(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5317061号
(P5317061)

(45)発行日　平成25年10月16日(2013.10.16)　(24)登録日　平成25年7月19日(2013.7.19)

(51)Int.Cl.　　F I
G06F 17/27 (2006.01)　　G06F 17/27 Z
G06F 17/30 (2006.01)　　G06F 17/30 １７０Ａ
G06F 17/30 ２２０Ｚ
G06F 17/30 ２１０Ｄ

請求項の数 8　(全 22 頁)

(21)出願番号　特願2009-177488 (P2009-177488)
(22)出願日　平成21年7月30日 (2009.7.30)
(65)公開番号　特開2011-34171 (P2011-34171A)
(43)公開日　平成23年2月17日 (2011.2.17)
審査請求日　平成24年6月19日 (2012.6.19)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　呉　鍾勲
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　内元　清貴
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　鳥澤　健太郎
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】単語間の意味的関係の有無についての、複数言語での同時分類器及びそのためのコンピュータプログラム。

(57)【特許請求の範囲】
【請求項１】
第１の言語の単語の対の間の所定の意味的関係の有無を判定し、信頼度を示すスコアとともに判定結果を出力する第１の分類器と、第２の言語の単語の対の間の前記意味的関係の有無を判定し、信頼度を示すスコアとともに出力する第２の分類器とを同時に機械学習により学習させるための同時学習装置であって、
前記第１及び第２の言語の分類器の学習のための学習データを記憶するための第１及び第２の学習データ記憶手段と、
前記第１及び第２の学習データ記憶手段に追加される候補となる第１及び第２の学習データ候補をそれぞれ記憶するための第１及び第２の候補記憶手段と、
前記第１及び第２の学習データ記憶手段に記憶された学習データを用いて前記第１及び第２の分類器の学習をそれぞれ行なうための第１及び第２の学習手段と、
前記第１及び第２の分類器を用いて、前記第１及び第２の候補記憶手段に記憶された前記第１及び第２の学習データ候補をそれぞれ分類させ、分類結果とスコアとをそれぞれ出力させるための第１及び第２の分類手段と、
前記第１及び第２の候補記憶手段に記憶された前記第１及び第２の学習データ候補をそれぞれ前記第２及び第１の言語に翻訳するための第１及び第２の翻訳手段と、
前記第１及び第２の翻訳手段によりそれぞれ翻訳された後の前記第２及び前記第１の言語の学習候補とを、前記第２及び第１の分類器を用いてそれぞれ分類させ、分類結果とスコアとをそれぞれ出力させるための第３及び第４の分類手段と、

前記第１の分類手段による分類結果及びスコアと、前記第３の分類手段による分類結果及びスコアとに基づいて、前記第１の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、前記第１の分類手段による分類結果とともに前記第２の学習データ記憶手段に追加するための第１の更新手段と、

前記第２の分類手段による分類結果及びスコアと、前記第４の分類手段による分類結果及びスコアとに基づいて、前記第２の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、前記第２の分類手段による分類結果とともに前記第１の学習データ記憶手段に追加するための第２の更新手段と、

前記第１及び第２の学習手段、前記第１及び第２の分類手段、前記第１及び第２の翻訳手段、前記第３及び第４の分類手段、ならびに前記第１及び第２の更新手段による処理を、所定の終了条件が成立するまで繰返させるための繰返し制御手段とを含む、同時学習装置。

【請求項２】

前記第１の更新手段は、

前記第１の翻訳手段による翻訳結果のうち、前記第１の分類手段によるスコアが所定の第１のしきい値以上の学習データに対する翻訳結果で、かつ前記第３の分類手段によるスコアが所定の第２のしきい値未満であるものを、前記第１の分類手段による分類結果とともに、前記第２の学習データ記憶手段に追加するための手段と、

前記第１の翻訳手段による翻訳結果のうち、前記第１の分類手段によるスコアが前記第１のしきい値以上の学習データに対する翻訳結果で、かつ前記第３の分類手段によるスコアが前記第２のしきい値以上であって、かつ前記第１及び第３の分類手段による分類結果が一致するものを、前記第１の分類手段による分類結果とともに、前記第２の学習データ記憶手段に追加するための手段とを含む、請求項１に記載の同時学習装置。

【請求項３】

前記第２の更新手段は、

前記第２の翻訳手段による翻訳結果のうち、前記第２の分類手段によるスコアが所定の第３のしきい値以上の学習データに対する翻訳結果で、かつ前記第４の分類手段によるスコアが所定の第４のしきい値未満であるものを、前記第２の分類手段による分類結果とともに、前記第１の学習データ記憶手段に追加するための手段と、

前記第２の翻訳手段による翻訳結果のうち、前記第２の分類手段によるスコアが前記第３のしきい値以上の学習データに対する翻訳結果で、かつ前記第４の分類手段によるスコアが前記第４のしきい値以上であって、かつ前記第２及び第４の分類手段による分類結果が一致するものを、前記第２の分類手段による分類結果とともに、前記第１の学習データ記憶手段に追加するための手段とを含む、請求項２に記載の同時学習装置。

【請求項４】

前記第１及び第２の分類器は、互いに同じ種類の機械学習モデルにより実現される、請求項１－請求項３のいずれかに記載の同時学習装置。

【請求項５】

前記第１及び第２の分類器は、互いに異なる種類の機械学習モデルにより実現される、請求項１－請求項３のいずれかに記載の同時学習装置。

【請求項６】

前記第１及び第２の言語は互いに異なる、請求項１－請求項６のいずれかに記載の同時学習装置。

【請求項７】

コンピュータにより実行されると、当該コンピュータを、請求項１－請求項６のいずれかに記載の同時学習装置として動作させる、コンピュータプログラム。

【請求項８】

請求項７に記載のコンピュータプログラムを記録した、コンピュータ読取可能な記録媒体。

【発明の詳細な説明】

【技術分野】
【０００１】
この発明は自然言語処理に関し、特に、単語間の意味的関係を精度よく獲得するための技術に関する。
【背景技術】
【０００２】
コンピュータを用いた情報処理技術、特に自然言語処理では、意味的知識をどのようにして獲得し集積するかに関する技術が必須である。たとえば質問に対する自動応答処理などにおいては、意味的関係を知ることは決定的に重要である。これ以外にも意味的知識が重要な役割を果たすことが多い。
【０００３】
たとえば、キーワードを用いた情報検索では、入力された単語の上位概念に相当する単語まで含めて検索が行なわれる場合がある。こうした場合、あらかじめ単語の上位下位（包摂）関係を記述した辞書（シソーラス）を準備しておく必要がある。シソーラスを手作業で準備してもよいが、現代のように変化の激しい社会では、意味の包摂関係を含めた言語に関する情勢の変化も速く、手作業ではそうした変化を辞書に的確に反映させることは事実上不可能である。そこで、自然言語処理技術を用い、そうしたシソーラスを自動的に、かつ精度高く作成する技術が求められている。
【０００４】
こうした要求は、単語の包摂関係にとどまらず、類語関係、症状とその原因、問題とその予防、問題とその対策、全体と部分、原因と結果など、語彙の間の意味的関係を用いる技術全般についても存在している。
【０００５】
語彙の意味的関係の自動的な獲得は、従来、任意の単語のペアに対し、ある特定の意味的関係があるか否かを二値分類するタスクとして扱われることが多い。二値分類のタスクには、教師あり学習がよく採用され、効果を挙げている。
【０００６】
図１に、後掲の非特許文献１に記載の、従来の意味的関係の分類システム３０の概略ブロック図を示す。図１を参照して、この分類システム３０は、たとえば日本語の２つの単語間に包摂関係があるか否かを判定するための、ＳＶＭ（Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ）、ＣＲＦ（Ｃｏｎｄｉｔｉｏｎａｌ　Ｒａｎｄｏｍ　Ｆｉｅｌｄｓ）又はＭＥＭ（Ｍａｘｉｍｕｍ　Ｅｎｔｒｏｐｙ　Ｍｏｄｅｌ）などの、機械学習による確率モデルを用いた分類器４４と、分類器４４の学習を行なうために、日本語の単語対と、それら単語対の間に包摂関係があるか否かを示すラベルとからなる学習データ４０を多数記憶するための記憶装置と、この学習データ４０を用いて分類器４４の機械学習を行なうための機械学習部４２とを含む。学習データ４０を用いて分類器４４の学習を行なうことにより、日本語の単語対４６が与えられると、分類器４４はこの単語の間に上記した意味的関係（包摂関係）が存在するか否かを示すラベル（真又は偽）と、その結果の信頼度を示すスコアとを出力する。信頼度としては、たとえばＳＶＭの場合には分類の境界となる超平面から、入力された単語対を示す点までの距離を用いることができる。一般的に機械学習モデルを分類器として用いる場合には、確率又はそれと等価な形でスコアが出力されるので、そのスコアを信頼度として用いることができる。
【先行技術文献】
【非特許文献】
【０００７】
【非特許文献１】ロクサナ・ガージュ他、２００７年、Ｓｅｍｅｖａｌ－２００７タスク０４：名詞類間の意味的関係の分類、第４回意味的評価に関する第４回国際ワークショップ予稿集（ＳｅｍＥｖａｌ－２００７），ｐｐ．１３－１８（Roxana Girju et al. 2007. Semeval-2007 task 04: Classification of Semantic relations between nominals. In Proceeding of the Fourth International Workshop on Seman

tic Evaluations (SemEval-2007), paes 13-18）

【発明の概要】

【発明が解決しようとする課題】

【０００８】

非特許文献１に記載されたような教師あり学習では、分類器の性能を高めるためには大量の学習データが必要である。学習データには正解のラベルを手作業で付す必要がある。そのため大量の学習データの準備に高いコストがかかるという問題がある。これは日本語だけではなく、英語又は他の言語における意味的知識の獲得においても直面する問題である。

【０００９】

それゆえに本発明の目的は、低コストで、言語にかかわらず意味的知識を効率よく分類できる分類器、及びコンピュータでそうした装置を実現することができるコンピュータプログラムを提供することである。

【００１０】

本発明の他の目的は、学習データの準備にかかる人手を削減しながら、言語にかかわらず意味的知識を効率よく分類できる分類器、及びコンピュータでそうした装置を実現することができるコンピュータプログラムを提供することである。

【００１１】

本発明のさらに他の目的は、学習データの準備にかかる人手を削減しながら、言語にかかわらず信頼性の高い学習データを集積して分類器の学習を行なうことが可能な分類器、及びコンピュータでそうした装置を実現することができるコンピュータプログラムを提供することである。

【課題を解決するための手段】

【００１２】

本発明の第１の局面に係る同時学習装置は、第１の言語の単語の対の間の所定の意味的関係の有無を判定し、信頼度を示すスコアとともに判定結果を出力する第１の分類器と、第２の言語の単語の対の間の意味的関係の有無を判定し、信頼度を示すスコアとともに出力する第２の分類器とを同時に機械学習により学習させるための同時学習装置であって、第１及び第２の言語の分類器の学習のための学習データを記憶するための第１及び第２の学習データ記憶手段と、第１及び第２の学習データ記憶手段に追加される候補となる第１及び第２の学習データ候補をそれぞれ記憶するための第１及び第２の候補記憶手段と、第１及び第２の学習データ記憶手段に記憶された学習データを用いて第１及び第２の分類器の学習をそれぞれ行なうための第１及び第２の学習手段と、第１及び第２の分類器を用いて、第１及び第２の候補記憶手段に記憶された第１及び第２の学習データ候補をそれぞれ分類させ、分類結果とスコアとをそれぞれ出力させるための第１及び第２の分類手段と、第１及び第２の候補記憶手段に記憶された第１及び第２の学習データ候補をそれぞれ第２及び第１の言語に翻訳するための第１及び第２の翻訳手段と、第１及び第２の翻訳手段によりそれぞれ翻訳された後の第２及び第１の言語の学習候補を、第２及び第１の分類器を用いてそれぞれ分類させ、分類結果とスコアとをそれぞれ出力させるための第３及び第４の分類手段と、第１の分類手段による分類結果及びスコアと、第３の分類手段による分類結果及びスコアとに基づいて、第１の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、第１の分類手段による分類結果とともに第２の学習データ記憶手段に追加するための第１の更新手段と、第２の分類手段による分類結果及びスコアと、第４の分類手段による分類結果及びスコアとに基づいて、第２の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、第２の分類手段による分類結果とともに第１の学習データ記憶手段に追加するための第２の更新手段と、第１及び第２の学習手段、第１及び第２の分類手段、第１及び第２の翻訳手段、第３及び第４の分類手段、ならびに第１及び第２の更新手段による処理を、所定の終了条件が成立するまで繰返させるための繰返し制御手段とを含む。

【００１３】

予め第１及び第２の学習データ記憶手段に、それぞれ第１及び第２の分類器を学習させるための学習データを記憶させておく。これら学習データに追加される候補を、第１及び第２の候補記憶装置に記憶させておく。第１および第２の分類手段は、第１及び第２の分類器で第１及び第２の候補記憶手段に記憶された候補を分類させ、分類結果とスコアとを出力させる。第１及び第２の翻訳手段は、第１及び第２の分類手段により分類された候補をそれぞれ第２及び第１の言語に翻訳する。第３及び第４の分類手段は、翻訳結果の第２及び第１の言語の候補をそれぞれ第２及び第１の分類器を用いて分類させ、分類結果とスコアとを出力させる。第１の更新手段は、第１の分類手段による分類結果及びスコアと、第３の分類手段による分類結果及びスコアとに基づいて、第１の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、第１の分類手段による分類結果とともに第２の学習データ記憶手段に追加する。第２の更新手段は、第２の分類手段による分類結果及びスコアと、第４の分類手段による分類結果及びスコアとに基づいて、第２の翻訳手段による翻訳結果のうち、所定の条件を充足するものを選択し、第２の分類手段による分類結果とともに第１の学習データ記憶手段に追加する。繰返し制御手段の制御にしたがい、第１及び第２の学習手段、第１及び第２の分類手段、第１及び第２の翻訳手段、第３及び第４の分類手段、ならびに第１及び第２の更新手段による処理が所定の終了条件が成立するまで繰返される。

【００１４】

このような構成により、第１の学習データ記憶手段に記憶される第１の言語の分類器のための学習データと、第２の学習データ記憶手段に記憶される第２の言語の分類器のための学習データとが追加される。第１の言語において意味的関係の有無が明確な単語対であっても、第２の言語では意味的関係が不明な場合がある。そうしたときでも、第１の言語の単語対を第２の言語に翻訳すると、得られた第２の言語の単語対の間に意味的関係が存在することが第１の言語側の情報から判明する。逆の場合も同様である。したがって、このように第１及び第２の言語の分類器を同時学習させることにより、それぞれの言語の学習データが効率よく、しかも精度高く集積でき、分類器の精度も高まる。学習データについて、多大な労力をかける必要はない。その結果、低コストで、言語にかかわらず意味的知識を効率よく分類できる分類器を提供できる。

【００１５】

好ましくは、第１の更新手段は、第１の翻訳手段による翻訳結果のうち、第１の分類手段によるスコアが所定の第１のしきい値以上の学習データに対する翻訳結果で、かつ第３の分類手段によるスコアが所定の第２のしきい値未満であるものを、第１の分類手段による分類結果とともに、第２の学習データ記憶手段に追加するための手段と、第１の翻訳手段による翻訳結果のうち、第１の分類手段によるスコアが第１のしきい値以上の学習データに対する翻訳結果で、かつ第３の分類手段によるスコアが第２のしきい値以上であって、かつ第１及び第３の分類手段による分類結果が一致するものを、第１の分類手段による分類結果とともに、第２の学習データ記憶手段に追加するための手段とを含む。

【００１６】

この構成により、第１の言語の候補についての第１の分類手段による分類結果のスコアが第１のしきい値以上であり、かつその候補を翻訳したものの第２の分類手段による分類のスコアが第２のしきい値未満の場合には、第１の分類手段によるスコアを信頼して翻訳後の候補が第２の言語の学習データに追加される。第１の分類手段による分類結果のスコアと、第２の分類手段によるスコアとがともにしきい値以上の場合には、両者の分類結果が一致しているときのみ、第２の言語の学習データに候補が追加される。分類結果がコンフリクトしているときにはその候補は追加されない。そのため、第２の言語の学習データには、分類結果の信頼性の高いもののみが集積されていく。この間に、人手で分類を行なったり、分類結果による候補の取捨選択を行なったりする必要はない。その結果、学習データの準備にかかる人手を削減しながら、言語にかかわらず意味的知識を効率よく分類できる分類器を提供できる。

【００１７】

より好ましくは、第２の更新手段は、第２の翻訳手段による翻訳結果のうち、第２の分類手段によるスコアが所定の第３のしきい値以上の学習データに対する翻訳結果で、かつ第４の分類手段によるスコアが所定の第４のしきい値未満であるものを、第２の分類手段による分類結果とともに、第１の学習データ記憶手段に追加するための手段と、第２の翻訳手段による翻訳結果のうち、第２の分類手段によるスコアが第３のしきい値以上の学習データに対する翻訳結果で、かつ第４の分類手段によるスコアが第４のしきい値以上であって、かつ第２及び第４の分類手段による分類結果が一致するものを、第２の分類手段による分類結果とともに、第１の学習データ記憶手段に追加するための手段とを含む。

【００１８】

第１及び第２の分類器は、互いに同じ種類の機械学習モデルにより実現されてもよいし、互いに異なる種類の機械学習モデルにより実現されてもよい。

【００１９】

好ましくは、第１及び第２の言語は互いに異なっている。

【００２０】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを、上記したいずれかの同時学習装置として動作させる。したがってこのコンピュータプログラムをコンピュータに実行させることにより、上記した同時学習装置により得られるものと同じ効果を得ることができる。

本発明の第３の局面に係る記録媒体は、このコンピュータプログラムを記録したものである。

【図面の簡単な説明】

【００２１】

【図１】従来の分類システム３０の概略ブロック図である。

【図２】本発明の一実施の形態に係る、日本語と英語との分類器の同時学習の概略を説明するための図である。

【図３】本発明の一実施の形態に係る日本語と英語との分類器の同時学習装置９０の概略ブロック図である。

【図４】図３に示す日本語・英語同時学習部１１６のより詳細なブロック図である。

【図５】（Ａ）は英語の初期学習データの例を示す図であり、（Ｂ）は日本語の初期学習データの例を示す図である。

【図６】日本語・英語同時学習部１１６をコンピュータで実現するためのコンピュータプログラムの制御構造を示すフローチャートである。

【図７】日本語の学習データの更新処理を実現するプログラムの制御構造を示すフローチャートである。

【図８】英語の学習データの更新処理を実現するプログラムの制御構造を示すフローチャートである。

【図９】英語のＷｉｋｉｐｅｄｉａの記載から包摂関係の単語対の候補を抽出する処理を説明するための図である。

【図１０】本発明の一実施の形態に係る分類器の同時学習装置９０を実現するためのコンピュータシステム５５０の外観を示す図である。

【図１１】図１０に示すコンピュータシステム５５０のハードウェア構成を示すブロック図である。

【図１２】実験における学習データサイズとＦ１値との関係を示すグラフである。

【発明を実施するための形態】

【００２２】

以下の説明では、同一の部品には同一の参照番号を付してある。それらの名称及び機能も同一である。したがって、それらについての詳細な説明は繰返さない。

【００２３】

＜基本的考え方＞

以下に説明する本実施の形態による学習方法は、以下のような考え方に基づくものであ

る。すなわち、ある量の第１の言語の学習データ及び第２の言語の学習データが予め存在するものとする。この第１の言語の学習データを別の第２の言語の学習データに翻訳し、第２の言語の学習データに追加することができれば、第２の言語の学習データを低コストに拡張することができる。逆に、第２の言語の学習データを第１の言語に翻訳することで、第１の言語の学習データを拡張することができる。

【００２４】

さらに、たとえばある学習データで学習済の第２の言語の分類器による分類を、第２の言語の単語対に対して適用することで、それら単語対の間に包摂関係があるか否かについての分類結果を得ることができる。この分類結果については、信頼性の比較的低いものから高いものまで存在しうる。そこで、信頼性の高い分類結果が得られた単語対を第１の言語に翻訳することで、第１の言語の学習データをさらに拡張することができる可能性がある。

【００２５】

異なる言語では、分類器のための素性（特徴量）としては異なるものが用いられることが通常である。したがって、第１の言語の分類器では信頼がおけないような結果しか得られない単語対であっても、対応する第２の言語の単語対を第２の言語の分類器に適用すると、信頼性の高い結果が得られるという場合もあり得るであろう。そうした場合、第２の言語の単語対を第１の言語に翻訳することで、第１の言語の学習データを拡張することができる。逆に第１の言語の分類結果から、第２の言語の学習データを拡張することも可能と考えられる。

【００２６】

こうして、第１の言語と第２の言語とを互いに入れ替えながら双方の言語の学習データを拡張していくことにより、双方の学習データを効率よく拡張でき、その結果、そうした学習データにより学習が行なわれた分類器の精度を高めることができる。このように、同種だが内容において異なる２つのタスクの確率モデルを互いの学習結果を用いて学習していくことを、英語ではｃｏ－ｔｒａｉｎｉｎｇと呼び、日本語では「同時学習」又は「共学習」と呼ぶ。

【００２７】

最初に、何らかの方法により予め日本語用学習データと、英語用学習データとを準備する。これら学習データの構成については図５を参照して後述するが、日本語の場合には、学習データは、任意の日本語の単語対と、それらが包摂関係にあるか否かを示すラベルとからなる。なお、単語対には順序があり、第１の単語が第２の単語の上位にあるか否かがラベルにより示されている。

【００２８】

以下の説明では、第１の言語として英語を、第２の言語として日本語を、それぞれ想定する。

【００２９】

図２は、本発明の一実施の形態に係る、日本語と英語との分類器の同時学習の基本的考え方を説明するための図である。図２を参照して、この実施の形態に係る分類器の同時学習では、日本語の包摂関係の分類器と、英語の包摂関係の分類器との同時学習を行なうものとする。また、本実施の形態では分類器としてＳＶＭを使用し、分類時のスコアとしてはＳＶＭの分類の境界を定める超平面から、超空間内で単語対を表す点までの距離を用いるものとする。

【００３０】

まず、日本語用学習データ６０を用いて日本語用分類器６４の学習を行なう。図示していない、学習データ追加候補である日本語の単語対の集合に対して日本語用分類器６４による分類を適用し、分類結果６８を得る。同様に、英語用学習データ６２を用いて英語用分類器６６の学習を行なう。図示していない、学習データ追加候補の英語の単語対の集合に対して英語用分類器６６による分類を適用し、分類結果７０を得る。

【００３１】

こうして得られた分類結果６８のうち、スコアが高いもの（分類結果の信頼性が高いもの）を、日英翻訳用の辞書を用いて英語の単語対に翻訳し、分類結果とともに英語用学習データ６２に追加することで、拡張した英語用学習データ７４が得られる。同様に、分類結果７０のうち、スコアが高いものを、英日翻訳用の辞書を用いて日本語の単語対に翻訳し、分類結果とともに日本語用学習データ６０に追加することで、拡張した日本語用学習データ７２が得られる。

【００３２】

こうして拡張した日本語用学習データ７２及び英語用学習データ７４は、初期の日本語用学習データ６０及び英語用学習データ６２には存在していなかった学習データを含む。しかもそれらに付されている、分類結果を示すラベルの信頼性は高い。その結果、拡張した日本語用学習データ７２及び拡張した英語用学習データ７４をそれぞれ使用して新たに日本語用分類器７６及び英語用分類器７８の学習を行なうことにより、日本語用分類器７６及び英語用分類器７８の精度は日本語用分類器６４及び英語用分類器６６より高くなることが期待される。さらにこれを繰返すことで、分類器の精度はさらに向上する。実際、後述する実験により、こうした予測と一致する結果を得ることができた。

【００３３】

なお、ＳＶＭの学習時及び判定時の素性としては以下を用いる。ここでは、ｈｙｐｅｒが上位語を表し、ｈｙｐｏが下位語候補を表し、（ｈｙｐｅｒ、ｈｙｐｏ）により包摂関係候補を表すものとする。特徴量として、次のテーブル１に示すものを用いた。

【００３４】

【表１】

テーブル1

タイプ	説明	例
LF1	形態素／単語	hyper: tiger, hypo: Siberian*, hypo: tiger*
LF2	形態素／単語の品詞	hyper: NN*, hypo: NP, hypo: NN*
LF3	hyper及びhypo自体	hyper: Tiger, hypo: Siberian tiger
LF4	使用した語彙パターン	hyper: "List of X", hypo: "Notable X"
LF5	典型的セクション見出し	hyper: History, hypo: Reference
SF1	hyperとhypoとの距離	3
SF2	レイアウト項目のタイプ	hyper: title, hypo: bulleted list
SF3	ツリーノードのタイプ	hyper: ルート、hypo: リーフ
SF4	hypoの親ノードのLF1及びLF3	LF3: Subspecies
SF5	hyperの子ノードのLF1及びLF3	LF3: Taxonomy
IF	hyper及びhypoの意味的属性	hyper:(分類ボックス、種) hypo:(分類ボックス、名称)

上のテーブルのＬＦ１及びＬＦ２で「＊」で示したものは先頭の形態素／単語とその品詞とを示す。ＬＦ４及びＬＦ５を除き、例は後に示す図６から得られるものを示してある。

【００３５】

＜構成＞

図３は、本発明の一実施の形態に係る分類器の同時学習装置９０の概略構成を示すブロック図である。図３を参照して、分類器の同時学習装置９０は、英語版のＷｉｋｉｐｅｄｉａのページデータ１００をそのレイアウト情報とともに記憶した記憶装置と、英語版のＷｉｋｉｐｅｄｉａのページデータ１００に対応した日本語版のＷｉｋｉｐｅｄｉａのページデータ１０２をそのレイアウト情報とともに記憶した記憶装置と、英語版のＷｉｋｉｐｅｄｉａのページデータ１００及び日本語版のＷｉｋｉｐｅｄｉａのページデータ１０２の文及び単語の対応関係に基づいて、公知の方法によって英語と日本語との対訳辞書（翻訳辞書１１４）を作成する翻訳辞書作成部１１２とを含む。Ｗｉｋｉｐｅｄｉａのｉｎｆｏｂｏｘと呼ばれるテンプレートは、文章の主題を属性とその値という組合せからなる

テーブル形式で記述するものであり、本実施の形態ではこのｉｎｆｏｂｏｘの性格を利用して、学習データ候補の抽出を行なっている。

【００３６】

分類器の同時学習装置９０はさらに、英語版のＷｉｋｉｐｅｄｉａのページデータ１００から、任意の単語対を多数抽出し、英語の包摂関係語候補１０８として記憶装置に記憶させるための、英語の包摂関係語候補抽出部１０４を含む。包摂関係語候補抽出部１０４により抽出される単語対は、必ずしも包摂関係にあるとは限らないが、その中には包摂関係にあるような単語対も含まれるはずである。本実施の形態では、そうした単語対を学習データに追加していく。

【００３７】

分類器の同時学習装置９０はさらに、日本語版のＷｉｋｉｐｅｄｉａのページデータ１０２から、任意の単語対を多数抽出し、日本語の包摂関係語候補１１０として記憶装置に記憶させるための、日本語の包摂関係語候補抽出部１０６を含む。

【００３８】

分類器の同時学習装置９０はさらに、英語の包摂関係語候補１０８、日本語の包摂関係語候補１１０、及び翻訳辞書１１４を用い、英語と日本語の包摂関係の分類器の学習を同時に行なう日本語・英語同時学習部１１６を含む。

【００３９】

日本語・英語同時学習部１１６は、英語の学習データを記憶するための英語学習データ記憶部１３４と、英語の分類器１３０と、英語学習データ記憶部１３４に記憶された英語の学習データを用いて英語分類器１３０の学習を行なうための学習部１３２と、日本語の学習データを記憶するための日本語学習データ記憶部１４４と、日本語分類器１４０と、日本語学習データ記憶部１４４に記憶された日本語の学習データを用いて日本語分類器１４０の学習を行なうための学習部１４２と、英語分類器１３０による英語の包摂関係語候補１０８の分類結果、日本語分類器１４０による日本語の包摂関係語候補１１０の分類結果、及び翻訳辞書１１４を用い、図２を参照して説明した方法によって英語学習データ記憶部１３４及び日本語学習データ記憶部１４４の更新を繰返し行なうための学習データ更新部１５０とを含む。英語学習データ記憶部１３４及び日本語学習データ記憶部１４４には、処理に先立って英語及び日本語の初期学習データが記憶されるものとする。これら初期学習データは、たとえば手作業によって準備された比較的少量のものでよい。

【００４０】

図４は、図３に示す学習データ更新部１５０のより詳細なブロック図である。図４では、学習データ更新部１５０内部の構成要素の関係、及び学習データ更新部１５０内部の構成要素と外部との関係のみを示してある。図４を参照して、学習データ更新部１５０は、英語分類器１３０による英語の包摂関係語候補１０８（図３）の分類結果と、翻訳辞書１１４（図３）とを用いて、日本語学習データ記憶部１４４に記憶された日本語の学習データに新たな学習データを追加するための日本語学習部１６０と、日本語分類器１４０による日本語の包摂関係語候補１１０（図３）の分類結果と、翻訳辞書１１４（図３）とを用いて、英語学習データ記憶部１３４に記憶された英語の学習データに新たな学習データを追加するための英語学習部１６２と、日本語学習部１６０及び英語学習部１６２が新たな学習データの選択の際に使用する信頼度のしきい値θを記憶するための記憶部１６４とを含む。

【００４１】

日本語学習部１６０は、英語分類器１３０により出力された英語対の分類結果（翻訳後の日本語学習データへの追加候補１８０となる。）のうち、信頼度が上位の所定個に入り、かつ英語学習データ記憶部１３４に記憶されておらず、かつその分類結果の信頼度が記憶部１６４に記憶されたしきい値θ以上のもののみを選択し選択結果１８４として出力する選択部１８２と、選択結果１８４内の英語の単語対の各々に対して、翻訳辞書１１４を用いて日本語の単語対への翻訳を行ない、翻訳辞書１１４に存在する訳語が見出された単語対のみを翻訳結果１８８として出力する英日翻訳部１８６とを含む。翻訳結果１８８内

の日本語の単語対の各々に対して日本語分類器１４０が分類を実行し、分類（真又は偽）のラベルがその信頼度とともに付された分類結果１９０を出力する。日本語学習部１６０はさらに、分類結果１９０内の日本語の単語対の各々について、追加候補１８０のうち対応する英語の単語対に付された信頼度がしきい値θ以上であり、かつ「分類結果１９０に付されたしきい値がθ未満である」及び「日本語分類器１４０による分類結果のラベルが追加候補１８０のうち対応する英語の単語対に付されたラベルと一致するとき」という条件のいずれか一方が充足されたときのみ、その日本語の単語対を選択し、追加候補１８０で対応する英語の単語対に付されたラベルとともに選択結果１９４として出力する選択部１９２と、選択結果１９４を日本語学習データ記憶部１４４に新たな学習データとして追加することにより日本語学習データを更新する更新部１９６とを含む。

【００４２】

ここで、「追加候補１８０のうち対応する英語の単語対に付された信頼度がしきい値θ以上」、かつ「分類結果１９０に付されたしきい値がθ未満である」という条件は、日本語の分類器では分類の信頼度が低いが、英語の分類器による分類の信頼度が高い、ということを意味する。このような条件を充足する場合、英語分類器１３０による分類結果と日本語分類器１４０による分類結果とが矛盾していても、英語分類器１３０による分類結果にしたがって、それらを翻訳した日本語の単語対を日本語学習データ記憶部１４４に追加すると、日本語のみによる処理では抽出できない日本語学習データを抽出することができると考えられる。一方、「追加候補１８０のうち対応する英語の単語対に付された信頼度がしきい値θ以上」、かつ「日本語分類器１４０による分類結果のラベルが追加候補１８０のうち対応する英語の単語対に付されたラベルと一致するとき」という条件は、追加候補１８０の判定結果と、日本語分類器１４０による判定結果とがコンフリクトする場合を排除するための条件である。両者の判定結果が互いに矛盾し、かつ両者の信頼度がしきい値以上の場合には、その単語対は学習データとしては採用しない。両者の判定結果がコンフリクトしない場合のみ、学習データを採用する。

【００４３】

同様に、英語学習部１６２は、日本語分類器１４０により出力された日本語対の分類結果（翻訳後の英語学習データへの追加候補２１０となる。）のうち、信頼度が上位所定個に入り、かつ日本語学習データ記憶部１４４に記憶されておらず、かつその分類結果の信頼度が記憶部１６４に記憶されたしきい値θ以上のもののみを選択し選択結果２１４として出力する選択部２１２と、選択結果２１４内の日本語の単語対の各々に対して、翻訳辞書１１４を用いて英語の単語対への翻訳を行ない、翻訳辞書１１４に存在する訳語が見出された単語対のみを翻訳結果２１８として出力する日英翻訳部２１６とを含む。翻訳結果２１８内の英語の単語対の各々に対して英語分類器１３０が分類を実行し、分類（真又は偽）のラベルがその信頼度とともに付された分類結果２２０を出力する。英語学習部１６２はさらに、分類結果２２０内の英語の単語対の各々について、追加候補２１０のうち対応する日本語の単語対に付された信頼度がしきい値θ以上であり、かつ「分類結果２２０に付されたしきい値がθ未満である」及び「英語分類器１３０による分類結果のラベルが追加候補２１０のうち対応する日本語の単語対に付されたラベルと一致するとき」という条件のいずれか一方が充足されたときのみ、その英語の単語対を選択し、追加候補２１０で対応する日本語の単語対に付されたラベルとともに選択結果２２４として出力する選択部２２２と、選択結果２２４を英語学習データ記憶部１３４に新たな学習データとして追加することにより英語学習データ記憶部１３４を更新する更新部２２６とを含む。

【００４４】

ここでの抽出条件も日本語の学習データの更新の場合と同様である。

【００４５】

図５（Ａ）は、図３に示す英語学習データ記憶部１３４に記憶される初期データの一例であり、図５（Ｂ）は日本語学習データ記憶部１４４の初期データの一例である。図５（Ａ）に示すように、英語学習データ記憶部１３４に記憶される初期データは、英語の単語対と、その単語対のうち前者が後者の上位語であるか否かを示す分類ラベルとの組からな

る。たとえば「Ｅｎｚｙｍｅ」（酵素）と「Ｈｉｓｔｏｒｙ　ｏｆ　ｂｉｏｃｈｅｍｉｓｔｒｙ」（生化学の歴史）という単語（名詞類）の対は無関係なのでそのラベルは「×」（偽）であり、「ｄｏｇ」（犬）と「Ａｋｂａｓｈ　Ｄｏｇ」（アクバシュ犬）という単語（名詞類）の対は上位下位の関係にあるのでそのラベルは「○」（真）である。同様に、図５（Ｂ）に示すように、日本語の「酵素」という単語と「酸化還元酵素」という単語とは上位下位の関係にあるのでそのラベルは「○」、「酵素」という単語と「歴史」という単語とは上位下位の関係にはないので、そのラベルは「×」である。このように、○のラベルを持つ学習データと×のラベルを持つ学習データとを予め手作業などにより準備しておく。

【００４６】

たとえば、英単語の対「Ｅｎｚｙｍｅ」と「ｏｘｙｒｅｄｕｃｔａｓｅ」の場合、両者が上位下位の関係にあることを容易に判定することはできない。それに対しこれらに対応する日本語である「酵素」と「酸化還元酵素」という単語対の場合、「酵素」という文字列を共有するため、両者が上位下位の関係にあることは文字列の構成を比較することで容易に判定できる。したがって、英語では学習データとして抽出できない単語対であっても、日本語を参考にすると、容易に上位下位の関係にあるか否かを判定し、英語の学習データに追加できる。日本語の学習データの追加の場合も同様である。こうした作用を有効に利用することで、英語と日本語との学習データを互いに効率よく集積できる。

【００４７】

図６は、日本語・英語同時学習部１１６をコンピュータで実現するためのコンピュータプログラムの制御構造を示すフローチャートである。以下、このフローチャートで使用する変数などの表現について説明する。

【００４８】

「ｉ」は、英語の学習データと日本語の学習データを抽出する処理（図２において日本語用学習データ６０及び英語用学習データ６２から拡張した日本語用学習データ７２及び拡張した英語用学習データ７４を得るまでの処理）を繰返す回数を制御するための変数である。

【００４９】

「ＭＡＸ」は上記した処理を繰返す回数として予め指定された定数である。

【００５０】

「L_S」と「L_T」はそれぞれソース言語（ここではソース言語として英語を考える。）及びターゲット言語（ここではターゲット言語は日本語である。）の初期学習データを示す。

【００５１】

「L^i_S」は、上記した処理のｉ番目の繰返しにおける、英語の学習データ（図３の英語学習データ記憶部１３４のデータ）を示す。「L^i_T」は同様にｉ番目の繰返しにおける、日本語の学習データ（図３の日本語学習データ記憶部１４４）を示す。

【００５２】

「c^i_S」は、英語の学習データL^i_Sを用いて学習した英語の分類器（図３に示す英語分類器１３０）を示す。「c^i_T」は、日本語の学習データL^i_Tを用いて学習した日本語の分類器（図３に示す日本語分類器１４０）を示す。

【００５３】

「CR^i_S」は、英語の包摂関係語候補（図３の英語の包摂関係語候補１０８）に対して分類器c^i_Sを適用して得られた結果を示す。「CR^i_T」は、日本語の包摂関係語候補（図３の日本語の包摂関係語候補１１０）に対して分類器c^i_Tを適用して得られた結果を示す。

【００５４】

図６を参照して、このプログラムは、変数ｉに０を代入するステップ２４０と、英語及び日本語の学習データL^0_S及びL^0_Tを初期学習データL_SおよびL_Tに設定するステップ２４２と、英語及び日本語の学習データL^i_S及びL^i_Tの同時学習処理２４６を、

ＭＡＸにより表される回数だけ繰返すステップ２４４とを含む。

【００５５】

同時学習処理２４６は、英語の学習データL^{i}_{S}により図３に示す英語分類器１３０（c^{i}_{S}）の学習を行ない、日本語の学習データL^{i}_{T}により図３に示す日本語分類器１４０（c^{i}_{T}）の学習を行なうステップ２５０と、英語の包摂関係語候補１０８（図３）に対して英語分類器１３０（c^{i}_{S}）（図３）を適用してその結果（CR^{i}_{S}）を得、日本語の包摂関係語候補１１０に対して日本語分類器１４０（c^{i}_{T}）（図３）を適用してその結果（CR^{i}_{T}）を得るステップ２５２と、次の繰返しの際に使用される学習データL^{i+1}_{S}及びL^{i+1}_{T}にそれぞれ現在の学習データL^{i}_{S}及びL^{i}_{T}を代入するステップ２５４と、英語の分類結果CR^{i}_{S}のうち、スコアが上位の所定個数の組を用いて、日本語の学習データL^{i+1}_{T}を更新するステップ２５６と、日本語の分類結果CR^{i}_{T}のうち、スコアが上位の所定個数の組を用いて、英語の学習データL^{i+1}_{S}を更新するステップ２５８と、変数ｉの値を１インクリメントするステップ２６０とを含む。

【００５６】

図７は、図６のステップ２５６の処理を実現するプログラムの制御構造を示すフローチャートである。図７を参照してこの処理は、分類結果CR^{i}_{S}のうちスコアが上位である所定個数の組のすべてに対し、以下に説明する日本語学習データの追加処理２７２を行なうステップ２７０を含む。

【００５７】

日本語学習データの追加処理２７２は、、英語対のスコアがしきい値θ以上か否かを判定し、しきい値未満であればこの英語対に対する処理を終了するステップ２８０と、ステップ２８０の判定結果がＹＥＳのときに実行され、英語の分類結果CR^{i}_{S}の中の処理対象の分類結果（英語の単語対＋分類ラベル）の英語の単語に対応する日本語単語を翻訳辞書１１４でルックアップするステップ２８２と、ステップ２８２で英語の単語の両者について、対応の日本語訳があるか否かを判定し、いずれか一方でも日本語訳が翻訳辞書１１４に存在していないときにはこの分類結果の英単語対に対する処理を終了するステップ２８４とを含む。

【００５８】

日本語学習データの追加処理２７２はさらに、ステップ２８４において英単語対の両者について対応する日本語が存在した場合に実行され、その日本語対が日本語の分類結果CR^{i}_{T}に存在するか否かを判定し、存在しない場合にはこの英単語対に対する処理を終了するステップ２８６と、ステップ２８６の判定結果がＹＥＳのときに実行され、翻訳により得られた日本語対に対して日本語分類器１４０による分類を適用するステップ２８７と、ステップ２８７で得られた分類結果のスコアがしきい値θ未満か否かを判定し、判定結果に応じて制御の流れを分岐させるステップ２８８と、ステップ２８８の判定結果がＹＥＳのときに実行され、この日本語対をステップ２８７における分類結果とともに日本語の学習データL^{i+1}_{T}に追加してこの英語の単語対に対する処理を終了するステップ２９２と、ステップ２８８の判定結果がＮＯであるときに実行され、処理対象の英語対の分類ラベルと、ステップ２８７における判定で得られた分類ラベルとが一致するか否かを判定し、一致する場合にはステップ２９２に制御を進め、一致しない場合にはこの英語対に対する処理を終了するステップ２９０とを含む。

【００５９】

図８は、図６のステップ２５８の処理を実現するプログラムの制御構造を示すフローチャートである。図８を参照してこの処理は、分類結果CR^{i}_{T}のうちスコアが上位である所定個数の組のすべてに対し、以下に説明する英語学習データの追加処理３０２を行なうステップ３００を含む。

【００６０】

英語学習データの追加処理３０２は、日本語の分類結果CR^{i}_{T}の中の日本語対のスコアがしきい値θ以上か否かを判定し、しきい値未満であればこの日本語対に対する処理を終了するステップ３１０と、ステップ３１０の判定結果がＹＥＳのときに実行され、日本

語の単語対中の日本語の単語に対応する英語単語を翻訳辞書１１４でルックアップするステップ３１２と、ステップ３１２で日本語の単語の両者について、対応の英語訳があるか否かを判定し、いずれか一方でも英語訳が翻訳辞書１１４に存在していないときにはこの分類結果の日本語単語対に対する処理を終了するステップ３１４と、ステップ３１４において日本語単語対の両者について対応する英語が存在した場合に実行され、その英語対が英語の分類結果CR^{i}_{S}に存在するか否かを判定し、存在しない場合にはこの日本語単語対に対する処理を終了するステップ３１６とを含む。

【００６１】

日本語学習データの追加処理３０２はさらに、ステップ３１６の判定結果がＹＥＳのときに実行され、翻訳により得られた英語対に対して英語分類器１３０による分類を適用するステップ３１７と、ステップ３１７で得られた分類結果のスコアがしきい値θ未満か否かを判定し、判定結果に応じて制御の流れを分岐させるステップ３１８と、ステップ３１８の判定結果がＹＥＳのときに実行され、この英語対をステップ３１７における分類結果とともに英語の学習データL^{i+1}_{S}に追加してこの日本語の単語対に対する処理を終了するステップ３２２と、ステップ３１８の判定結果がＮＯであるときに実行され、処理対象の日本語対の分類ラベルと、ステップ３１７における判定で得られた分類ラベルとが一致するか否かを判定し、一致する場合にはステップ３２２に制御を進め、一致しない場合にはこの英語対に対する処理を終了するステップ３２０とを含む。

【００６２】

なお、図３に示す英語の包摂関係語候補１０８及び日本語の包摂関係語候補１１０としては、任意の英単語対及び日本語単語対でよい。しかし、学習データとしては、ラベルが真のものと偽のものとが適度に含まれていると、学習の効率が高くなる。いずれか一方の単語対のみが大量に存在する場合には、学習データの学習効率が低下する可能性が高く、処理に要する時間も長くなる。ランダムに選んだ単語からなる単語対のみでは、偽の単語対のみが大量に得られることになり、学習の効率が悪い。そこで、英語の包摂関係語候補１０８及び日本語の包摂関係語候補１１０の中には、以下に述べるような方法により、真の分類結果になる可能性が高い単語対が多く含まれるようにする。

【００６３】

図９は、そのような単語対を抽出する処理を説明するための図である。図９を参照してＷｉｋｉｐｅｄｉａに限らず、一般的にＨＴＭＬ形式の文書４００では、テキスト内にレイアウト情報を含む。レイアウト情報は、たとえば第１レベルの見出し４０２、第２レベルの見出し４０４及び４０６、第３レベルの見出し４０８、などのように、レベル別の見出しタグを含む。またＨＴＭＬ形式の文書には、リスト４１０が含まれることがあり、リストを形成する見出しはリストのためのタグにより識別できる。

【００６４】

このような見出し及びリストは、上位語及び下位語の関係にある単語を含むことが多い。そこで、本実施の形態では、こうしたレイアウト情報に基づき、見出し及びリストを構成する単語を抽出し、見出し相互の間の関係に基づいてツリー構造４２０を形成する。このツリー構造４２０において、上位ノードにある単語を、その単語の下位ノードにある単語全てと組合せることにより、単語対を形成する。このような処理によって、包摂関係を充足する単語対を比較的多く含む単語対の集合を得ることができる。これらを英語の包摂関係語候補１０８及び日本語の包摂関係語候補１１０（図３）として使用することにより、学習データを効率よく集積できる。

【００６５】

＜コンピュータによる実現＞

上述の実施の形態は，コンピュータシステムと、当該システム上で実行されるコンピュータプログラムとによって実現可能である。図１０はこれら実施の形態で用いられるコンピュータシステム５５０の外観を示し、図１１はコンピュータシステム５５０のブロック図である。ここで示すコンピュータシステム５５０は単なる例示であって、さまざまな他の構成が利用可能である。

【００６６】
図１０を参照して、コンピュータシステム５５０は、コンピュータ５６０と、モニター５６２と、キーボード５６６と、マウス５６８と、スピーカ５５８と、マイクロフォン５９０とを含む。さらに、コンピュータ５６０は、ＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｃ）ドライブ５７０及び半導体メモリポート５７２を含む。
【００６７】
図１１を参照して、コンピュータ５６０はさらに、ＤＶＤドライブ５７０及び半導体メモリポート５７２に接続されたバス５８６と、上述した装置を実現するコンピュータプログラムを実行するためのＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）５７６と、コンピュータ５６０のブートアッププログラムなどを記憶するＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ）５７８と、ＣＰＵ５７６によって使用される作業領域及びＣＰＵ５７６によって実行されるプログラムの記憶領域を提供するＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）５８０と、英語版のＷｉｋｉｐｅｄｉａのページデータ１００、日本語版のＷｉｋｉｐｅｄｉａのページデータ１０２、英語の包摂関係語候補１０８、日本語の包摂関係語候補１１０、翻訳辞書１１４、英語学習データ、日本語学習データ、及び処理途中で一時的に作成されるデータを記憶するためのハードディスク（ＨＤ）５７４と、コンピュータ５６０にネットワーク５５２との接続を提供するためのネットワークインターフェース（Ｉ／Ｆ）５９６とを含み、これらは全てバス５８６に接続されている。
【００６８】
上述の実施の形態に係るシステムを実現するソフトウェアはＤＶＤ５８２又は半導体メモリ５８４等の記憶媒体に記憶されたオブジェクトコードの形で流通し、ＤＶＤドライブ５７０又は半導体メモリポート５７２等の読出装置を介してコンピュータ５６０に提供され、ハードディスク５７４に記憶される。ＣＰＵ５７６がプログラムを実行する際には、プログラムはハードディスク５７４から読出されてＲＡＭ５８０に記憶される。図示しないプログラムカウンタによって指定されたアドレスから命令がフェッチされ、ＣＰＵ５７６によりその命令が実行される。ＣＰＵ５７６はハードディスク５７４から処理すべきデータを読出し、処理の結果をこれもまたハードディスク５７４に記憶する。
【００６９】
コンピュータシステム５５０の一般的動作は周知であるので、ここでは詳細な説明は行なわない。
【００７０】
ソフトウェアの流通の方法に関して、ソフトウェアは必ずしも記憶媒体上に固定されたものでなくてもよい。例えば、ソフトウェアはネットワーク５５２に接続された別のコンピュータから配布されてもよい。ソフトウェアの一部がハードディスク５７４に記憶され、ソフトウェアの残りの部分をネットワークを介してハードディスク５７４に取込み、実行の際に統合する様にしてもよい。
【００７１】
典型的には、現代のコンピュータはコンピュータのオペレーティングシステム（ＯＳ）によって提供される汎用の関数を利用し、所望の目的に従って制御された態様でこれら関数を実行する。従って、ＯＳ又は第３者から提供されうる汎用関数を含まず、一般的な関数の実行順序の組合せのみを指定したプログラムであっても、そのプログラムが全体として所望の目的を達成する制御構造を有する限り、そのプログラムがこの発明の範囲に包含されることは明らかである。
【００７２】
また、プログラムは必ずしもオブジェクトコード形式でなくともよい。コンピュータシステム５５０にコンパイラが存在する場合には、ソースコードで提供されたプログラムをコンパイルしてオブジェクトコードとすることで、上記した処理を実現するオブジェクトプログラムが得られる。
【００７３】

コンピュータシステム５５０に特定の言語のスクリプトの実行系が備えられている場合、プログラムはスクリプト形式でこのコンピュータに提供されてもよい。複数個のスクリプトにより上記した処理が実現される場合、それらスクリプトがどこに存在しているかにかかわらず、それらをまとめてコンピュータシステム５５０に格納可能とするようなサービスをたとえばネットワーク上で提供した場合、そうしたサービスは本発明の実施に相当する。

【００７４】

さらに、プログラムを分割可能な複数個のユニットに分割し、それらを別々のコンピュータで実行することで、上記した処理を実現する場合にも、本発明の実施に相当することはいうまでもない。

【００７５】

＜動作＞

以上に構成を説明した分類器の同時学習装置９０（図３）は以下のように動作する。最初に、英語版のＷｉｋｉｐｅｄｉａのページデータ１００及び日本語版のＷｉｋｉｐｅｄｉａのページデータ１０２をＨＤ５７４などの記憶媒体に集積する。この作業は手作業でもよいし、いわゆるロボットプログラムでＷｉｋｉｐｅｄｉａのページを巡回することで集積してもよい。

【００７６】

次いで、翻訳辞書作成部１１２により翻訳辞書１１４を準備する。翻訳辞書１１４の作成には、既存の方法、たとえば特開2007-280122号公報、特開2005-250746号公報、特開2002-366546号公報などに開示されたものを使用することができる。本実施の形態では、単純に１つの英単語と１つの日本語単語とを対訳形式で割当てることにより翻訳辞書１１４を作成すればよい。

【００７７】

英語及び日本語の包摂関係語候補抽出部１０４及び１０６により、英語の包摂関係語候補１０８及び日本語の包摂関係語候補１１０を作成し、ＨＤ５７４に記憶させる。

【００７８】

英語学習データ記憶部１３４及び日本語学習データ記憶部１４４に、初期学習データを準備する。この初期学習データの形式は図５に示したとおりである。これらは手作業で新たに準備してもよいし、既存の学習データを用いてもよい。初期学習データの量はそれほど多くなくてもよい。

【００７９】

以下、英語学習データ及び日本語学習データの同時集積と、英語分類器１３０と日本語分類器１４０との同時学習を開始する。図６に示すコンピュータプログラムでは最初に変数ｉに０が代入され（ステップ２４０）、英語及び日本語の初期学習データが指定される（ステップ２４２）。

【００８０】

図３を参照して、学習部１３２により、英語学習データ記憶部１３４に記憶された英語学習データを用いて英語分類器１３０の学習が行なわれる（図６、ステップ２５０）。これと同時、又はこの処理に続き、学習部１４２により、日本語学習データ記憶部１４４に記憶された日本語学習データを用いて日本語分類器１４０の学習が行なわれる（ステップ２５０）。

【００８１】

以下、日本語学習部１６０の動作について説明する。英語学習部１６２の動作は英語と日本語とを交換することを除き、日本語学習部１６０と同じである。

【００８２】

英語分類器１３０による分類を英語の包摂関係語候補１０８に適用することにより、追加候補１８０（図４）が得られる（図６、ステップ２５２）。このとき、追加候補１８０内の単語対の各々にはスコアが付されている。選択部１８２は、追加候補１８０のうち、英語学習データ記憶部１３４に存在せず、かつスコアがしきい値θ以上のものの上位所定

個までを選択し、選択結果１８４として出力する（図７、ステップ２８０）。
【００８３】
英日翻訳部１８６は、選択結果１８４内の各単語対を構成する単語の各々について翻訳辞書１１４を参照して翻訳を試みる（ステップ２８２）。単語対内の単語の双方について日本語の訳語が存在した場合、英日翻訳部１８６はその日本語対を翻訳結果１８８として出力する（図７、ステップ２８４でＹＥＳ）。単語対内の単語のいずれか一方でも対応の日本語訳が翻訳辞書１１４に存在しない場合、英日翻訳部１８６はこの単語対を無視する（ステップ２８４でＮＯ）。この日本語訳が日本語の分類結果中にない場合にも処理対象の単語対は無視される（ステップ２８６でＮＯ）。
【００８４】
日本語分類器１４０は、英日翻訳部１８６の処理の結果得られた翻訳結果１８８を構成する日本語の単語対の各々について分類を行ない、分類ラベル（真／偽）とそのスコアとを付して分類結果１９０として出力する（ステップ２８７）。
【００８５】
選択部１９２は、分類結果１９０のうち、（１）日本語対の分類スコアがしきい値θ未満のもの（図７、ステップ２８８でＹＥＳ）、又は（２）日本語対の分類スコアがしきい値θ以上で、かつ追加候補１８０における対応する英語対の分類ラベルと、日本語分類器１４０による分類ラベルとが一致するもの（ステップ２８８でＮＯ、かつステップ２９０でＹＥＳ）、を分類結果１９０の中から選択し、その日本語単語対に、追加候補１８０の対応する英語の単語対のラベルを付したものを選択結果１９４として出力する。それ以外については何もしない。
【００８６】
更新部１９６は、選択結果１９４を新たな日本語学習データとして日本語学習データ記憶部１４４に追加する（ステップ２９２）。
【００８７】
こうして、所定回数だけ上記した処理を繰返す。最終的に英語学習データ記憶部１３４及び日本語学習データ記憶部１４４には、同時学習により、初期の状態と比較してより多くの学習データが記憶されている。その精度は高い。このように同時学習した英語学習データ及び英語分類器１３０及び日本語分類器１４０についても、その分類精度は高くなる。これは、以下に述べるように実験によって確認された。
【００８８】
＜実験＞
２００８年５月の英語版Ｗｉｋｉｐｅｄｉａと、２００８年６月版の日本語版Ｗｉｋｉｐｅｄｉａとを用いて以下に述べるような実験を行なった。両言語について２４０００個の包摂関係語候補を抽出し、手作業で図３に示す初期英語学習データ、初期日本語学習データ、英語及び日本語の包摂関係語候補データとを作成し、さらに同様にしてテストデータを作成した。両言語について、これら候補の中で８０００個の包摂関係にある単語対が存在した。２００００個の単語対を初期トレーニングデータとし、英語分類器１３０及び日本語分類器１４０の学習に用いた。残りの単語対は、両言語についてそれぞれ等分し、一方は包摂関係語候補１０８及び１１０として用い、他方はテストデータとして用いた。
【００８９】
この実験では、分類器（英語分類器１３０及び日本語分類器１４０）として、２次多項式カーネルのＴｉｎｙＳＶＭを用いた。最大繰返し数ＭＡＸ＝１００とした。しきい値θ＝１とし、包摂関係語候補としては９００個を選択することにした。
【００９０】
実験ではＷｉｋｉｐｅｄｉａの対応する日英のリンクから抽出したバイリンガル翻訳辞書を用いた。
【００９１】
ここでは、精度（Ｐ）、再現率（Ｒ）、およびＦ１値（Ｆ１）を次の式のように定めた。ただし、Ｒｅｌは手作業で検査した包摂関係の集合を表し、ＨＲｂｙＳは実験対象のシ

ステムにより包摂関係にあると判定された包摂関係語候補の集合を表す。

【００９２】

【数１】

$$P = |Rel \cap H\ RbyS| / |H\ RbyS| \qquad (1)$$

$$R = |Rel \cap H\ RbyS| / |Rel|$$

$$F_1 = 2 \times (P \times R) / (P + R)$$

［実験］

【００９３】

【表２】

テーブル2

	英語			日本語		
	P	R	F1	P	R	F1
SYT	78.5	63.8	70.4	75.0	77.4	76.1
INIT	77.9	67.4	72.2	74.5	78.5	76.6
TRAN	76.8	70.3	73.4	76.7	79.3	78.0
BICO	78.0	83.7	80.7	78.3	85.2	81.6

テーブル２は、４種類の分類システムの結果をパーセントで示す。ＳＹＴは従来例としてスミダら（アスカ　スミダ他、「包摂関係の獲得のためのＷｉｋｉｐｅｄｉａのハッキング」．自然言語処理に関する第３回国際合同会議（ＩＪＣＮＬＰ）、ｐｐ．８８３－８８８、２００８年１月）によるシステムを発明者らが実装したものによる分類結果を示す。ＩＮＩＴは上記システムでの初期学習データにより学習をした分類器を用いたシステムの分類結果を示す。英語及び日本語の学習データのサイズは、それぞれ２０，７２９語と２０，４８６語であった。ＴＲＡＮは、上記した初期学習データをそれぞれ相手側言語に翻訳して相手側の学習データに追加したものを用いて学習した分類器を用いたものである。ＢＩＣＯは上記した実施の形態によるものである。

【００９４】

上記結果を参照して、日本語についてはＳＹＴの性能は上記スミダらによる報告結果より低い。これは学習データのサイズによるものと思われる（本実験では２０，０００、スミダらによる実験では２９，９００）。テストデータのサイズも異なっている（本実験では２，０００、スミダらの実験では１，０００）。

【００９５】

ＩＮＩＴとＳＹＴとの比較により、ＳＶＭの素性として使用したもの（テーブル１を参照）を比較すると、ＳＦ３－ＳＦ５とＩＦとの影響がわかる。ＩＮＩＴは、Ｆ１値にしてわずか０．５－１．８％ではあるが、常にＳＹＴの性能を上回っている。

【００９６】

ＢＩＣＯにより、ＳＹＴ、ＩＮＩＴ及びＴＲＡＮに比して、Ｆ１値にして３．６－１０．３％というかなりの性能の改善が得られた。ＴＲＡＮとＢＩＣＯとの比較により、このようなバイリンガル同時学習が、学習データの拡張に有効であること、及びこのようなバイリンガル同時学習により得られた性能向上は、既存の学習データを単に翻訳するだけでは得られないことがわかった。

【００９７】

図１２は、手操作により準備したものと、バイリンガル同時学習により拡張されたものとを含む学習データのサイズに対する、Ｆ１値の関係を示す。図１２を参照して、このグラフは、サイズ＝２０，０００からスタートして日本語の場合５０，０００個まで、英語の場合６２，０００個まで続く。学習データのサイズが大きくなるにつれて、Ｆ１曲線はいずれの言語の場合にも上昇していく傾向にあることが分かる。このグラフから、２言語

の分類器の同時学習により、互いに協働して性能が向上していくことが分かる。

【００９８】

最終的には、この実験により英語で５４０万、日本語で２４１万の包摂関係が得られた。

【００９９】

以上のとおり、本実施の形態によれば、英語及び日本語の包摂関係の分類器の学習において、同時学習を行なうことで効率的に学習データを追加し、分類器の性能を向上させることができる。

【０１００】

＜可能な変形例＞

上記した実施の形態は、英語と日本語との組合せに関するものであった。しかし、自然言語処理の技術分野における技術者であれば容易に分かるように、この手法及びシステムは、任意の言語の組合せに対しても適用することができる。確率モデルの学習を行なうときの素性は、各言語の特徴に応じて適切なものを選択すればよい。

【０１０１】

なお、図６に示す処理では、一定回数ＭＡＸだけ学習処理を繰返すと同時学習を終了する。しかし繰返しの終了条件はこのような条件には限定されない。たとえば、英語と日本語との双方において、新たに追加する単語対が得られなかったときに終了してもよいし、いずれか一方において新たな単語対が得られないときに終了してもよい。それに代えて、新たに追加する単語対の数が所定のしきい値以下となったときに終了してもよい。この場合、英単語及び日本語単語の一方がそうした終了条件を満たしたときでもよいし、その双方ともその条件を満たしたときでもよい。さらには、英単語及び日本語単語で新たに追加すべき単語対の数の合計が終了条件を満たしたときに繰返しを終了するようにしてもよい。その他、終了条件としては種々のものを想定することができる。

【０１０２】

上記した実施の形態では、分類器としてＳＶＭを用いた。しかし本発明はそのような実施の形態には限定されない。分類器としては、分類結果とともに、分類結果の信頼性（確率）を示すスコアを出力可能な、機械学習による確率モデルであれば、どのようなものでも用いることができる。たとえば従来技術の項で述べたＣＲＦ及びＭＥＭを用いたものでもよい。

【０１０３】

さらに、上記した実施の形態では、分類結果は真／偽の２値であったが、本発明はそのような実施の形態には限定されない。３値以上の分類を行なう分類器についても、同様に本発明を適用することができる。

【０１０４】

上記実施の形態では、本発明を包摂関係（単語の上位下位関係）に適用した場合を説明した。しかし本発明はそのような実施の形態に限定されるわけではなく、単語対の間に定義される意味的関係であれば、どのような関係についても適用することができる。たとえば、原因と結果、類語関係、状況と対策、状況（トラブル）とその原因、部分と全体、問題と解決のためのツールなど、単語の間の様々な関係の判定に本願発明を適用することができる。

【０１０５】

さらに、上記した実施の形態では、英語の分類器と、日本語の分類器として同種の確率モデル（ＳＶＭ）を用いている。しかし本発明はそのような実施の形態には限定されない。第１の言語の分類器と、第２の言語の分類器として、異種のものを用いてもよい。この場合、第１の言語と第２の言語とが一致していてもよい。

上記した実施の形態では、２言語について分類器の同時学習を行なっている。しかし本発明はそのような実施の形態には限定されない。３言語以上の何らかの意味的関係の分類器の同時学習にも容易に適用可能である。たとえば３言語の場合には、第１の言語の分類器による分類結果を用いて第２の言語の学習データの更新及び分類器の学習を行ない、第

２の言語の分類器による分類結果を用いて第３の言語の学習データの更新及び分類器の学習を行ない、第３の言語の分類器による分類結果を用いて第１の言語の学習データの更新を行ない、というように巡回的に学習データの更新と分類器の学習とを行なってもよい。

【０１０６】

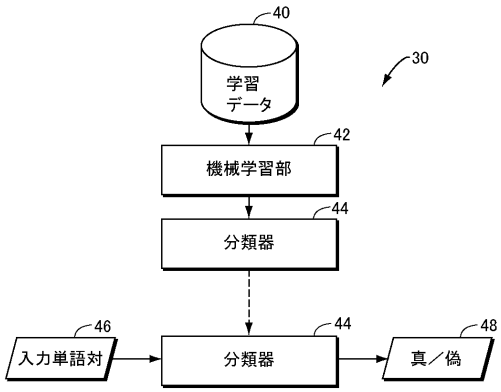
今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内での全ての変更を含む。

【符号の説明】

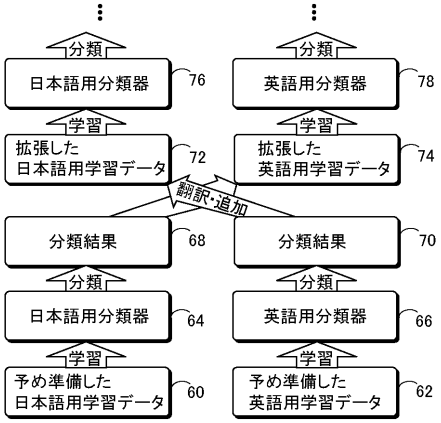
【０１０７】

９０　分類器の同時学習装置
１１４　翻訳辞書
１１６　日本語・英語同時学習部
１３０　英語分類器
１３２，１４２　学習部
１３４　英語学習データ記憶部
１４０　日本語分類器
１４４　日本語学習データ記憶部
１５０　学習データ更新部
１６０　日本語学習部
１６２　英語学習部
１８２，１９２，２１２，２２２　選択部
１８６　英日翻訳部
１９６，２２６　更新部
２１６　日英翻訳部

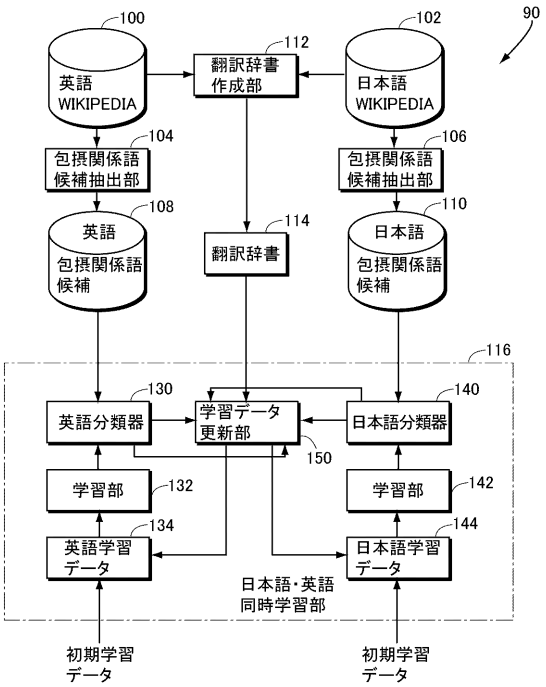
【図１】



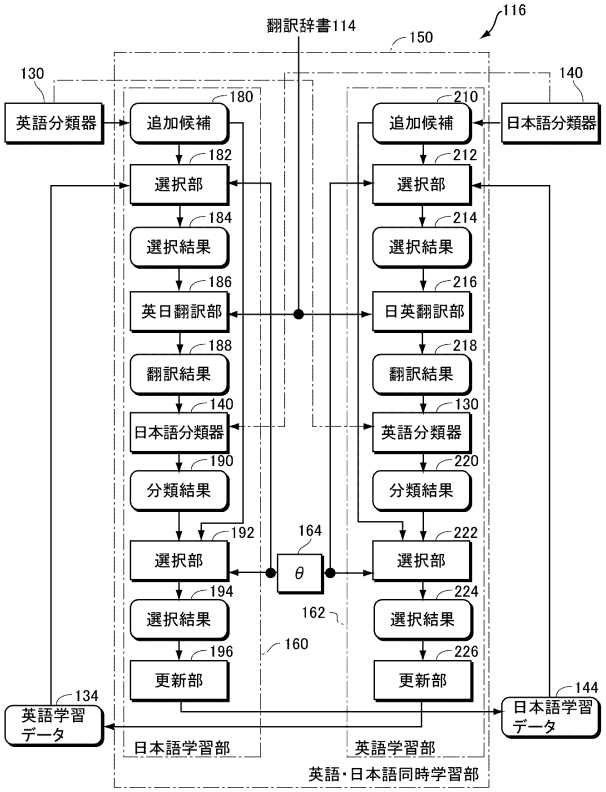
【図２】



【図３】



【図４】



【図５】

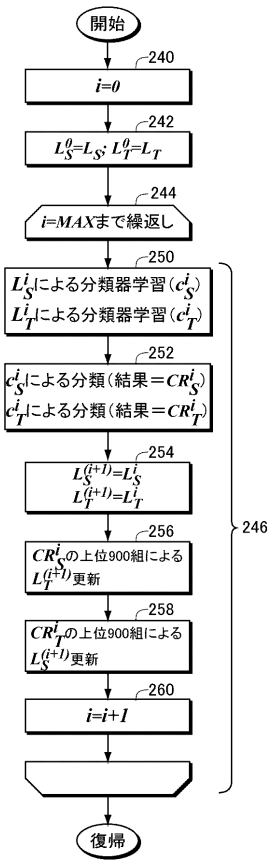
(A) 134

分類	上位	下位
×	Enzyme	History of biochemistry
×	Enzyme	History
×	Dog	Distribution
○	dog	Akbash Dog
○	dog	Terrier
×	Soviet space dogs	Training
×	Akita inu	History
×	Shinto Shrine	Koma inu
…		

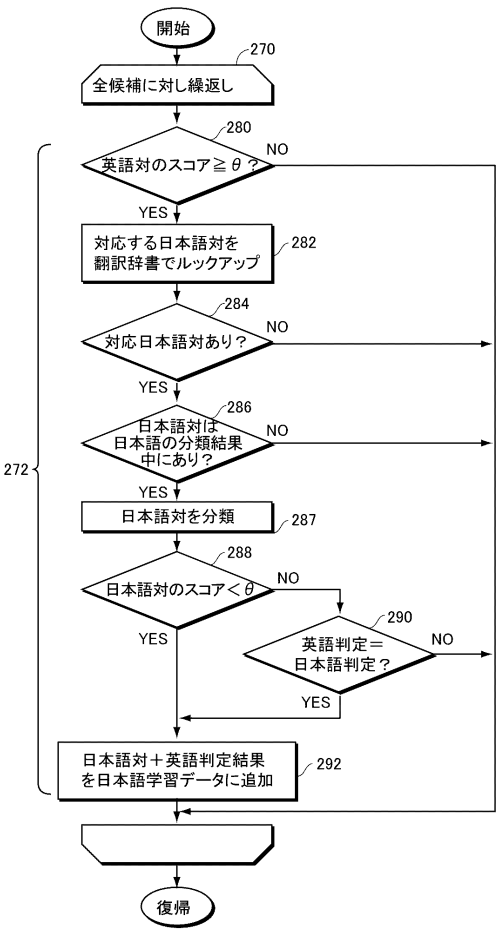
(B) 144

分類	上位	下位
×	酵素	生化学の歴史
×	酵素	歴史
○	酵素	酸化還元酵素
○	酵素	転移酵素
×	イヌ	分布
×	ソ連の宇宙犬	トレーニング
×	秋田犬	歴史
×	神社	狛犬
…		

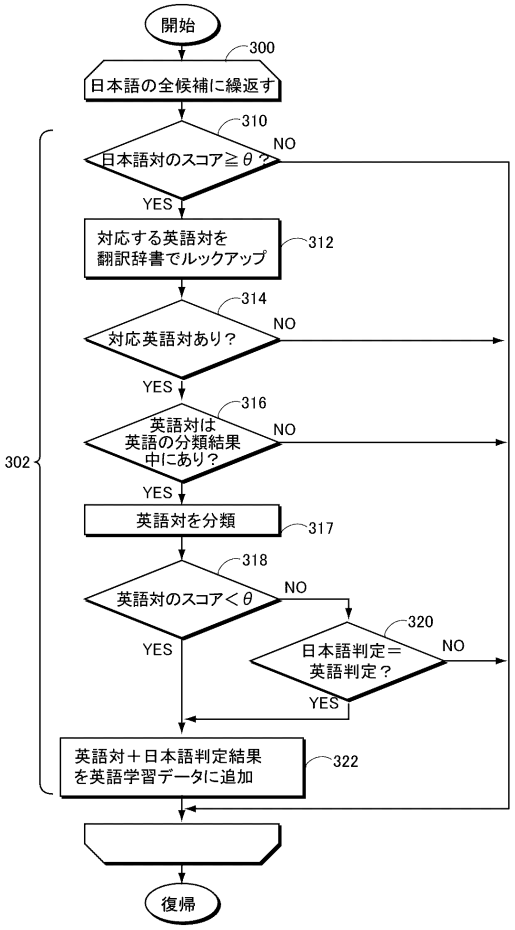
【図６】



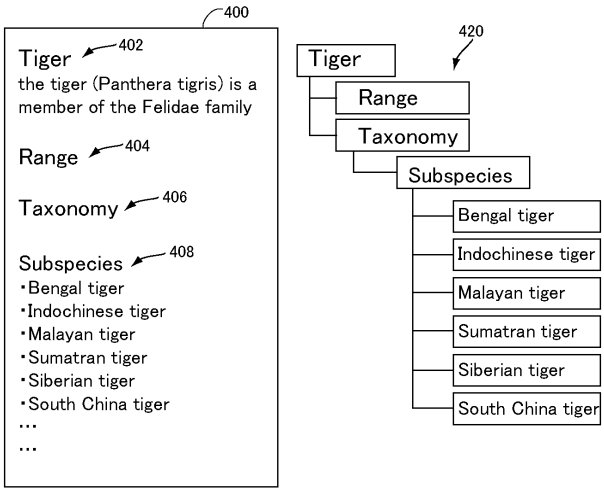
【図７】



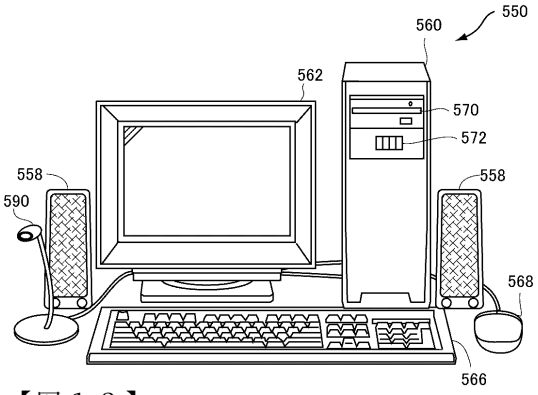
【図８】



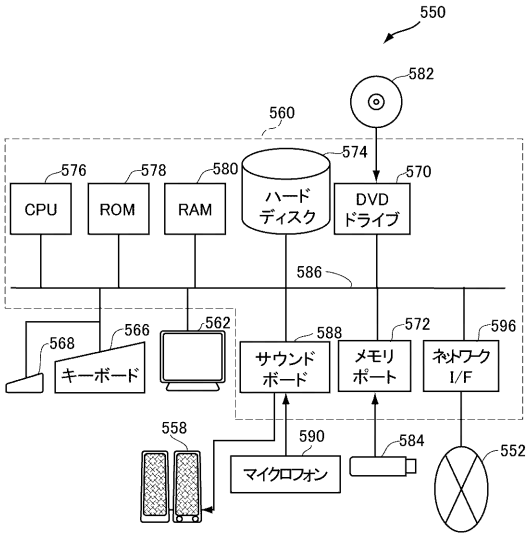
【図９】



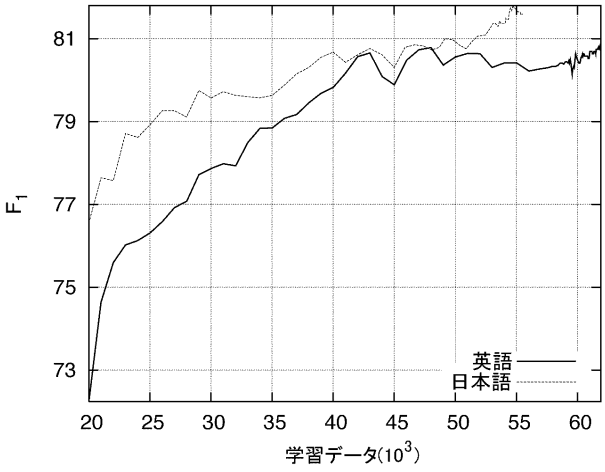
【図１０】



【図１１】



【図１２】



フロントページの続き

審査官　長　由紀子

(56)参考文献　特開２００６－００４３９９（ＪＰ，Ａ）
隅田　飛鳥　外2名，Ｗｉｋｉｐｅｄｉａの記事構造からの上位下位関係抽出，自然言語処理，日本，言語処理学会，２００９年　7月１０日，第16巻第3号，P.3～24
中川　哲治　外2名，事例の重み付けに基づく自動獲得されたコーパスの効果的な利用法と評価極性分類への応用，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，２００９年　1月１９日，第108巻第408号，P.25～30

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０－２８
Ｇ０６Ｆ　１７／３０