(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2012-506596
(P2012-506596A)
(43)公表日　平成24年3月15日(2012.3.15)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	350C	5B091
G06F 17/28	(2006.01)	G06F	17/28	C	
		G06F	17/30	170A	

審査請求 未請求　予備審査請求 未請求　（全 16 頁）

(21)出願番号　特願2011-533276 (P2011-533276)
(86)(22)出願日　平成21年10月20日 (2009.10.20)
(85)翻訳文提出日　平成23年4月21日 (2011.4.21)
(86)国際出願番号　PCT/US2009/061352
(87)国際公開番号　WO2010/048204
(87)国際公開日　平成22年4月29日 (2010.4.29)
(31)優先権主張番号　12/255,372
(32)優先日　平成20年10月21日 (2008.10.21)
(33)優先権主張国　米国 (US)

(71)出願人　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　110001243
特許業務法人　谷・阿部特許事務所
(72)発明者　ラガベンドラ　ウドゥパ　ユー
アメリカ合衆国　９８０５２－６３９９　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　エルシーエー－インターナショナル　パテンツ内

最終頁に続く

(54)【発明の名称】コンパラブルコーパスを使用する固有表現の翻字

(57)【要約】
第１の言語の文書、および第２の言語の追加文書を検査することができる。追加文書が、第１の言語の文書に十分類似しているかどうか判定することができる。追加文書が第１の言語の文書に十分類似していると判定される場合には、文書内の固有表現を選択することができる。この方法では、文書内の固有表現と追加文書内の単語とを比較し、文書内の固有表現と単語が十分類似しているかどうか判定することにより、類似している固有表現を調査することができる。文書内の固有表現に類似する単語が検出されると、文書内の固有表現および類似する固有表現を、固有表現の翻字として記憶することができる。

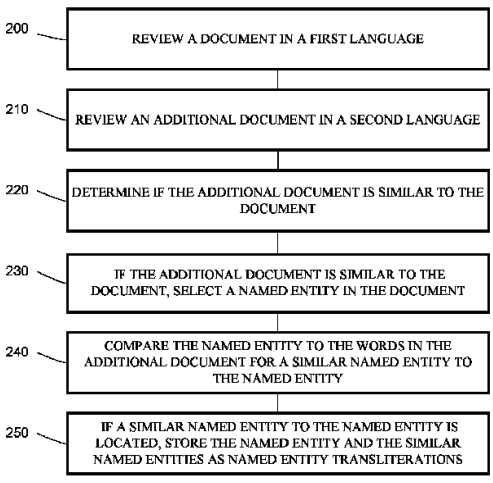


FIGURE 2

【特許請求の範囲】
【請求項１】
多言語の固有表現の翻字を探索する方法であって、
第１の言語の文書を検査するステップと、
第２の言語の追加文書を検査するステップと、
前記追加文書が、前記文書に十分類似しているかどうか判定するステップと、
前記追加文書が、前記文書に十分類似している場合に、
前記文書内の固有表現を選択するステップと、
前記文書内の固有表現を前記追加文書内の単語と比較することを含む、十分類似している固有表現を探索するステップと、
前記文書内の固有表現に十分類似している単語が検索される場合に、前記文書内の固有表現および前記類似している単語を、固有表現の翻字として記憶するステップと
を含むことを特徴とする方法。
【請求項２】
前記追加文書が前記文書に十分類似しているかを判定するステップが、言語横断の類似度スコアを計算するステップをさらに含むことを特徴とする請求項１に記載の方法。
【請求項３】
前記言語横断の類似度スコアを、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒダイバージェンスを使用して計算することを特徴とする請求項２に記載の方法。
【請求項４】
前記言語横断の類似度スコアを、複数の文書と追加文書のペアに対して計算することを特徴とする請求項２に記載の方法。
【請求項５】
最も高い類似度スコアを有する前記文書のペアを選択するステップをさらに含むことを特徴とする請求項４に記載の方法。
【請求項６】
類似している固有表現を探索するステップが、前記単語および前記文書内の固有表現に対して言語横断の類似度スコアを計算するステップを含むことを特徴とする請求項１に記載の方法。
【請求項７】
前記言語横断の類似度スコアが、前記文書内の固有表現と前記単語の間の翻字の同等性の度合いを測定することを特徴とする請求項１に記載の方法。
【請求項８】
前記言語横断の類似度スコアが、複数の固有表現のペアに対して計算され、固有表現のペアが、前記文書内の固有表現および前記追加文書内の前記単語を含むことを特徴とする請求項７に記載の方法。
【請求項９】
前記追加文書内の前記単語が、前記追加文書内の単語のグループから順次選択され、前記グループが、前記追加文書内の前置詞、動詞および形容詞を含まないことを特徴とする請求項８に記載の方法。
【請求項１０】
最大の前記言語横断のスコアを有する前記固有表現のペアが、互いの翻字として選択されることを特徴とする請求項９に記載の方法。
【請求項１１】
多言語の固有表現の翻字を探索するためのコンピュータ実行可能な命令を含むコンピュータ記憶媒体であって、前記コンピュータ実行可能な命令が、
第１の言語の文書を検査する命令と、
第２の言語の追加文書を検査する命令と、
言語横断の類似度スコアを計算することにより、前記追加文書が、前記文書に十分類似しているかどうか判定する命令と、

前記追加文書が、前記文書に十分類似している場合に、
前記文書内の固有表現を選択する命令と、
前記文書内の固有表現を前記追加文書内の単語と比較することを含む、十分に類似している固有表現を探索する命令と、
前記文書内の固有表現に十分類似している単語が検出される場合に、前記文書内の固有表現および前記類似している単語を、固有表現の翻字として記憶する命令と
を含むことを特徴とするコンピュータ記憶媒体。
【請求項１２】
前記言語横断の類似度スコアを、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒダイバージェンスを使用して計算することを特徴とする請求項１１に記載のコンピュータ記憶媒体。
【請求項１３】
前記言語横断の類似度スコアを複数の文書と追加文書のペアに対して計算して、最も高い類似度スコアを有する前記文書のペアを選択することを特徴とする請求項１２に記載のコンピュータ記憶媒体。
【請求項１４】
類似している固有表現を探索する命令が、前記単語および前記文書内の固有表現に対して言語横断の類似度スコアを計算する命令を含み、前記言語横断の類似度スコアが、前記文書内の固有表現と前記単語の間の翻字の同等性の度合いを測定することを特徴とする請求項１１に記載のコンピュータ記憶媒体。
【請求項１５】
前記言語横断の類似度スコアが、複数の固有表現のペアに対して計算され、固有表現のペアが、前記文書内の固有表現および前記追加文書内の前記単語を含むことを特徴とする請求項１４に記載のコンピュータ記憶媒体。
【請求項１６】
前記追加文書内の前記単語が、前記追加文書内の単語のグループから順次選択され、前記グループが、前記追加文書内の前置詞、動詞および形容詞を含まないことを特徴とする請求項１５に記載のコンピュータ記憶媒体。
【請求項１７】
最大の前記言語横断のスコアを有する前記固有表現のペアが、互いの翻字として選択されることを特徴とする請求項１６に記載のコンピュータ記憶媒体。
【請求項１８】
多言語の固有表現の翻字を探索するためのコンピュータ実行可能な命令を実行するプロセッサ、前記プロセッサと通信するメモリ、および入力／出力回路を備えるコンピュータ・システムであって、前記コンピュータ実行可能な命令が、

第１の言語の文書を検査する命令と、
第２の言語の追加文書を検査する命令と、
言語横断の類似度スコアを計算することにより、前記追加文書が、前記文書に十分類似しているかどうか判定する命令であって、前記言語横断の類似度スコアを、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒダイバージェンスを使用して計算する命令と、
前記追加文書が、前記文書に十分類似している場合に、
前記文書内の固有表現を選択する命令と、
前記文書内の固有表現を前記追加文書内の単語と比較することを含む、十分に類似している固有表現を探索する命令と、
前記文書内の固有表現に十分類似している単語が検出される場合に、前記文書内の固有表現および前記類似している単語を、固有表現の翻字として記憶する命令と
を含むことを特徴とするコンピュータ・システム。
【請求項１９】
前記言語横断の類似度スコアが、複数の文書と追加文書のペアに対して計算され、最も高い類似度スコアを有する前記文書のペアを選択することを特徴とする請求項１８に記載

のコンピュータ・システム。
【請求項２０】
類似している固有表現を探索する命令が、前記単語および前記文書内の固有表現に対して言語横断の類似度スコアを計算する命令を含み、
前記言語横断の類似度スコアが、前記文書内の固有表現と前記単語の間の翻字の同等性の度合いを測定し、
前記追加文書内の前記単語が、前記追加文書内の単語のグループから順次選択され、前記グループが、前記追加文書内の前置詞、動詞および形容詞を含まず、
前記言語横断の類似度スコアが、複数の固有表現のペアに対して計算され、固有表現のペアが、前記文書内の固有表現および前記追加文書内の前記単語を含むことを特徴とする請求項１８に記載のコンピュータ・システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンパラブルコーパスを使用する固有表現の翻字に関する。
【背景技術】
【０００２】
この背景は、本特許出願の基本的な事実関係を提供するものであり、解決すべき具体的な問題を説明するものではない。
【０００３】
固有表現は、言語横断情報検索（ＣＬＩＲ）における質問語の重要な部分を形成し、言語横断情報検索システムの性能に大きな影響を及ぼす。機械翻訳（ＭＴ）では、未知語の多くが固有表現である。しかし、対訳辞書では十分な範囲の固有表現に乏しく、機械翻訳システムは、しばしば間違った翻字を生成する。
【発明の概要】
【０００４】
この概要は、発明を実施するための形態において以下でさらに説明する選ばれた概念を、簡略化した形態で紹介するために提供される。この概要は、特許請求される主題の重要な特徴または本質的な特徴を特定するものではなく、特許請求される主題の範囲を限定するために使用するものでもない。
【０００５】
多言語の固有表現の翻字を探索するための方法およびシステムが開示されている。第１の言語の文書を検査し、そして第２の言語の追加文書を検査する。追加文書が、第１の言語の文書に十分類似しているかどうか判定することができる。追加文書が第１の言語の文書に十分類似していると判定される場合には、この文書内の固有表現を選択することができる。この方法では、固有表現を追加文書内の単語と比較し、この固有表現と単語が十分類似しているかどうか判定することにより、類似している固有表現を検索することができる。固有表現に類似している単語が検出されると、この固有表現および類似している固有表現を、固有表現の翻字として記憶することができる。
【図面の簡単な説明】
【０００６】
【図１】携帯用コンピューティング装置の図である。
【図２】多言語の固有表現の翻字を探索する方法を示す図である。
【図３】コンパラブルコーパスを示す図である。
【発明を実施するための形態】
【０００７】
以下の説明では、数多くの様々な実施形態の詳細な説明を行っているが、この説明の法的な範囲は、この特許の最後で述べる特許請求の範囲の説明によって規定されるものと理解すべきである。実現可能な実施形態をすべて説明することは、不可能ではないとしても実際的ではないので、詳細な説明は例示的なものに過ぎないと解釈すべきであり、必ずし

もあらゆる実現可能な実施形態を説明するものではない。現行技術、またはこの発明の出願日以降に開発される技術のいずれかを使用して、数多くの代替実施形態を実施することもできるが、こうした技術もやはり、特許請求の範囲に記載の範囲の中にある。

【０００８】

「本明細書では、用語「＿」は．．．を意味するものとしてここに定義する」といった文、または同様の文を使用して、この特許において用語が明示的に定義されない場合には、明示的にせよ暗示的にせよ、そのありのままの意味または通常の意味を超えて、その用語の意味を限定する意図はなく、こうした用語は、（特許請求の範囲の言葉以外の）この発明の任意のセクションでなされる任意の記述に基づいて範囲を限定的されるものと解釈すべきではないことも理解されたい。この特許の最後にある特許請求の範囲に記載された任意の用語が、この特許において単一の意味と合致するように参照される限りにおいて、それは、読み手を混乱させないよう明確にするためになされるものであり、こうした特許請求の範囲の用語が、暗にまたは他の方法で、その単一の意味に限定されるものではない。最後に、請求項の要素が、任意の構造を列挙することなく、用語「手段」と機能を列挙することによって定義されない限り、任意の請求項の要素の範囲は、米国特許法第１１２条第６項の適用に基づいて解釈されるものではない。

【０００９】

図１に、本明細書で説明するユーザ・インターフェースを表示し提供するように動作することのできる、適切なコンピューティング・システム環境１００の一例を示す。コンピューティング・システム環境１００は、適切なコンピューティング環境のほんの一例に過ぎず、特許請求の範囲に記載の方法および装置の、使用法または機能の範囲に関して、何ら制限を示唆するものではないことに留意されたい。コンピューティング環境１００は、例示的な動作環境１００に示した構成要素のうちの任意の１つの構成要素もしくはそれらの組合せに関して、いかなる依存性または要求をも、有すると解釈すべきではない。

【００１０】

図１を参照すると、特許請求された方法および装置の各ブロックを実施するための例示的なシステムは、コンピュータ１１０の形態の汎用コンピューティング装置を備える。コンピュータ１１０の構成部品には、それだけには限らないが、処理ユニット１２０、システム・メモリ１３０、および、システム・メモリを含む様々なシステム構成部品を処理ユニット１２０に結合するシステム・バス１２１が含まれ得る。

【００１１】

コンピュータ１１０は、モデム１７２または他のネットワーク・インターフェース１７０を用いて、ローカル・エリア・ネットワーク（ＬＡＮ）１７１および／またはワイド・エリア・ネットワーク（ＷＡＮ）１７３を介して、リモート・コンピュータ１８０など、１つまたは複数のリモート・コンピュータとの論理的な接続を使用して、ネットワーク環境で動作することができる。

【００１２】

コンピュータ１１０は、通常、コンピュータ１１０がアクセスすることのできる任意の使用可能な媒体でよい様々なコンピュータ読取り可能な媒体を備え、揮発性および不揮発性の媒体、取外し可能および取外し不可能な媒体の両方を含む。システム・メモリ１３０には、読取り専用メモリ（ＲＯＭ）１３１およびランダム・アクセス・メモリ（ＲＡＭ）１３２など、揮発性および／または不揮発性のメモリの形態のコンピュータ記憶媒体が含まれる。ＲＯＭには、基本入出力システム（ＢＩＯＳ）１３３が含まれ得る。ＲＡＭ１３２は、通常、オペレーティング・システム１３４、アプリケーション・プログラム１３５、他のプログラム・モジュール１３６、およびプログラム・データ１３７を含む、データおよび／またはプログラム・モジュールを含む。コンピュータ１１０にはまた、ハード・ディスク・ドライブ１４１、磁気ディスク１５２との間で読み書きする磁気ディスク・ドライブ１５１、光ディスク１５６との間で読み書きする光ディスク・ドライブ１５５など、他の取外し可能／取外し不可能な、揮発性／不揮発性コンピュータ記憶媒体が含まれ得る。ハード・ディスク・ドライブ１４１、１５１、および１５５は、インターフェース１

４０、１５０を介して、システム・バス１２１とインターフェースすることができる。
【００１３】
ユーザは、キーボード１６２や、マウス、トラックボールまたはタッチ・パッドと普通呼ばれているポインティング装置１６１などの入力装置を用いて、コンピュータ２０にコマンドおよび情報を入力することができる。他の入力装置（図示せず）は、マイクロホン、ジョイスティック、ゲーム・パッド、衛星用パラボラ・アンテナ、スキャナまたは同様のものを含んでもよい。上記その他の入力装置は、システム・バスに結合されたユーザ入力インターフェース１６０を介して、処理ユニット１２０にしばしば接続されているが、パラレル・ポート、ゲーム・ポート、またはユニバーサル・シリアル・バス（ＵＳＢ）など、他のインターフェースおよびバス構造によって接続されてもよい。ビデオ・インターフェース１９０などのインターフェースを介して、モニタ１９１または他のタイプの表示装置もシステム・バス１２１に接続してもよい。モニタに加えて、コンピュータにはまた、出力周辺インターフェース１９０を介して接続してもよいスピーカ１９７やプリンタ１９６など、他の周辺出力装置が含まれ得る。
【００１４】
図２は、多言語の固有表現の翻字のために文書を探索する方法を示すことができる。固有表現（ＮＥ）は、情報検索（ＩＲ）システムにおける質問語の重要な部分を形成し、その性能に大きな影響を及ぼすことがある。これら固有表現は、言語横断情報検索（ＣＬＩＲ）において、さらに重要になることがある。さらに、未知語の多くが実際には固有表現なので、固有表現は、機械翻訳（ＭＴ）システムの性能においても重要な役割を演じる。固有表現は、言語横断情報検索システムがうまく働くかどうかにとって重要であり、機械翻訳の性能に著しい影響を及ぼすが、対訳辞書は、固有表現を十分に網羅していないので、それが手作りであれ統計に基づくものであれ、限られたサポートしか提供しない。ニュース記事およびインターネットにより、毎日、新規の固有表現が言語の語彙に取り入れられる。機械翻字の代替手法では、スペルミスまたは間違った翻字がしばしば生じるが、それらは、言語横断情報検索において雑音として働き、ＭＴにおける翻訳品質を劣化させる。
【００１５】
最近では、多言語で同時にニュース記事が利用可能になることにより、こうしたニュースのコーパスから、固有表現の翻字、特に固有表現の翻字の相当語句（ＮＥＴＥ）を探索することへの見込みのある選択肢に関心が集まってきた。形式的には、ニュースのコンパラブルコーパスとは、適度に長い期間にわたる、一対の言語の時間通りに並んだニュース記事である。世界中の多くの報道機関が、毎日、こうしたニュース・コンテンツを多言語で作成する。ニュースのコンパラブルコーパスから探索された固有表現の翻字の相当語句は、対訳辞書および機械翻字システムを効果的に補完するのに、ＣＬＩＲおよびＭＴを含む多くのタスクにおいて貴重なものになり得る。
【００１６】
ニュース記事は、通常、固有表現３１０を多く含み、したがって、ニュースのコンパラブルコーパスには、固有表現の翻字の相当語句が豊富に存在する。多くの世界中の言語でのニュース・コーパスの膨大な量と永久の可用性により、こうしたコーパスから固有表現３１０の相当語句を探索する効果的な方式を考え出すことができる場合には、固有表現３１０の相当語句を探索するための、膨大で貴重で尽きることのないデータ・ソースが示される。この可能性が、説明する方法およびシステムの推進要因の１つである。
【００１７】
大規模なコンパラブルコーパスから固有表現の翻字の相当語句を効果的に探索するには、いくつかの課題がある。第１に、固有表現の識別および検証には、多くの言語で利用可能でない言語ツールおよび言語学的資源が必要となることがある。第２に、コンパラブルコーパスにおける固有表現の大部分が散在しており、コーパス内の固有表現の署名の頻度にはほとんど依存する必要がない。第３に、探索する方法は、より大きいコーパスを探索するときに効果的に計算できるように、候補の生成を制限しなければならない。さらに、

候補を制限は、誤判定による精度の劣化が減少することになる。最後に、各言語にわたって探索を効果的にするために、言語特有の知識をほとんど使用しないことが重要である。

【００１８】

本出願は、大規模なコンパラブルコーパスから、固有表現の翻字の相当語句を効果的に探索するための、ＭＩＮＴと呼ばれる新規な方法を紹介し、上記に掲載されたあらゆる課題に取り組む。ＭＩＮＴは、ただ１つの言語について固有表現認識装置（ＮＥＲ）の可用性を想定しており、したがって、固有表現認識装置が利用可能である言語とペアを組むときに、リソースが乏しい言語からでさえ固有表現の翻字の相当語句を探索するのに適用可能である。それに応じて、本出願では以下のことを行う。

【００１９】

類似しているコンテンツを有するニュース記事が先験的に知られている場合、これらのニュース記事を、効果的かつ徹底的に探索することができると認識する。

【００２０】

標準のコンパラブルコーパスの場合など、類似している記事が先験的に知られていない場合も、ＭＩＮＴが、上記洞察および言語横断情報検索技法を使用して、最新技術よりもはるかに良好に探索できることを実験的に証明する。

【００２１】

様々な特性を有する多くのコーパスにわたって、その有効性を証明する。

【００２２】

最後に、本出願は、３つの異なる言語ファミリー（すなわち、スラブ語、インドヨーロッパ語およびドラビダ語）からの互いに異なる言語セット（すなわち、ロシア語、ヒンディー語、カンナダ語およびタミル語）間でのいくつかのコンパラブルコーパス上でその性能を説明することにより、この方法が言語に依存しないことを証明する。

【００２３】

ＭＩＮＴ法は、ニュースが、人々、場所、組織、および他の固有表現を含む出来事に関するものなので、類似しているコンテンツを有する多言語のニュース記事が、非常にオーバラップする固有表現のセットを含んでいるはずである、という重要な考えに基づく。同じニュースの出来事を報告する多言語のニュース記事は、それぞれの言語の固有表現に言及するはずであり、したがって、固有表現の翻字の相当語句を豊富に生成することが予想される。図３には、ＢＢＣによって公開された、ヒンディー語および英語での一対の類似している記事における固有表現の翻字の相当語句を示してある。１つのソースによって公開された、英語とタミル語での２００対の類似しているニュース記事を分析すると、英語側の単一語の固有表現３１０のうちの８７％において、タミル語側で少なくとも１つの相当語句が存在したことが分かっている。ＭＩＮＴ法は、この経験に裏打ちされた考えを用いて、こうしたコーパスから固有表現の翻字の相当語句を探索する。

【００２４】

ＭＩＮＴは２つの段階を有することができる。第１の段階で、各文書を比較して、ソース側のあらゆる文書について、類似しているコンテンツを有するターゲット側の文書のセットを識別する。類似している文書が識別されると、それらの文書を入力として第２の段階に与え、そこで、それらの文書から固有表現の翻字の相当語句を探索する。

【００２５】

再び図２を参照すると、ブロック２００で、第１の言語の文書３００（図３）を検査することができる。理想的には、この文書は、対象となるいくつかの固有表現の翻字の相当語句を含むものとする。この文書は、問題となる固有表現の翻字の相当語句を含むときに選択してもよく、または、特定の日向けに書かれたニュース記事を用いた逐次探索でもよい。もちろん、第１の言語の文書を選択する他の方法も可能であり、またそのように意図されている。

【００２６】

ブロック２１０で、第２の言語の追加文書３０５を検査することができる。理想的には、追加文書３０５も、固有表現の翻字の相当語句を有するように選択することができる。

たとえば、第１の言語の文書がスポーツ記事である場合、第２の言語の追加文書として学術論文を検査することは、ほとんど意味がない。文書３００と追加文書３０５の間で、類似している固有表現の翻字の相当語句が存在する確率が低いからである。

【００２７】

ブロック２２０で、追加文書３０５が文書３００に類似しているかどうか判定することができる。この判定は、様々な方式で行うことができる。実施形態によっては、言語横断の文書類似度モデルを使用して、言語横断の類似度スコアを計算する。言語横断の文書類似度モデルにより、ソース言語とターゲット言語の一対の文書間における類似度を測定することができる。文書の確率分布と追加文書の確率分布との間の負のＫｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒ（ＫＬ）ダイバージェンスを、類似度測定値として使用してもよい。

【００２８】

確率論および情報理論では、Ｋｕｌｌｂａｃｋ－Ｌｅｉｂｌｅｒダイバージェンス（また、情報ダイバージェンス、情報利得（ｉｎｆｏｒｍａｔｉｏｎ　ｇａｉｎ）、または相対エントロピー）は、２つの確率分布ＰとＱの間の差の非可換尺度である。ＫＬは、Ｐに基づく符号を使用するとき、およびＱに基づく符号を使用するとき、Ｐからのサンプルを符号化するのに必要となるビットの数の予想される差を測定する。通常Ｐは、データの「真」の分布、観測値、または精密に計算された理論分布を表す。測定値Ｑは通常、Ｐの理論、モデル、記述、または近似を表す。

【００２９】

本出願では、ソース言語およびターゲット言語の語彙を示すＶｓ、Ｖｔを有する、それぞれソース言語およびターゲット言語の２つの文書ＤＳ３００、ＤＴ３０５が与えられている場合、これら２つの文書３００と３０５の間の類似度は、ＫＬ（Ｄｓ‖Ｄｔ）で与えられ得る。

【００３０】

$$\sum_{w_T \in V_T} P(w_T \mid D_s) \log \frac{p(w_T \mid D_T)}{p(w_T \mid D_S)}$$

【００３１】

ここで、ｐ（ｗ｜Ｄ）は、単語ｗが文書Ｄ内に存在する確率である。所与のソース言語の文書３００に類似しているターゲットの文書３０５を見つけることが関心であるので、分子は、ターゲット言語の文書と無関係であるときには無視してもよい。最後に、以下のようにｐ（w_t｜Ｄｓ）を展開する。

【００３２】

$$\sum_{w_s e V_s} p(w_S \mid D_S) p(w_T \mid w_S)$$

【００３３】

言語横断の類似度スコアは、以下のように指定することができる。

【００３４】

$$CrossLanguageDocumentSimilarity(D_S, D_T, MD) =$$

【００３５】

$$\sum_{w_T \in V_T} \sum_{w_S \in V_S} p(w_S D_S) p(w_T \mid w_S) \log p(w_T \mid D_T)$$

【００３６】

擬似コードでは、文書３００、３０５の比較を、以下のように進めることができる。

【００３７】

```
Input: Comparable news corpora (C_S, C_T) in languages (S,T)
       Crosslanguage Document Similarity Model MD for (S,T)
       Threshold score a.
Output: Set A_S,T of pairs of similar articles (D_S, D_T) from (C_S, C_T).
1 A_ST←ø;      //Set of Similar articles (D_S,D_T)
2 for each article D_S in C_S do
3       XS←ø;          //Set of candidates for D_S.
4       for each article d_T in C_T do
5               score = CrossLanguageDocumentSimilarity(D_S,d_T,MD);
6               if (score ≥ α) then X_S ← X_S ∪ (d_T, score);
7       end
8       D_T = BestScoringCandidate(X_S);
9 if (DT≠ø) then A_ST←A_ST ∪ (D_S,D_T);
10 end
```

【００３８】

擬似コードから分かるように、複数の追加文書３０５を文書３００と比較することができる。実施形態によっては、追加文書３０５は事前選別して、固有表現３１０に類似している単語３１５を有する可能性のある追加文書３０５のみを検査することを確実にする。例として、Ｍｉｃｈａｅｌ　Ｐｈｅｌｐｓに着目した文書３００は、おそらくスポーツに関連したものになる。この知識を使用して、検査される追加文書３０５のタイプを減らしてもよい。各追加文書／文書のペア３２０（元のソース文書３００および個別の各追加文書３０５）に対して類似度スコアを計算することができ、最も高い類似度を有するペア３２０を、さらに分析すべき文書／追加文書のペア３２０として使用することができる。

【００３９】

ブロック２３０で、追加文書３０５が文書３００に十分類似している場合、文書内の固有表現３１０を選択することができる。たとえば、水泳選手Ｍｉｃｈａｅｌ　Ｐｈｅｌｐｓは、アメリカ人の名前であり、多くのスポーツ記事で認識が容易である可能性がある。しかし、Ｍｉｃｈａｅｌ　Ｐｈｅｌｐｓは、他の言語では作成が難しいことがある。したがって、Ｍｉｃｈａｅｌ　Ｐｈｅｌｐｓは、この方法が探索しようとし得る固有表現３１０の一例であってよい。

【００４０】

複数の追加文書３０５が、文書３００と比較された場合、最も高いと判定された類似度を有する文書３００／追加文書３０５のペア３２０を、さらに分析すべき選択されたペア３２０として選ぶことができる。いずれのペア３２０も十分な類似度スコアに達しない場合、追加文書３０５を戻さなくてもよく、この方法は、終了してもよく、また新規の文書に関して再び開始してもよい。

【００４１】

ブロック２４０で、固有表現３１０に類似している単語３１５を得るために、固有表現３１０を追加文書３０５内の単語３１５と比較することができる。想定できるように、単語は、句でも、文章の断片でも、実体名でもよい。この方法は、コレクション$A_{s,t}$内の記事（Ｄｓ，Ｄｔ）の各ペア上で働くことができ、固有表現の翻字の相当語句のセット$P_{s,t}$を生成する。$P_{s,t}$内の各ペア（ε_S，e_T）は、言語Ｓの固有表現ε_S３１０および言語Ｔのトークンe_T３１５から構成され、これらは互いの翻字の相当語句である。さらに、翻字の類似度モデルＭＴによって測定される、ε_S３１０とe_T３１５の間の翻字の類似度は、少なくともβ≧０とすることができる。

【００４２】

擬似コードでは、この方法の一実施形態を以下のように進めることができる。

【００４３】

Input:
Set A_{ST} of similar documents (D_S, D_T) in languages (S,T)
Transliteration Similarity Model MT for (S,T)
Threshold score ß.
Output: Set $P_{S,T}$ of NETEs (s_S, s_T) from A_{ST};
1 $P_{ST} \leftarrow \emptyset$;
2 for each pair of articles (D_S, D_T) in A_{ST} do
3 for each named entity s_T in D_s do
4 $Y_S \leftarrow \emptyset$; // Set of candidates for s_S
5 for each candidate e_T in D_T do
6 score = TransliterationSimilarity(s_S,e_T,MT);
7 if (score ≥ ß) then $Y_S \leftarrow Y_S \cup (c_T, score)$;
8 end
9 s_T = BestScoringCandidate(Y_S);
10 if ($s_T \neq$ null) then $P_{ST} \leftarrow P_{ST} \cup (s_S, e_T)$;
11 end
12 end

【００４４】
翻字の類似度モデルは、ソースの固有表現３１０とターゲット言語の単語３１５の間の翻字の同等性の度合いを測定する。以下のように、翻字の類似度モデルＭＴとして、ロジスティック関数を利用してもよい。

【００４５】

$$TransliterationSimiliarity(\varepsilon_S, e_T, MT) = \frac{1}{1+e^{-w\cdot\phi(s_s, e_T)}}$$

【００４６】
ここで、φ（εS，eT）は、ペア（εS，eT）についての特徴ベクトルであり、ｗは、重みベクトルである。翻字の類似度は、［０．．１］の範囲の値をとるとし得る。このモデルで利用されている特徴は、ある種の文字配列の出現、εSとeTのサブストリングの結合、文字の配列の単調性、および２つのストリング内の文字数の差など、（εS，eT）において観察される関心を引く言語横断の関連性を取り込んでもよい。重みベクトルｗは、既知の翻字の相当語句のトレーニング・コーパスにわたって、識別しながら学習される。もちろん、固有表現３１０と単語３１５の類似度を判定する他の方式が可能でもよく、また企図されている。

【００４７】
実施形態によっては、追加文書３０５内のすべての単語３１５を、文書３００の固有表現３１０と比較する。他の実施形態では、追加文書３０５を詳しく調べて、多くの用語を分析から排除する。たとえば、「ｔｈｅ」、「ａ」、「ａｎ」など英語の冠詞は、固有表現３１０の一部である可能性は非常に低いので、これらの単語は分析しなくてもよい。さらに、固有表現３１０内に動詞が存在する可能性は低いので、動詞は分析しなくてもよい。さらに他の例として、形容詞が固有表現３１０の一部である可能性は低いので、形容詞も分析しなくてもよい。追加文書３０５内の単語３１５をさらにふるい分けることが可能であり、また企図されている。その結果、固有表現３１０と比較される、追加文書３０５内の単語３１５の数は、極めて少ない可能性があり、まさに対象とすべきである。

【００４８】
ブロック２５０で、固有表現３１０に類似している単語３１５が検出される場合、固有表現３１０および類似している単語３１５を、固有表現の翻字として格納することができる。複数の単語３１５が固有表現と比較された場合、最も高いと判定された類似度を有する単語３１５／固有表現３１０のペアを、固有表現の翻字として選択することができる。

いずれのペアも十分な類似度スコアに達しない場合、いずれの単語３１５も、固有表現３１０の翻字として戻さないとし得る。

【００４９】

次いで、この翻字を様々な目的に使用してもよい。一実施形態では、翻訳ソフトウェアがこの翻字を使用して、翻訳を改善してもよい。他の実施形態では、この翻字を探索ソフトウェアで使用して、複数の言語の関連する結果を探索する助けとし得る。もちろん、他の使用法も可能であり、またそのように意図されている。

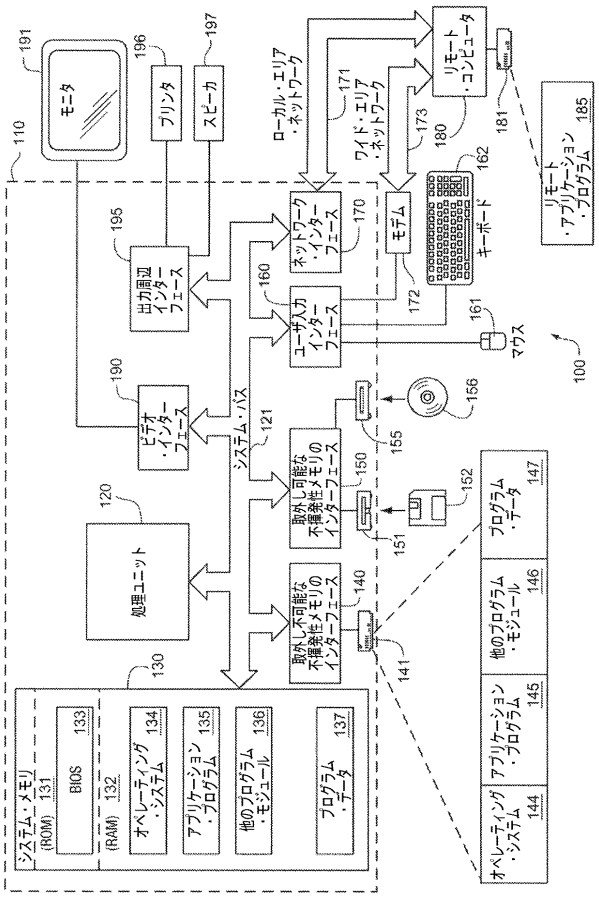
【００５０】

前述の説明では、数多くの様々な実施形態の詳細な説明を行っているが、この特許の範囲は、この特許の最後で述べる特許請求の範囲の説明によって規定されるものと理解すべきである。あらゆる実現可能な実施形態を説明することは、不可能ではないとしても、実際的ではないはずなので、詳細な説明は例示的なものに過ぎないと解釈すべきであり、あらゆる実現可能な実施形態を説明するものではない。現行技術、またはこの発明の出願日以降に開発される技術のいずれかを使用して、数多くの代替実施形態を実施することもできるが、こうした技術もやはり、特許請求の範囲に記載の範囲内に収まるものとする。

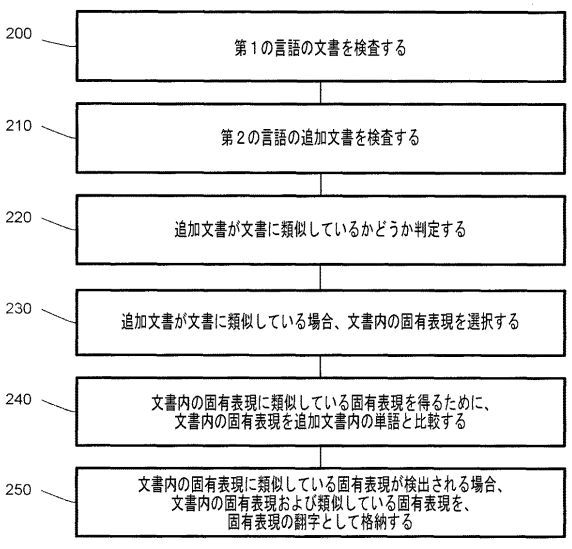
【００５１】

このように、この特許請求の範囲に記載の精神および範囲から逸脱することなく、本明細書において説明し図示した技法および構造物において、多くの修正形態および変形形態を実施してもよい。したがって、本明細書において記載された方法および装置は、例示的なものに過ぎず、特許請求の範囲に記載の範囲を限定するものではないことを理解されたい。

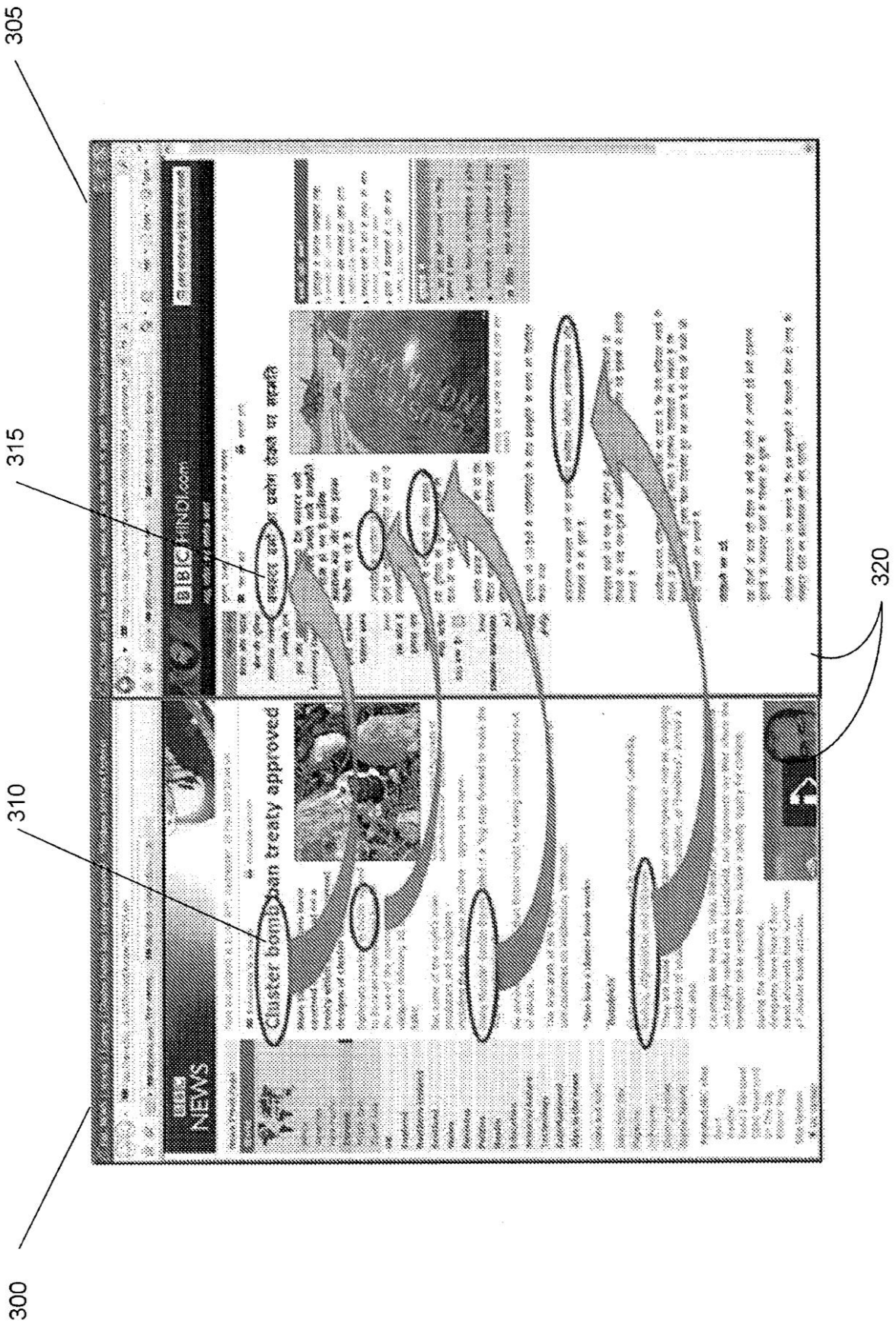
【図１】



【図２】



【図３】


305
315
310
320
300
BBC NEWS
BBC HINDI.com
Cluster bomb ban treaty approved

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2009/061352

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/28(2006.01)i, G06F 17/21(2006.01)i

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)
G06F 17/28

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models
(Chinese Patents and application for patent)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: entity and document and (translate or interpret) and (pronunciation or phonetic or phoneme)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	AHMED HASSAN et al., "IMPROVING NAMED ENTITY TRANSLATION BY EXPLOITING COMPARABLE AND PARALLEL CORPORA", Conference on Recent Advances in Natural Language Processing (RANLP 2007), AMML Workshop, 2007, pp.1-6. See abstract; chapter 3; figs.2-3. (http://www.computing.dcu.ie/~hhasan/ranlp07.pdf)	1-2,4-5,11,13,18-19
A	RICHARD SPROAT et al., "NAMED ENTITY TRANSLITERATION WITH COMPARABLE CORPORA", Proceeding of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, July 2006, pp.73-80. See abstract; chapter 3 (3.1-3.2). (http://www.mt-archive.info/Coling-ACL-2006-Sproat.pdf)	1-20

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* 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 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
03 JUNE 2010 (03.06.2010)	**03 JUNE 2010 (03.06.2010)**

Name and mailing address of the ISA/KR
Korean Intellectual Property Office
Government Complex-Daejeon, 139 Seonsa-ro, Seo-gu, Daejeon 302-701, Republic of Korea
Facsimile No. 82-42-472-7140

Authorized officer
KIM, SANG CHEOL
Telephone No. 82-42-481-8521

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2009/061352

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DAN GOLDWASSER et al., "ACTIVE SAMPLE SELECTION FOR NAMED ENTITY TRANSLITERATION", Proceeding of ACL-08: HLT, Short papers, June 2008, pp.53-56. See abstract; chapter 2; fig.2. (http://www.aclweb.org/anthology/P/P08/P08-2014.pdf)	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2009/061352

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
None			

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(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,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,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

(72)発明者 サラバナン クリシュナン
アメリカ合衆国 ９８０５２－６３９９ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション エルシーエー－インターナショナル パテンツ内
(72)発明者 アルムガン クマラン
アメリカ合衆国 ９８０５２－６３９９ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション エルシーエー－インターナショナル パテンツ内

Ｆターム(参考) 5B091 AA01 BA02 CC15 EA02