と、分割パターンの出現頻度と、分割パターンで分割した後に得られる全ての漢字列の品詞列パターンとが関連付けられて記録されていてもよい。

【００７６】

単語分割パターンとしては、例えば、漢字複合語の字数、分割数、分割パターンの順に、６Ｐ（漢字複合語の字数）３Ｂ（分割数）２２２（分割パターン）、６Ｐ４Ｂ２１２１、７Ｐ４Ｂ２２２１、７Ｐ５Ｂ２１１２１などと表示できる。

【００７７】

単語分割パターン辞書２の記録データの一例としては、例えば、６Ｐ３Ｂ２２２　１　２０６６　名詞・名詞・名詞、６Ｐ３Ｂ２２２　１　１７２５　名詞・サ変・名詞、６Ｐ３Ｂ２２２　１　８３８　名詞・名詞・サ変、６Ｐ４Ｂ２１２１　２　６９８　名詞・接尾・名詞・名詞、６Ｐ３Ｂ２２２　１　５２０　サ変・サ変・名詞、６Ｐ３Ｂ２２２　１　５０７　サ変・名詞・名詞、６Ｐ３Ｂ２２２　１　４２９　名詞・サ変・サ変、６Ｐ４

Ｂ２１２１　２　２８１　名詞・名詞・サ変・名詞などを挙げることができる。この場合、単語分割パターン辞書２は、主記憶にロードされた後に、単語分割パターンとその単語分割パターンに含まれる分割パターンに存在する複数の品詞列パターンとの構成に編成される。なお、単語分割パターンと、単語分割パターンの出現順位と、分割パターンの出現頻度と、分割パターンで分割した後に得られる全ての漢字列の品詞列パターンの順番は、単語分割パターン、単語分割パターンの出現順位、分割パターンの出現頻度、分割パターンで分割した後に得られる全ての漢字列の品詞列パターンの順番で配列してもよく、それ以外の順番で配列してもよい。

【００７８】

抽出照合手段１１は、漢字複合語の語頭の漢字又は漢字複合語の直前に決定した区切位置の直後にある漢字から、予め設定した抽出字数の順番に従って、抽出字数分の漢字列を順次抽出し、日本語辞書１を参照して、抽出した漢字列を基本単語と照合する。

【００７９】

区切決定手段１２は、抽出照合手段１１で抽出した漢字列と一致する基本単語が見つかった場合には、日本語辞書１に従い、基本単語と一致する抽出した漢字列に品詞を付与し、基本単語と一致する抽出した漢字列の後方に漢字があるときは、基本単語と一致する抽出した漢字列の語尾とその直後の漢字の間を、漢字複合語を分割する区切位置として決定する。

【００８０】

未知語決定手段１３は、抽出照合手段１１で予め設定した全ての抽出字数から抽出した漢字列の全部と一致する基本単語が見つからなかった場合には、抽出した漢字１字を日本語辞書１に存在しない１字未知語と定め、抽出した漢字１字の後方に漢字があるときは、抽出した漢字１字とその直後の漢字の間を、漢字複合語を分割する区切位置として決定する。

【００８１】

未知語連接手段１４は、二以上の１字未知語を連接する。二以上の１字未知語が存在する場合には、常に未知語連接手段１４で未知語を連接する処理を行う必要はなく、未知語を連接する処理を行うオプションが付加されているときのみ、未知語連接手段１４で未知語を連接する処理を行えばよい。

【００８２】

未知語連接手段１４では、例えば、漢字複合語のｐ番目の漢字列に未知語決定手段１３で定義した未知語が存在する場合には、ｐ＋１番目以降の漢字列に未知語が存在していないか検索した後、ｐ番目の漢字列から連続するｋ個の未知語を連接して連接未知語とし、未知語決定手段１３で決定した区切位置を、連接未知語の語尾とその直後にある漢字の間に変更する。ここで、連接未知語が日本語辞書１に存在するかどうか検索してもよく、連接未知語が日本語辞書１に存在する場合には、連接未知語に品詞を付与して、未知語決定手段１３で決定した区切位置を、連接未知語の語尾とその直後にある漢字の間に変更し、連接未知語が日本語辞書１に存在しない場合には、未知語の連接は行わないようにしてもよい。

【００８３】

隣接語連接手段１５は、１字未知語を含む隣接する漢字列を連接する。１字未知語を含む隣接する漢字列が存在する場合には、常に隣接語連接手段１５で隣接する漢字列を連接する処理を行う必要はなく、未知語を含む隣接する漢字列を連接する処理を行うオプションが付加されているときのみ、隣接語連接手段１５で隣接する漢字列を連接する処理を行えばよい。

【００８４】

隣接語連接手段１５では、例えば、漢字複合語のｐ番目の漢字列に未知語決定手段１３で定義した未知語が存在する場合には、ｐ番目の漢字列とｐ＋１番目の漢字列を連接して、第一の隣接語とし、第一の隣接語が日本語辞書１に存在するかどうか検索する。第一の隣接語が日本語辞書１に存在する場合には、第一の隣接語に品詞を付与して、未知語決定

手段１３で決定した区切位置を、第一の隣接語の語尾とその直後にある漢字の間に変更する。第一の隣接語が日本語辞書１に存在しない場合には、ｐ番目の漢字列とｐ－１番目の漢字列を連接して、第二の隣接語とし、第二の隣接語が日本語辞書１に存在するかどうか検索する。第二の隣接語が日本語辞書１に存在する場合には、第二の隣接語に品詞を付与して、未知語決定手段１３で決定した区切位置を、第二の隣接語の語尾とその直後にある漢字の間に変更する。第二の隣接語が日本語辞書１に存在しない場合には、隣接する漢字列の連接は行わない。

【００８５】

区切位置確定手段１６は、区切決定手段１２、未知語決定手段１３、未知語連接手段１４、隣接語連接手段１５で決定した区切位置を、単語分割パターン辞書２を参照して、漢字複合語を分割する区切位置として確定する。

【００８６】

区切位置確定手段１６では、第一段階として、単語分割パターン辞書２のうち、分割対象となる漢字複合語の字数に属する分割パターンを検索して、決定した区切位置の各漢字列の字数の配列と一致する分割パターンが存在するか判定する。決定した区切位置の各漢字列の字数の配列と一致する分割パターンが単語分割パターン辞書２に存在する場合には、第二段階として、単語分割パターン辞書２のうち、分割対象となる漢字複合語の字数に属する品詞列パターンを検索して、決定した区切位置で分割した各漢字列に該当する品詞の配列と一致する品詞列パターンが存在するか判定する。決定した区切位置で分割した各漢字列に該当する品詞の配列と一致する品詞列パターンが単語分割パターン辞書２に存在する場合には、決定した区切位置を、漢字複合語を分割する区切位置として確定する。

【００８７】

なお、決定した区切位置の各漢字列の字数の配列と一致する分割パターンが単語分割パターン辞書２に存在しない場合には、一致する分割パターンがないことを示す出力マーカーを付与してもよく、決定した区切位置で分割した各漢字列に該当する品詞の配列と一致する品詞列パターンが単語分割パターン辞書２に存在しない場合には、一致する品詞列パターンがないことを示す出力マーカーを付与してもよい。

【００８８】

次に、本発明の第二の漢字複合語分割装置について説明する。なお、上述した漢字複合語分割装置と同様の事項は記載を省略する。漢字複合語分割装置２０は、日本語辞書１と、単語分割パターン辞書２と、仮分割照合手段２１と、分割決定手段２２と、未知語分割決定手段２３と、分割位置確定手段２４とを備える（図２）。

【００８９】

仮分割照合手段２１は、単語分割パターン辞書２を参照して、分割パターンの出現頻度の高い順に、漢字複合語を複数の漢字列に順次仮分割した後、日本語辞書１を参照して、仮分割した全ての漢字列を基本単語と照合する。

【００９０】

分割決定手段２２は、仮分割照合手段２１で仮分割した全ての漢字列について一致する基本単語が見つかった場合には、日本語辞書１に従い、基本単語と一致する全ての漢字列に品詞を付与して、仮分割した全ての漢字列と一致する基本単語が見つかった分割パターンに従い、漢字複合語を分割する分割位置を決定する。

【００９１】

未知語分割決定手段２３は、仮分割照合手段２１で仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかった場合には、日本語辞書１に存在しない漢字列を未知語と定め、全ての分割パターンについて仮分割した漢字列のいずれかの漢字列に一致する基本単語が見つからなかったときは、未知語の個数が最小であり、かつ分割パターンの出現頻度の最も高い分割パターンに従い、漢字複合語を分割する分割位置を決定する。なお、全ての分割パターンについて仮分割したか確認する過程を設けてもよい。この場合、全ての分割パターンについて仮分割していないときは、仮分割照合手段２１に戻り、全ての分割パターンについて仮分割したときは、未知語の個数が最小であり、かつ分割パタ

ーンの出現頻度の最も高い分割パターンに従い、漢字複合語を分割する分割位置を決定する。

【００９２】

分割位置確定手段２４は、決定した分割位置を、単語分割パターン辞書２を参照して、漢字複合語を分割する分割位置として確定する。

【００９３】

分割位置確定手段２４では、単語分割パターン辞書２のうち、分割対象となる漢字複合語の字数に属する品詞列パターンを検索して、決定した分割位置で分割した各漢字列の品詞の配列と一致する品詞列パターンが存在するか判定する。決定した分割位置で分割した各漢字列の品詞の配列と一致する品詞列パターンが単語分割パターン辞書２の中に存在する場合には、決定した分割位置を、漢字複合語を分割する区切位置として確定する。

【００９４】

次に、本発明の第三の漢字複合語分割装置について説明する。なお、上述した漢字複合語分割装置と同様の事項は記載を省略する。漢字複合語分割装置３０は、漢字列抽出処理手段３１と、フラグ付与判定処理手段３２と、基本単語照合処理手段３３と、第一の照合結果処理手段３４と、第二の照合結果処理手段３５と、区切位置確定処理手段３６とを備える（図３）。

【００９５】

漢字列抽出処理手段３１は、漢字複合語から抽出する漢字列の先頭の文字位置としての抽出先頭位置を漢字複合語の語頭又は漢字複合語の語頭から設定変更した最新の抽出先頭文字の位置とし、漢字複合語の中から、抽出先頭位置から設定した抽出字数分の漢字列を抽出する。

【００９６】

フラグ付与判定処理手段３２は、漢字列抽出処理手段３１で抽出した漢字列のいずれかの漢字に変更したフラグが付与されているか判定し、漢字列抽出処理手段３１で抽出した漢字列のいずれかの漢字に変更したフラグが付与されている場合には、抽出先頭文字を抽出先頭位置から一字分後方のものに設定変更して、漢字列抽出処理手段３１に戻る。

【００９７】

基本単語照合処理手段３３は、日本語辞書１を参照して、漢字列抽出処理手段３１で抽出した漢字列を基本単語と照合する。

【００９８】

第一の照合結果処理手段３４は、基本単語照合処理手段３３において、漢字列抽出処理手段３１で抽出した漢字列と一致する基本単語が見つかった場合には、日本語辞書１に従い、抽出した漢字列に品詞を付与してから、抽出した漢字列の語尾とその直後の漢字の間を、漢字複合語を分割する区切位置として決定すると共に、抽出した漢字列を構成する各々の漢字に付与されたフラグを変更した後、漢字複合語において、抽出した漢字列の後方に抽出字数以上の文字数の漢字があるか確認し、抽出した漢字列の後方に抽出字数未満の文字数の漢字しかないときは、抽出字数を一つ減らして設定すると共に、抽出先頭文字を漢字複合語の語頭に設定変更して、漢字列抽出処理手段３１に戻り、漢字複合語において、抽出した漢字列の後方に抽出字数以上の文字数の漢字があるときは、抽出先頭文字を抽出先頭位置から抽出字数分後方のものに設定変更して、漢字列抽出処理手段３１に戻る。

【００９９】

第二の照合結果処理手段３５は、基本単語照合処理手段３３において、漢字列抽出処理手段３１で抽出した漢字列と一致する基本単語が見つからなかった場合には、漢字複合語において、抽出した漢字列の後方に漢字があるか確認し、抽出した漢字列の後方に漢字がないときは、抽出字数を一つ減らして設定すると共に、抽出先頭文字を漢字複合語の語頭に設定変更して、漢字列抽出処理手段３１に戻り、漢字複合語において、抽出した漢字列の後方に漢字があるときは、抽出先頭文字を抽出先頭位置から一字分後方のものに設定変更して、漢字列抽出処理手段３１に戻る。

【０１００】

区切位置確定処理手段３６は、漢字複合語を構成するすべての漢字に変更されたフラグが付与されている場合又は設定した抽出字数が０になった場合には、第一の照合結果処理手段３４で決定した区切位置を、漢字複合語を分割する区切位置として確定し、変更したフラグが付与されていない漢字については１字未知語と定める。

【０１０１】

本発明の漢字複合語分割方法は、連続する漢字列で構成された漢字複合語を、日本語辞書と単語分割パターン辞書を参照して、単語に分割する。以下、手法１～手法３を例として説明する。

【０１０２】

手法１では、抽出字数の順番は、日本語辞書に記録された基本単語の字数の大きい順に設定する。例えば、基本単語の長さが１字～４字であった場合、日本語辞書１に記録された基本単語の字数は、大きい順に、４字、３字、２字、１字となるため、最初に漢字複合語から４字抽出された後、４字の基本単語との照合が行われ、一致しない場合には、漢字複合語から３字抽出された後、３字の基本単語との照合が行われ、一致しない場合には、漢字複合語から２字抽出された後、２字の基本単語との照合が行われ、一致しない場合には、漢字複合語から１字抽出された後、１字の基本単語との照合が行われる。

【０１０３】

６字の漢字複合語「遠隔早期警戒」は、手法１を用いると、以下の手順で分割される。なお、Ｎは漢字複合語から抽出される漢字列の語頭が漢字複合語の語頭から何番目に位置しているかを示し、Ｌは漢字複合語から適宜抽出される漢字列の字数を示す。

【０１０４】

漢字複合語の語頭（Ｎ＝１）（遠）（Ｓ１０１）から４字（Ｌ＝４）（Ｓ１０２）を取り出し（遠隔早期）（Ｓ１０３）、日本語辞書中の４字の基本単語と照合する（Ｓ１０４）。「遠隔早期」は４字の基本単語に存在しない（Ｓ１０４／Ｎｏ）ため、漢字複合語の語頭（遠）から３字（Ｌ＝３）（Ｓ１０５／Ｎｏ，Ｓ１０６）を取り出し（遠隔早）（Ｓ１０３）、日本語辞書中の３字の基本単語と照合する（Ｓ１０４）。「遠隔早」は３字の基本単語に存在しない（Ｓ１０４／Ｎｏ）ため、漢字複合語の語頭（遠）から２字（Ｌ＝２）（Ｓ１０５／Ｎｏ，Ｓ１０６）を取り出し（遠隔）（Ｓ１０３）、日本語辞書中の２字の基本単語と照合する（Ｓ１０４）。「遠隔」は２字の基本単語に存在する（Ｓ１０４／Ｙｅｓ）ため、第一ステップから第二のステップに進み、漢字列「遠隔」に品詞が付与され（遠隔（形動）　早期警戒）（Ｓ１０７）、基本単語と一致する抽出した漢字列「遠隔」の語尾「隔」とその直後にある漢字「早」との間を単語に分割する区切位置として決定する（遠隔（形動）｜早期警戒）（Ｓ１０９）。

【０１０５】

ここで、Ｎ＝１，Ｌ＝２であるため、Ｎは、１＋２＝３となり（Ｓ１１０）、漢字複合語の数－３（６－３＝３）と同じである（Ｓ１１１／Ｎｏ）ため、次に、直前に分割した区切位置の直後（Ｎ＝１＋２＝３）（早）（Ｓ１１０）から４字（Ｌ＝４）（Ｓ１１２）を取り出し（早期警戒）（Ｓ１０３）、日本語辞書中の４字の基本単語と照合する（Ｓ１０４）。「早期警戒」は４字の基本単語に存在しない（Ｓ１０４／Ｎｏ）ため、直前に分割した区切位置の直後（早）から３字（Ｌ＝３）（Ｓ１０５／Ｎｏ，Ｓ１０６）を取り出し（早期警）（Ｓ１０３）、日本語辞書中の３字の基本単語と照合する（Ｓ１０４）。「早期警」は３字の基本単語に存在しない（Ｓ１０４／Ｎｏ）ため、直前に分割した区切位置の直後（早）から２字（Ｌ＝２）（Ｓ１０５／Ｎｏ，Ｓ１０６）を取り出し（早期）（Ｓ１０３）、日本語辞書中の２字の基本単語と照合する（Ｓ１０４）。「早期」は２字の基本単語に存在する（Ｓ１０４／Ｙｅｓ）ため、第一ステップから第二のステップに進み、漢字列「早期」に品詞が付与され（遠隔（形動）｜早期（形動）　警戒）（Ｓ１０７）、基本単語と一致する抽出した漢字列「遠隔」の語尾「隔」とその直後にある漢字「早」との間を単語に分割する区切位置として決定する（遠隔（形動）｜早期（形動）｜警戒）（Ｓ１０９）。

【０１０６】

ここで、Ｎ＝３，Ｌ＝２であるため、Ｎは、３＋２＝５となり（Ｓ１１０）、漢字複合語の数－３（６－３＝３）より大きい（Ｓ１１１／Ｙｅｓ）が、漢字複合語の語数（６）より小さい（Ｓ１１３／Ｎｏ）ため、次いで、直前に分割した区切候補の直後（Ｎ＝３＋２＝５）（警）（Ｓ１１０）から２字（Ｌ＝６－５＋１＝２）（Ｓ１１４）を取り出し（警戒）、２字の基本単語と照合する（Ｓ１０４）。「警戒」は２字の基本単語に存在するため、第一のステップから第二のステップに進み、漢字列「警戒」に品詞が付与され（遠隔（形動）｜早期（形動）｜警戒（動詞））（Ｓ１０７）、各漢字列に品詞が付与され、かつすべての区切位置が決定した状態となる。この時点では、Ｎ＝５，Ｌ＝２であるため、Ｎは、５＋２＝７となり（Ｓ１１０）、漢字複合語の基本単語と一致する抽出した漢字列の後方に漢字がない（Ｎ＝５＋２＞６）（Ｓ１１３／Ｙｅｓ）ことになる。

【０１０７】

なお、上述の場合には、未知語が全くないため、二以上の１字未知語を連接する第四のステップ（未知語連接）や１字未知語を含む隣接する漢字列を連接する第五のステップ（隣接語連接）は必要とされない。

【０１０８】

しかしながら、日本語辞書に「早期」、「早」及び「期」が存在しないという場合（Ｓ１０５／Ｙｅｓ）には、第三のステップで「早」と「期」は未知語と定義され（Ｓ１０８）、すべての区切位置が決定した状態は、遠隔｜早（未知）｜期（未知）｜警戒となる（Ｓ１０９／Ｙｅｓ）。ここで、第四のステップの未知語連接を行うと、連続する複数の未知語が１つの未知語となり、遠隔｜早期（未知）｜警戒となる。

【０１０９】

また、日本語辞書に「早期」及び「期」が存在しないという場合には、第三のステップで「期」は未知語と定義され（Ｓ１０８）、すべての区切位置が決定した状態は、遠隔｜早｜期（未知）｜警戒となる（Ｓ１０９／Ｙｅｓ）。ここで、第五のステップの隣接語連接を行うと、１字未知語を含む隣接する漢字列が１つの未知語となり、遠隔｜早期（未知）｜警戒となる。

【０１１０】

手法２は、単語分割パターン辞書の情報に基づいて漢字複合語を複数の漢字列に仮分割し、次に仮分割されたすべての漢字列に対して日本語辞書の基本単語と照合する。６字の漢字複合語「遠隔早期警戒」は、手法２を用いると、以下の手順で分割される。

【０１１１】

単語分割パターン辞書に記録された漢字複合語６字の分割パターンのうち、出現頻度が最も高い分割パターンは２・１・２・１であり、出現頻度が二番目に高い分割パターンは２・２・２であるため、最初に、第六のステップで、一番目（ｉ＝１）の分割パターン２・１・２・１（Ｓ２０１）を用いて、「遠隔早期警戒」を「遠隔／早／期警／戒」と仮分割し（Ｓ２０２）、先頭の漢字列から日本語辞書中の１字及び２字の基本単語に対して照合を行う（Ｓ２０３）。

【０１１２】

仮分割した漢字列のうち「期警」については一致する基本単語が見つからない（Ｓ２０３／Ｎｏ）ため、第八のステップで、日本語辞書に存在しない漢字列（期警）は未知語と定義され（遠隔／早／期警（未知）／戒）（Ｓ２０４）、全ての分割パターンについて仮分割されていないことを確認し（Ｓ２０５／Ｙｅｓ）、第六のステップに戻る。

【０１１３】

次に、二番目（ｉ＝１＋１）の分割パターン２・２・２（Ｓ２０８）を用いて、「遠隔早期警戒」を「遠隔／早期／警戒」と仮分割し（Ｓ２０２）、先頭の漢字列から日本語辞書中の２字の基本単語に対して照合を行う（Ｓ２０３）。

【０１１４】

仮分割した漢字列の全部が日本語辞書に存在する、即ち、仮分割した漢字列の全てに一致する基本単語が見つかった（Ｓ２０３／Ｙｅｓ）ため、第七のステップで、すべての漢字列に品詞が付与され（遠隔（形動）｜早期（形動）｜警戒（動詞））（Ｓ２１０）、仮

分割した全ての漢字列と一致する基本単語が見つかった分割パターンの区切位置（２・２・２）を、漢字複合語を分割する分割位置として決定する（遠隔（形動）｜早期（形動）｜警戒（動詞））。

【０１１５】

手法３は、日本語辞書に含まれる基本単語に基づき、漢字複合語の抽出位置を順次移動させながら、分割位置を決定する。８字の漢字複合語「良性副腎皮質腫瘍」は、手法３を用いると、以下の手順で分割される。ここでは、照合方向を前方から後方としている。

【０１１６】

分割対象の漢字複合語「良性副腎皮質腫瘍」に対し、日本語辞書を構成する基本単語の長さ順、例えば、４字の基本単語、３字の基本単語、２字の基本単語、１字の基本単語の順で照合する。漢字複合語から抽出する抽出字数、即ち照合する基本単語の長さ（Ｌｗ）、漢字複合語の語数（Ｌｅｎ）、抽出先頭位置（Ｐｏｓ）、漢字複合語を構成する各々の漢字の解析状態（Ｆｌａｇ）の変数を用意する。ここでＦｌａｇは、漢字複合語を構成する各々の漢字に対する解析状態を表し、０は、初期状態であり、抽出した漢字列と日本語辞書中の基本単語とが一致しなかったことを示し、例えば、１から４は抽出した漢字列と一致した基本単語の長さ（Ｌｗ）を示す。初期状態では、全ての文字のＦｌａｇを０とする。初期設定として、照合方向を前方から後方としたので、抽出先頭位置は漢字複合語の語頭（Ｐｏｓ＝１）、漢字複合語「良性副腎皮質腫瘍」の長さＬｅｎは８となる（Ｓ３０１）。

【０１１７】

まず、漢字複合語の中から、漢字複合語の語頭（Ｐｏｓ＝１）から最初に設定した抽出字数４字（Ｌｗ＝４）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。漢字複合語の語頭から後方４（＝Ｌｗ）文字の個々の漢字のすべてのＦｌａｇが０である（Ｓ３０２／Ｙｅｓ）ため、抽出した漢字列「良性副腎」が日本語辞書中の４字の基本単語と一致するか照合する（Ｓ３０３）。「良性副腎」と一致する４字の基本単語がない（Ｓ３０３／Ｎｏ）ため、抽出先頭文字を漢字複合語の語頭（Ｐｏｓ＝１）から１文字後ろに設定変更する（Ｐｏｓ＝１＋１＝２）（Ｓ３０８）。設定変更した抽出先頭文字の位置と基本単語の長さの和（Ｐｏｓ＋Ｌｗ）は６で、漢字複合語の語数８を超えない（Ｓ３０９／Ｎｏ）ので、１文字後ろに設定変更した抽出先頭位置（Ｐｏｓ＝２）から設定した抽出字数４字（Ｌｗ＝４）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。抽出した漢字列の個々の漢字のすべてのＦｌａｇが０である（Ｓ３０２／Ｙｅｓ）ため、抽出した漢字列「良性副腎」が日本語辞書中の４字の基本単語と一致するか照合する（Ｓ３０３）。「性副腎皮」と一致する４字の基本単語がない（Ｓ３０３／Ｎｏ）ため、抽出先頭文字（Ｐｏｓ＝２）を１文字後ろに設定変更する（Ｐｏｓ＝２＋１＝３）（Ｓ３０８）。ここで、設定変更した抽出先頭文字の位置と基本単語の長さの和（Ｐｏｓ＋Ｌｗ）は７で、漢字複合語の語数８を超えない（Ｓ３０９／Ｎｏ）ので、１文字後方に設定変更した抽出先頭位置（Ｐｏｓ＝３）から設定した抽出字数４字（Ｌｗ＝４）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。抽出した漢字列の個々の漢字のすべてのＦｌａｇが０である（Ｓ３０２／Ｙｅｓ）ため、抽出した漢字列「副腎皮質」が日本語辞書中の４字の基本単語と一致するか照合する（Ｓ３０３）。

【０１１８】

「副腎皮質」と一致する４字の基本単語がある（Ｓ３０３／Ｙｅｓ）ため、「副腎皮質」に品詞（名詞）を付与し（Ｓ３０４）、抽出した漢字列の語尾とその直後の漢字の間を区切位置として決定すると共に、「副腎皮質」の４個の漢字のＦｌａｇに４を付与する（Ｓ３０５）。ここで、漢字複合語を構成する全ての漢字のＦｌａｇは０より大きくない（Ｓ３０６／Ｎｏ）ため、抽出先頭文字を４文字分後方に設定変更する（Ｓ３０７）。抽出先頭位置は７となる（Ｐｏｓ＝３＋４）。ここで、設定変更した抽出先頭文字の位置と基本単語の長さの和（Ｐｏｓ＋Ｌｗ）は１１で、漢字複合語の語数８を超える（Ｓ３０９／

Ｙｅｓ）ので、抽出字数が一字減らした３字に設定変更され、照合する基本単語の長さ（Ｌｗ）は３になる（Ｓ３１０）。

【０１１９】

抽出字数は０でない（Ｓ３１１／Ｎｏ）ため、抽出先頭文字は漢字複合語の語頭（Ｐｏｓ＝１）にする（Ｓ３１２）。漢字複合語の語頭（Ｐｏｓ＝１）から設定変更した抽出字数３字（Ｌｗ＝３）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。漢字複合語の語頭から後方３文字のうち、「副」のＦｌａｇが４である（Ｓ３０２／Ｎｏ）ため、抽出先頭位置は２となる（Ｓ３０８）。ここで、設定変更した抽出先頭文字の位置と基本単語の長さの和（Ｐｏｓ＋Ｌｗ）は６で、漢字複合語の語数８を超えない（Ｓ３０９／Ｎｏ）ので、１文字後方に設定変更した抽出先頭位置（Ｐｏｓ＝２）から設定した抽出字数３字（Ｌｗ＝３）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。漢字複合語の語頭から後方３文字のうち、「副」と「腎」のＦｌａｇが４である（Ｓ３０２／Ｎｏ）ため、抽出先頭位置は３となる（Ｓ３０８）。その後、抽出先頭位置が５となるまで全く同じステップが繰り返され、抽出先頭位置が６のとき、抽出先頭文字の位置と基本単語の長さの和（Ｐｏｓ＋Ｌｗ）が９になり、漢字複合語の語数８を超える（Ｓ３０９／Ｙｅｓ）ため、抽出字数が一字減らした２字に設定変更され、照合する基本単語の長さ（Ｌｗ）は２になる（Ｓ３１０）。

【０１２０】

抽出字数は０でない（Ｓ３１１／Ｎｏ）ため、抽出先頭文字は漢字複合語の語頭（Ｐｏｓ＝１）にする（Ｓ３１２）。漢字複合語の語頭（Ｐｏｓ＝１）から設定変更した抽出字数２字（Ｌｗ＝２）分を抽出し、抽出した漢字列を構成する各々の漢字について、０以外のフラグが付与されているか判定する（Ｓ３０２）。漢字複合語の語頭から後方２（＝Ｌｗ）文字の個々の漢字のすべてのＦｌａｇが０である（Ｓ３０２／Ｙｅｓ）ため、抽出した漢字列「良性」が日本語辞書中の２字の基本単語と一致するか照合する（Ｓ３０３）。「良性」と一致する２字の基本単語がある（Ｓ３０３／Ｙｅｓ）ため、「良性」に品詞（名詞）を付与し（Ｓ３０４）、抽出した漢字列の語尾とその直後の漢字の間を区切位置として決定すると共に、「良性」の２個の漢字のＦｌａｇに２を付与する（Ｓ３０５）。

【０１２１】

以降、ステップ３０６、ステップ３０７、ステップ３０９、ステップ３０２と進み、ステップ３０２でＮｏとなり、ステップ３０８に進み、抽出先頭位置は１字後方に設定変更され、３となる。その後、ステップ３０９、ステップ３０２、ステップ３０８のループが繰り返され、抽出先頭位置が６のときに、漢字複合語の語頭（Ｐｏｓ＝６）から設定変更した抽出字数２字（Ｌｗ＝２）分を抽出し、抽出した漢字列を構成する各々の漢字については、漢字複合語の語頭から後方２（＝Ｌｗ）文字の個々の漢字のすべてのＦｌａｇが０である（Ｓ３０２／Ｙｅｓ）ため、抽出した漢字列「腫瘍」が日本語辞書中の２字の基本単語と一致するか照合する（Ｓ３０３）。「腫瘍」と一致する２字の基本単語がある（Ｓ３０３／Ｙｅｓ）ため、「腫瘍」に品詞（名詞）を付与する（Ｓ３０４）と共に、「腫瘍」の２個の漢字のＦｌａｇに２を付与する（Ｓ３０５）。この段階で、全ての文字のＦｌａｇの値は２又は４となった（Ｓ３０６／Ｙｅｓ）ため、漢字複合語の分割処理は終了する（良性（名詞）｜副腎皮質（名詞）｜腫瘍（名詞））。

【０１２２】

上記の処理において、日本語辞書中の１字の基本単語にない１字の漢字がある場合には、Ｆｌａｇの値は０のままとなり、ステップ３１１が真（Ｓ３１１／Ｙｅｓ）となり、終了する。この場合、Ｆｌａｇの値が０の１字の漢字は未知語と判断される。

【実施例】

【０１２３】

（１）分割精度の評価実験その１

手法１、手法２及び手法３の３つの手法の分割精度を客観的に測定するため、図６に示す手順で評価実験を行った。具体的には、辞書から取り出した６～１０字の漢字複合語（

６字：７７７６語、７字：４３１５語、８字：２０８６語、９字：１１１７語、１０字：５４３語）を漢字熟語ファイルに記録した。漢字熟語ファイルに記録した漢字複合語１５８３７語について、自動単語分割プログラムを用い、上述した手法１、手法２及び手法３のそれぞれを実行して、漢字複合語を分割し、分割した漢字複合語に品詞を付与した。使用した日本語辞書及び単語分割ファイル辞書は上述したフォーマットのファイルを用い、単語分割パターンは、異なる字数のものを比較することができないようにした。その後、予め人手により分割された漢字複合語との比較を判定プログラムで行って、分割の成否を調べた。

【０１２４】

手法１、手法２及び手法３のそれぞれの手法を用いて漢字複合語を分割した結果を表６に示す。また、漢字複合語の字数を横軸とし、漢字複合語を分割したときの成功の確率を縦軸として、グラフ化した結果を図７に示す。

【０１２５】

【表６】

複合語字数	対象語数		手法１	確率	手法２	確率	手法３	確率
6	7776	成功	7590	97.61	7541	96.98	7589	97.6
		失敗	186	2.39	235	3.02	187	2.4
7	4315	成功	4161	96.43	4033	93.46	4164	96.5
		失敗	154	3.57	282	6.54	151	3.5
8	2087	成功	1996	95.64	1933	92.62	1997	95.69
		失敗	91	4.36	154	7.38	90	4.31
9	1117	成功	1082	96.87	1016	90.96	1077	96.42
		失敗	35	3.13	101	9.04	40	3.58
10	543	成功	522	96.13	467	86	522	96.13
		失敗	21	3.87	76	14	21	3.87

【０１２６】

その結果、手法１～手法３のいずれについても、一部の例外（漢字複合語が１０字の場合における手法２及び手法３）はあるが、ほぼ９０％以上の非常に高い確率で漢字複合語の分割が成功していることがわかった。これにより、本発明を用いることによって、日本語文書に含まれる漢字複合語を超高精度で正しく分割することができ、かつ分割した単語の信頼性が非常に高くなることが証明された。

【０１２７】

（２）分割精度の評価実験その２

非特許文献２～４の手法についても、６～１０字の漢字複合語の分割精度を求めてみた。表７に本発明の手法１と非特許文献２～４の手法の分割精度を示す。ただし、分割対象の漢字複合語の特性は本発明の手法１と非特許文献２～４では同一ではないことを考慮されたい。

【０１２８】

【表７】

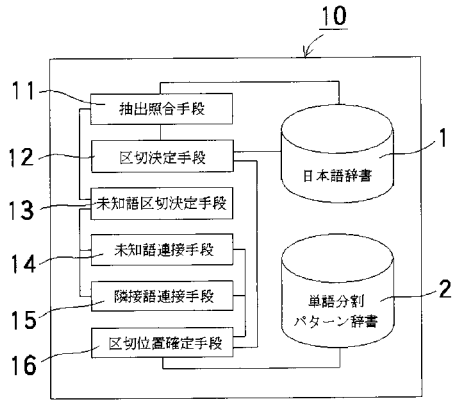
	分割精度（％）			
複合語字数	本発明	非特許文献２	非特許文献３	非特許文献４
6	97.76	89.65	61	
7	96.43	91.17	７文字以上	94
8	95.64	88.00		
9	96.87	86.09	32	
10	96.13	83.19		90

【０１２９】
表７から、全ての漢字複合語の字数で、本発明の手法１が最も高精度であることがわかった。また、本発明の手法１では、漢字複合語の字数が１０字であっても分割精度は９５％以上であるが、非特許文献２～４の手法では最高でも９４％以下であった。さらに、本発明の手法１では総計１５０００語の漢字複合語を対象としており、非特許文献２～４で用いられた漢字複合語と比較しても数倍以上大きい。それ故、本発明の手法は、非特許文献２～４と比較して、学術・特許データベースはもちろんのこと、インターネット上の膨大のｗｅｂページなどの大規模なデータに対しても、相対的に最も有効であることは明らかである。

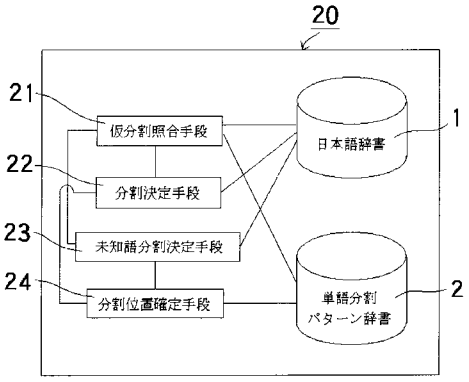
【産業上の利用可能性】
【０１３０】
本発明は、例えば、形態素解析、構文解析は勿論のこと、Ｗｅｂ検索エンジン、音声認識、文字認識、仮名漢字変換などに有用である。

【符号の説明】
【０１３１】
１　日本語辞書
２　単語分割パターン辞書
１０　漢字複合語分割装置
１１　抽出照合手段
１２　区切決定手段
１３　未知語区切決定手段
１４　未知語連接手段
１５　隣接語連接手段
１６　区切位置確定手段
２０　漢字複合語分割装置
２１　仮分割照合手段
２２　分割決定手段
２３　未知語分割決定手段
２４　分割位置確定手段
３０　漢字複合語分割装置
３１　漢字列抽出処理手段
３２　フラグ付与判定処理手段
３３　基本単語照合処理手段
３４　第一の照合結果処理手段
３５　第二の照合結果処理手段
３６　区切位置確定処理手段

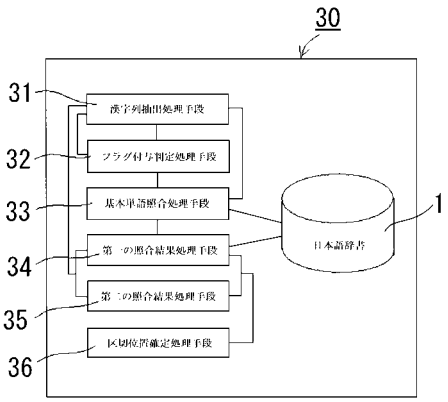
【図１】



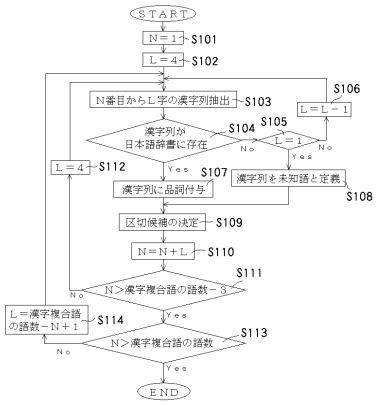
【図２】



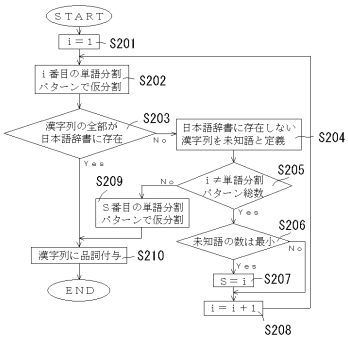
【図３】



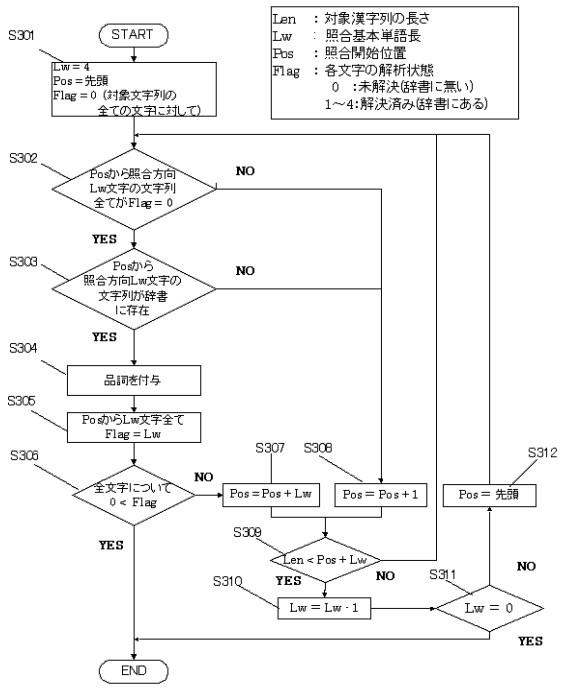
【図４】



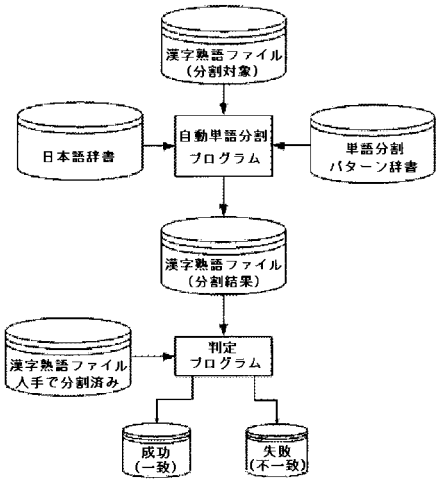
【図５】



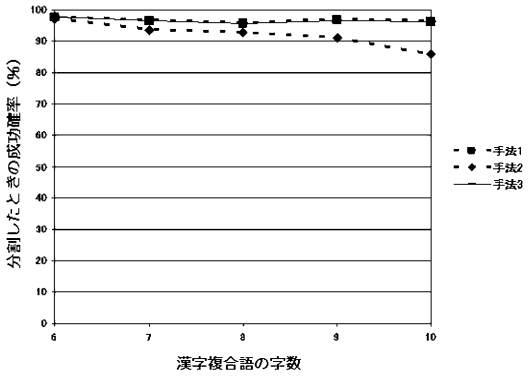
【図６】



【図７】


漢字熟語ファイル
(分割対象)
日本語辞書
自動単語分割
プログラム
単語分割
パターン辞書
漢字熟語ファイル
(分割結果)
漢字熟語ファイル
人手で分割済み
判定
プログラム
成功
(一致)
失敗
(不一致)

【図８】


分割したときの成功確率（%）
100
90
80
70
60
50
40
30
20
10
0
6
7
8
9
10
漢字複合語の字数
手法1
手法2
手法3