

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第5648956号  
(P5648956)

(45) 発行日 平成27年1月7日(2015.1.7)

(24) 登録日 平成26年11月21日(2014.11.21)

(51) Int. Cl. F 1  
G 0 6 F 17/27 (2006.01) G 0 6 F 17/27 E

請求項の数 10 (全 29 頁)

(21) 出願番号	特願2010-222057 (P2010-222057)	(73) 特許権者	592218300
(22) 出願日	平成22年9月30日 (2010. 9. 30)		学校法人神奈川大学
(65) 公開番号	特開2011-96245 (P2011-96245A)		神奈川県横浜市神奈川区六角橋3丁目27番1号
(43) 公開日	平成23年5月12日 (2011. 5. 12)	(74) 代理人	100131679
審査請求日	平成25年8月30日 (2013. 8. 30)		弁理士 ▲高▼橋 幸夫
(31) 優先権主張番号	特願2009-228800 (P2009-228800)	(72) 発明者	後藤 智範
(32) 優先日	平成21年9月30日 (2009. 9. 30)		神奈川県横浜市栄区庄戸5-20-13
(33) 優先権主張国	日本国(JP)	(72) 発明者	梅木 定博
			神奈川県鎌倉市笛田4-8-5
		審査官	成瀬 博之

最終頁に続く

(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) 【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漢字複合語分割方法であって、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候補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、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た後、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候補として決定し、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

10

20

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候補として決定し、第一のステップに戻る第三のステップと、

前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、

10

を含み、該第一のステップ乃至該第三のステップ及び該第九のステップをコンピュータに実行させ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漢字複合語分割方法であって、

20

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候補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、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た後、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候補として決定し、第一のステップに戻る第二のステップと、

30

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候補として決定し、第一のステップに戻る第三のステップと、

前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、

40

を含み、該第一のステップ乃至該第三のステップ及び該第九のステップをコンピュータに実行させ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、前記漢字複合語の漢字列に第三のステップで定義した未知語が存在する場合には、当該未知語の後方の漢字列に未知語が存在していないか検索した後、当該未知語から連続する未知語を連結して接続未知語とし、該接続未知語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を接続未知語の語尾とその直

50

後にある漢字の間に変更すると共に、接続未知語に品詞を付与するが、接続未知語が日本語辞書に存在しなかったときは、未知語の連結を行わない第四のステップをさらに含み、該第四のステップをコンピュータに実行させる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】

前記漢字複合語分割方法は、前記漢字複合語の漢字列に第三のステップで定義した未知語が存在する場合には、当該未知語の直後にある漢字列と連結して、第一の隣接語とし、該第一の隣接語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を該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と共に、該第一の隣接語に品詞を付与するが、該第一の隣接語が前記日本語辞書に存在しなかったときは、当該未知語の直前にある漢字列と連結して、第二の隣接語とし、該第二の隣接語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を該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と共に、該第二の隣接語に品詞を付与するが、該第二の隣接語が前記日本語辞書に存在しなかったときは、隣接する漢字列の連結は行わない第五のステップをさらに含み、該第五のステップをコンピュータに実行させることを特徴とする請求項 1～3 のいずれか 1 項に記載の漢字複合語分割方法。

【請求項 5】

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漢字複合語分割方法であって、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、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た後、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候補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た後、前記漢字複合語を分割する区切候補を決定する第八のステップと、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第六のステップ乃至該第九のステップをコンピュータに実行させることを特徴とする漢字複合語分割方法。

【請求項 6】

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

10

20

30

40

50

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漢字複合語分割方法であって、

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、該仮分割した全ての漢字列に品詞を付与した後、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候補を決定する第七のステップと、

10

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、該仮分割した全ての漢字列に品詞を付与した後、前記漢字複合語を分割する区切候補を決定する第八のステップと、

前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第六のステップ乃至該第九のステップをコンピュータに実行させることを特徴とする漢字複合語分割方法。

20

#### 【請求項7】

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、

30

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候補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候補として決定する区切決定手段と、

40

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候補として決定する未知語区切決定手段と、

前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補の各漢字列の字数の配列と

50

一致する分割パターンが存在した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する分割位置確定手段と、

を含むことを特徴とする漢字複合語分割装置。

【請求項 8】

前記漢字複合語分割装置は、前記漢字複合語の漢字列に未知語区切決定手段で定義した未知語が存在する場合には、当該未知語の後方の漢字列に未知語が存在していないか検索した後、当該未知語から連続する未知語を連結して接続未知語とし、該接続未知語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を接続未知語の語尾とその直後にある漢字の間に変更すると共に、接続未知語に品詞を付与するが、接続未知語が日本語辞書に存在しなかったときは、未知語の連結を行わない未知語接続手段をさらに含むことを特徴とする請求項 7 に記載の漢字複合語分割装置。

10

【請求項 9】

前記漢字複合語分割装置は、前記漢字複合語の漢字列に未知語区切決定手段で定義した未知語が存在する場合には、当該未知語の直後にある漢字列と連結して、第一の隣接語とし、該第一の隣接語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を該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と共に、該第一の隣接語に品詞を付与するが、該第一の隣接語が前記日本語辞書に存在しなかったときは、当該未知語の直前にある漢字列と連結して、第二の隣接語とし、該第二の隣接語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を該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と共に、該第二の隣接語に品詞を付与するが、該第二の隣接語が前記日本語辞書に存在しなかったときは、隣接する漢字列の連結は行わない隣接語接続手段をさらに含むことを特徴とする請求項 7 又は 8 に記載の漢字複合語分割装置。

20

【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、

30

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候補を決定する分割決定手段と、

40

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、前記漢字複合語を分割する分割候補を決定する未知語分割決定手段と、

前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補の各漢字列の字数の配列と

50

一致する分割パターンが存在した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する分割位置確定手段と、

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。

10

【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。

20

【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  $f[I, n]$  ( $I = 1, 2, n = 0, \dots, N$ ) 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 ( $p = 1, \dots, N - 1$ ) 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。

30

【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。

40

(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

$$w[1, i] = (f[2, i] \cdot f[1, i]) / 2$$

$$w[2, i] = (f[1, i] + f[2, i]) / (f[1, i] - f[2, i])$$

【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】

50

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)の値の大きさに基づいてなされる。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、ペ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。

10

#### 【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。

20

#### 【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。

30

#### 【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)。

40

#### 【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]、・・・、[[11

50



9 : 1 2 4 ] , 1 4 7 ]、・・・、[ 1 4 5 , [ 1 2 4 : 1 4 7 ] ] の合計 1 0 種類に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 . 3 6 となる [ [ 1 3 3 : 1 1 0 ] , 1 4 7 ] 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。

10

#### 【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。

20

#### 【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, Vol 25, No 6, 970-979(1984)

30

【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, Vol 28, No 9, 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, Vol 3, No 1, 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, Vol 101, No 40, 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。

40

#### 【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、分割候補の生成に概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

50

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り。

10

【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。

20

【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漢字複合語分割方法であって、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候補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、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た後、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候補として決定し、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候補として決定し、第一のステップに戻る第三のステップと、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第一のステップ

30

40

50

乃至該第三のステップ及び該第九のステップをコンピュータに実行させることを特徴とする（以下、「手法1」ということがある。）。

【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漢字複合語分割方法であって、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候補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、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た後、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候補として決定し、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候補として決定し、第一のステップに戻る第三のステップと、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第一のステップ乃至該第三のステップ及び該第九のステップをコンピュータに実行させることを特徴とする。

【0032】

上述した本発明の漢字複合語分割方法においては、前記漢字複合語の漢字列に第三のステップで定義した未知語が存在する場合には、当該未知語の後方の漢字列に未知語が存在していないか検索した後、当該未知語から連続する未知語を連結して接続未知語とし、該接続未知語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を接続未知語の語尾とその直後にある漢字の間に変更すると共に、接続未知語に品詞を付与するが、接続未知語が日本語辞書に存在しなかったときは、未知語の連結を行わない第四のステップをさらに含み、該第四のステップをコンピュータに実行させる構成を採用することができ、前記漢字複合語の漢字列に第三のステップで定義した未知語が存在する場合には、当該未知語の直後にある漢字列と連結して、第一の隣接語とし、該第一の隣接語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を該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と共に、該第一の隣接語に品詞を付与するが、該第一の隣接語が前記日本語辞書に存在しなかったときは、当該未知語の直前にある漢字列と連結して、第二の隣接語とし、該第二の隣接語が前記日本語辞書に存在したときは、第三のステップで決定した区切候補を該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と共に、該第二の隣接語に品詞を付与するが、該第二の隣接語が前記日本語辞書に存在しなかったときは、隣接する漢字列の連結は行わない第五のステップをさらに含み、該第五のステップをコンピュータに実行させる構成を採用することができる。

【0033】

10

20

30

40

50

さらに、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漢字複合語分割方法であって、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、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た後、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候補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た後、前記漢字複合語を分割する区切候補を決定する第八のステップと、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第六のステップ乃至該第九のステップをコンピュータに実行させることを特徴とする（以下、「手法2」ということがある。）。

【0034】

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漢字複合語分割方法であって、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、該仮分割した全ての漢字列に品詞を付与した後、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候補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、該仮分割した全ての漢字列に品詞を付与した後、前記漢字複合語を分割する区切候補を決定する第八のステップと、前記単語分割パターン辞書の分割対

10

20

30

40

50

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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する第九のステップと、を含み、該第六のステップ乃至該第九のステップをコンピュータに実行させることを特徴とする。

【 0 0 3 5 】

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候補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候補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候補として決定する未知語区切決定手段と、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する分割位置確定手段と、を含むことを特徴とする。

【 0 0 3 6 】

本発明の第一の漢字複合語分割装置については、前記漢字複合語の漢字列に未知語区切決定手段で定義した未知語が存在する場合には、当該未知語の後方の漢字列に未知語が存在していないか検索した後、当該未知語から連続する未知語を連結して接続未知語とし、該接続未知語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を接続未知語の語尾とその直後にある漢字の間に変更すると共に、接続未知語に品詞を付与するが、接続未知語が日本語辞書に存在しなかったときは、未知語の連結を行わない未知語接続手段をさらに含む構成を採用することができ、前記漢字複合語の漢字列に未知語区切決定手段で定義した未知語が存在する場合には、当該未知語の直後にある漢字列と連結して、第一の隣接語とし、該第一の隣接語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を該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と共に、該第一の隣接語に品詞を付与するが、該第一の隣接語が前記日本語辞書に存在しなかったときは、当該未知語の直前にある漢字列と連結して、第二の隣接語とし、該第二の隣接語が前記日本語辞書に存在したときは、未知語区切決定手段で決定した区切候補を該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と共に、該第二の隣接語に品詞を付与するが、該第二の隣接語が前記日本語辞書に存在しなかったときは、隣接する漢字列の連結は行わない隣接語接続手段をさらに含む構成を採用することができる

## 【0037】

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候補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、前記漢字複合語を分割する分割候補を決定する未知語分割決定手段と、前記単語分割パターン辞書の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た結果、前記単語分割パターン辞書中に決定した区切候補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した場合には、前記単語分割パターン辞書の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、前記単語分割パターン辞書中に決定した区切候補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したときは、決定した区切候補を漢字複合語を分割する区切位置として確定する分割位置確定手段と、を含むことを特徴とする。

## 【発明の効果】

## 【0038】

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。

## 【0039】

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。

## 【0040】

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】

## 【0041】

【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他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の手法1及び手法2についての分割精度の評価実験

10

20

30

40

50

の手順を示す図である。

【図6】本発明の漢字複合語分割方法の手法1及び手法2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】

【0042】

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。

【0043】

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)。

【0044】

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。

【0045】

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。

【0046】

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」)。

【0047】

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。

【0048】

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。

【0049】

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。

【0050】

10

20

30

40

50

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。

## 【0051】

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、化学物質名等を除き、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。

## 【0052】

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。

## 【0053】

## 【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

## 【0054】

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。

## 【0055】

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。

## 【0056】



【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

10

20

## 【0057】

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% となるのがわかる。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% を占めるのがわかる。

30

## 【0058】

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。

## 【0059】

【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

10

20

## 【0060】

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%となることからわかる。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。

30

## 【0061】

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。

## 【0062】

40

【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			

10

20

## 【0063】

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

30

## 【0064】

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。

## 【0065】

【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

10

20

## 【0066】

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。

## 【0067】

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。

30

40

## 【0068】

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・

50

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。

【 0 0 6 9 】

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・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、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。

10

20

30

【 0 0 7 0 】

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。

40

【 0 0 7 1 】

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、6 P（漢字複合語の字数）3 B（分割数）2 2 2（分割パターン）、6 P 4 B 2 1 2 1、7 P 4 B 2 2 2 1、7 P 5 B 2 1 1 2 1 などと表示できる。

【 0 0 7 2 】

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、6 P 3 B 2 2 2 1 2 0 6 6 名詞・名詞・名詞、6 P 3 B 2 2 2 1 1 7 2 5 名詞・サ変・名詞、6 P 3 B 2 2 2 1 8 3 8 名詞・名詞・サ変、6 P 4 B 2 1 2 1 2 6 9 8 名詞・接尾・名詞・名詞、6 P 3 B 2 2 2 1 5 2 0 サ変・サ変・名詞、6 P 3 B 2 2 2 1 5 0 7 サ変・名詞・名詞、6 P 3 B 2 2 2 1 4 2 9 名詞・サ変・サ変、6 P 4

50

B 2 1 2 1 2 2 8 1 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。

【 0 0 7 3 】

抽出照合手段 1 1 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候補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。

10

【 0 0 7 4 】

区切決定手段 1 2 は、抽出照合手段 1 1 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候補として決定する。

【 0 0 7 5 】

未知語決定手段 1 3 は、抽出照合手段 1 1 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候補として決定する。

20

【 0 0 7 6 】

未知語接続手段 1 4 は、二以上の 1 字未知語を連結する。二以上の 1 字未知語が存在する場合には、常に未知語接続手段 1 4 で未知語を連結する処理を行う必要はなく、未知語を連結する処理を行うオプションが付加されているときのみ、未知語接続手段 1 4 で未知語を連結する処理を行えばよい。

【 0 0 7 7 】

30

未知語接続手段 1 4 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 1 3 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 個の未知語を連結して接続未知語とし、未知語決定手段 1 3 で決定した区切候補を、接続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、接続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、接続未知語が日本語辞書 1 に存在する場合には、接続未知語に品詞を付与して、未知語決定手段 1 3 で決定した区切候補を、接続未知語の語尾とその直後にある漢字の間に変更し、接続未知語が日本語辞書 1 に存在しない場合には、未知語の連結は行わないようにしてもよい。

【 0 0 7 8 】

40

隣接語接続手段 1 5 は、1 字未知語を含む隣接する漢字列を連結する。1 字未知語を含む隣接する漢字列が存在する場合には、常に隣接語接続手段 1 5 で隣接する漢字列を連結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結する処理を行うオプションが付加されているときのみ、隣接語接続手段 1 5 で隣接する漢字列を連結する処理を行えばよい。

【 0 0 7 9 】

隣接語接続手段 1 5 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 1 3 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 番目の漢字列を連結して、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

50

手段 1 3 で決定した区切候補を、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 番目の漢字列を連結して、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 1 3 で決定した区切候補を、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結は行わない。

【 0 0 8 0 】

区切位置確定手段 1 6 は、区切決定手段 1 2、未知語決定手段 1 3、未知語連接手段 1 4、隣接語連接手段 1 5 で決定した区切候補を、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。

10

【 0 0 8 1 】

区切位置確定手段 1 6 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、決定した区切候補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。決定した区切候補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、決定した区切候補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。決定した区切候補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、決定した区切候補を、漢字複合語を分割する区切位置として確定する。

20

【 0 0 8 2 】

なお、決定した区切候補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を付与してもよく、決定した区切候補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を付与してもよい。

【 0 0 8 3 】

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 2 0 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 2 1 と、分割決定手段 2 2 と、未知語分割決定手段 2 3 と、分割位置確定手段 2 4 とを備える（図 2）。

30

【 0 0 8 4 】

仮分割照合手段 2 1 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。

【 0 0 8 5 】

分割決定手段 2 2 は、仮分割照合手段 2 1 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区切候補を決定する。

40

【 0 0 8 6 】

未知語分割決定手段 2 3 は、仮分割照合手段 2 1 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区切候補を決定する。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 2 1 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

50

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区切候補を決定する。

【0087】

分割位置確定手段24は、決定した区切候補を、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。

【0088】

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、決定した区切候補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。決定した区切候補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、決定した区切候補を、漢字複合語を分割する区切位置として確定する。

10

【0089】

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及び手法2を例として説明する。

【0090】

手法1では、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。

20

【0091】

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。

【0092】

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候補として決定する(遠隔(形動)|早期警戒)(S109)。

30

【0093】

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候補の直後(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候補の直後(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候補の直後(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

40

50



、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候補とする（遠隔（形動）|早期（形動）|警戒）（S109）。

【0094】

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、かつすべての区切候補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。

10

【0095】

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

【0096】

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）、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接続を行うと、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。

20

【0097】

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）、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接続を行うと、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。

30

【0098】

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。

【0099】

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 $2 \cdot 1 \cdot 2 \cdot 1$ であり、出現頻度が二番目に高い分割パターンは $2 \cdot 2 \cdot 2$ 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 $2 \cdot 1 \cdot 2 \cdot 1$ （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）。

40

【0100】

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。

【0101】

次に、二番目（ $i = 1 + 1$ ）の分割パターン $2 \cdot 2 \cdot 2$ （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）。

50

## 【 0 1 0 2 】

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（S 2 0 3 / Y e s）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））（S 2 1 0）、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区切候補位置として決定する（遠隔（形動） | 早期（形動） | 警戒（動詞））。

## 【実施例】

## 【 0 1 0 3 】

## （ 1 ）分割精度の評価実験その 1

手法 1 及び手法 2 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 ~ 1 0 字の漢字複合語（6 字：7 7 7 6 語、7 字：4 3 1 5 語、8 字：2 0 8 6 語、9 字：1 1 1 7 語、1 0 字：5 4 3 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 1 5 8 3 7 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 及び手法 2 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。

## 【 0 1 0 4 】

手法 1 及び手法 2 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。

## 【 0 1 0 5 】

## 【表 6】

複合語字数	対象語数		手法 1	確率	手法 2	確率
6	7776	成功	7590	97.61	7541	96.98
		失敗	186	2.39	235	3.02
7	4315	成功	4161	96.43	4033	93.46
		失敗	154	3.57	282	6.54
8	2087	成功	1996	95.64	1933	92.62
		失敗	91	4.36	154	7.38
9	1117	成功	1082	96.87	1016	90.96
		失敗	35	3.13	101	9.04
10	543	成功	522	96.13	467	86
		失敗	21	3.87	76	14

## 【 0 1 0 6 】

その結果、手法 1 と手法 2 のいずれについても、一部の例外（漢字複合語が 1 0 字の場合における手法 2）はあるが、ほぼ 9 0 % 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。

## 【 0 1 0 7 】

## （ 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 ~ 1 0 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。

【 0 1 0 8 】

【表 7】

複合語字数	分割精度 (%)			
	本発明	非特許文献 2	非特許文献 3	非特許文献 4
6	97.61	89.65	61	
7	96.43	91.17	7 文字以上 32	94
8	95.64	88.00		
9	96.87	86.09		
10	96.13	83.19		90

10

【 0 1 0 9 】

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な 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。

20

【産業上の利用可能性】

【 0 1 1 0 】

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】

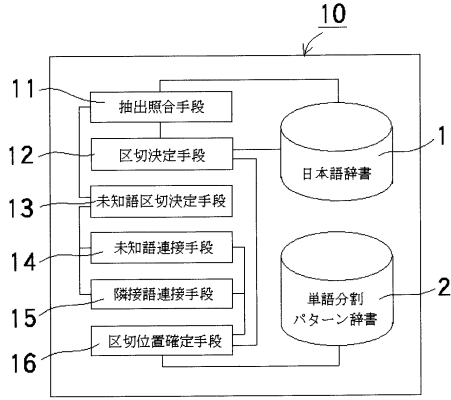
【 0 1 1 1 】

- 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

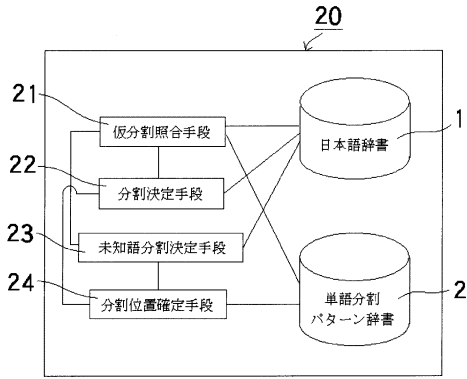
30

40

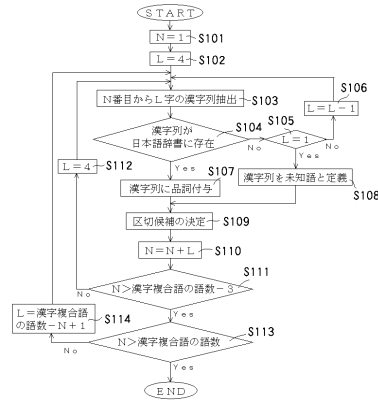
【図1】



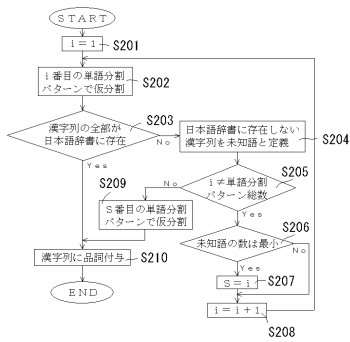
【図2】



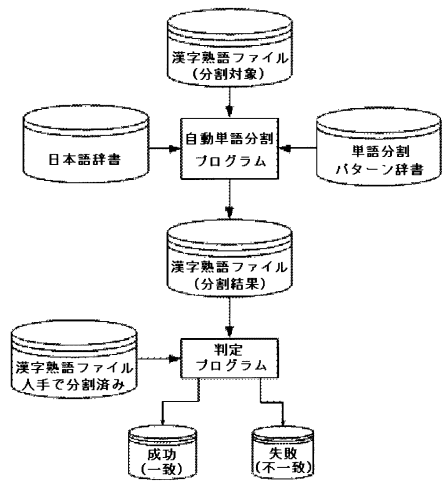
【図3】



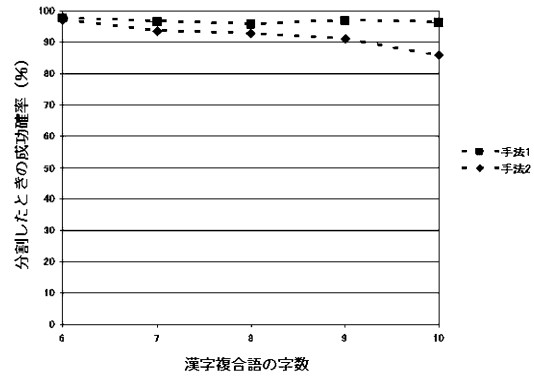
【図4】



【図5】



【図6】



---

フロントページの続き

(56)参考文献 特開平07-262191(JP,A)  
特開平05-020302(JP,A)  
特開平06-161995(JP,A)

(58)調査した分野(Int.Cl., DB名)  
G06F 17/20 - 17/28