

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2011-34171

(P2011-34171A)

(43) 公開日 平成23年2月17日(2011.2.17)

(51) Int.Cl.	F I	テーマコード (参考)
<b>G06F 17/28 (2006.01)</b>	G06F 17/28 Z	5B075
<b>G06F 17/27 (2006.01)</b>	G06F 17/27 M	5B091
<b>G06F 17/30 (2006.01)</b>	G06F 17/30 170A	
	G06F 17/30 220Z	
	G06F 17/30 210D	

審査請求 未請求 請求項の数 8 O L (全 23 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)

(71) 出願人 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  
 Fターム(参考) 5B075 ND03 NR12  
 5B091 BA16 CA12 EA01

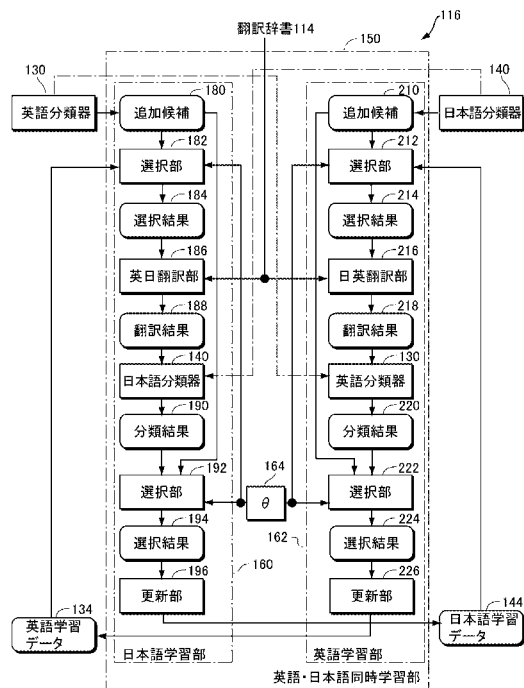
(54) 【発明の名称】 単語間の意味的関係の有無についての、複数言語での同時分類器及びそのためのコンピュータプログラム。

(57) 【要約】

【課題】 低コストで意味的知識を効率よく分類できる分類器を提供する。

【解決手段】 同時学習装置150は、英単語間の意味的関係の有無を判定し、スコアとともに出力する英語分類器130、日本語単語について同様に動作する日本語分類器140、英日の学習データ134及び144、英日の学習データにより英語分類器130と日本語分類器140との学習を行なう学習部、英日の学習データの候補を日英に翻訳する英日翻訳部186及び日英翻訳部216、英日の翻訳結果188を日本語分類器140で分類した結果190及び元の英語に対する分類結果180に基づき、翻訳結果140のうち信頼性の高いものを日本語学習データ144に追加する更新部196、日英の翻訳結果218を英語分類器130で分類した結果220及び元の日本語に対する分類結果210に基づき、翻訳結果218のうち信頼性の高いものを英語学習データ134に追加する更新部226を含む。

【選択図】 図4



**【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

10

20

30

40

50

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。

10

【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】

20

【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。

30

【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。

40

【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

50

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關係(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。

10

【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：名詞類間の意味的關係の分類、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 Semantic Evaluations (SemEval-2007), pages 13-18)

20

【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。

30

【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。

40

【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關係の有無を判定し、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關係の有無を判定し、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

50

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、人手で分類を行ったり、分類結果による候補の取捨選択を行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。

10

【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。

20

【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。

30

【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。

40

【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。

50

【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、

10

20

30

40

50

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。

10

【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。

20

【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。さらにこれを繰り返すことで、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。

30

【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】

40



【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)

10

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。

20

## 【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。

30

## 【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。

40

## 【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

50

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。

10

#### 【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を記憶するための記憶部 164 とを含む。

20

#### 【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以上であり、かつ「分類結果 190 に付されたしきい値が未満である」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。

30

40

#### 【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以上」、かつ「分類結果 190 に付されたしきい値が未満である」という条件は、日本語の分類器では分類の信頼度が低い、英語の分類器による分類の信頼度が高い、ということの意味する。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

50

がしきい値 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 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 以上であり、かつ「分類結果 220 に付されたしきい値が 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「x」（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「x」である。このように、のラベルを持つ学習データと x 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】

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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 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 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。

10

20

【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 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。

30

【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 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。

40

【0062】

50

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の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。

10

#### 【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。

20

#### 【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。

30

#### 【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。

40

50

## 【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み出し、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含ま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

10

20

30

40

50

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。

10

【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)。これと同時に、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)。

20

【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)。

30

【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も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)。

40

【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

50



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。

10

【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 として使い、他方はテストデータとして用いた。

20

【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 Tiny SVM を用いた。最大繰返し数 MAX = 100 とした。しきい値 = 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。

30

【0092】

【数 1】

$$P = |Rel \cap HRbyS| / |HRbyS| \quad (1)$$

$$R = |Rel \cap HRbyS| / |Rel|$$

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

[ 実験 ]

40

【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

10

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。

20

【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。

30

【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。

40

【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 >

50

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。

10

【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。

20

【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關係であれば、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。

30

【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關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。

40

【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の全ての変更を含む。

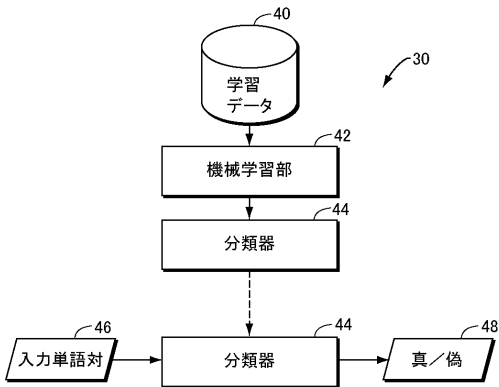
【符号の説明】

【0107】

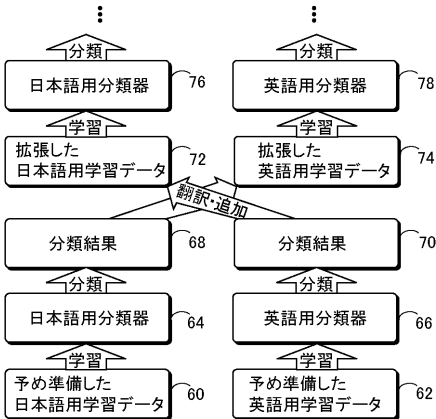
50

- 9 0 分類器の同時学習装置
- 1 1 4 翻訳辞書
- 1 1 6 日本語・英語同時学習部
- 1 3 0 英語分類器
- 1 3 2 , 1 4 2 学習部
- 1 3 4 英語学習データ記憶部
- 1 4 0 日本語分類器
- 1 4 4 日本語学習データ記憶部
- 1 5 0 学習データ更新部
- 1 6 0 日本語学習部
- 1 6 2 英語学習部
- 1 8 2 , 1 9 2 , 2 1 2 , 2 2 2 選択部
- 1 8 6 英日翻訳部
- 1 9 6 , 2 2 6 更新部
- 2 1 6 日英翻訳部

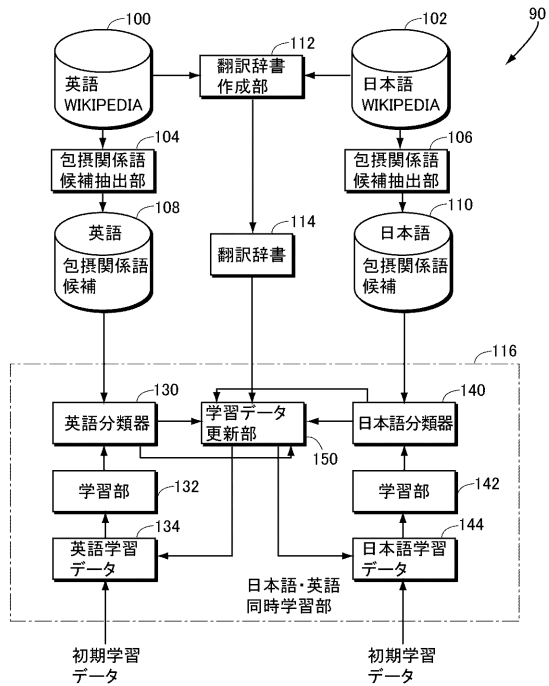
【 図 1 】



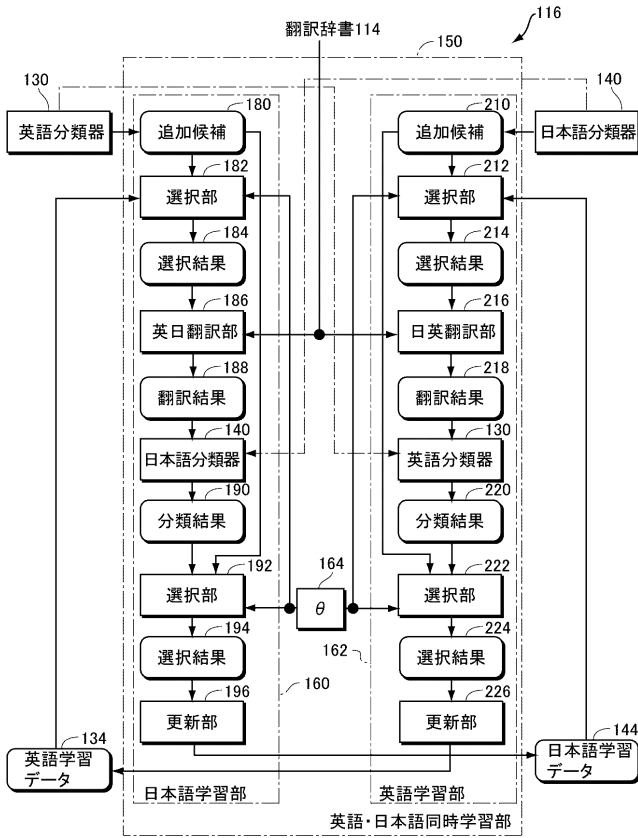
【 図 2 】



【 図 3 】



【図4】



【図5】

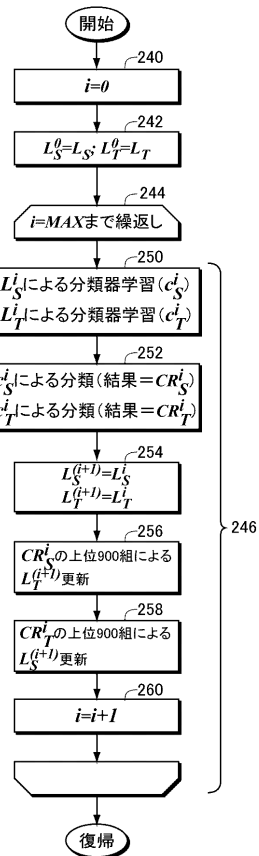
(A)

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

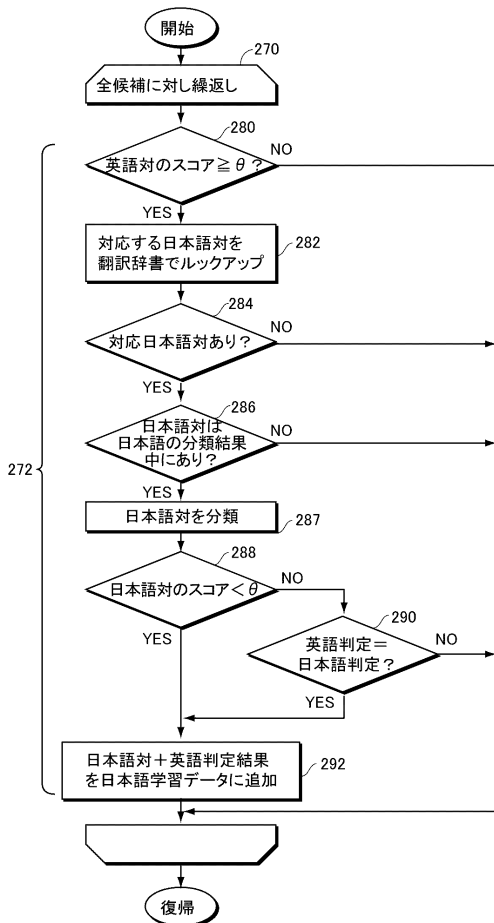
(B)

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
...		

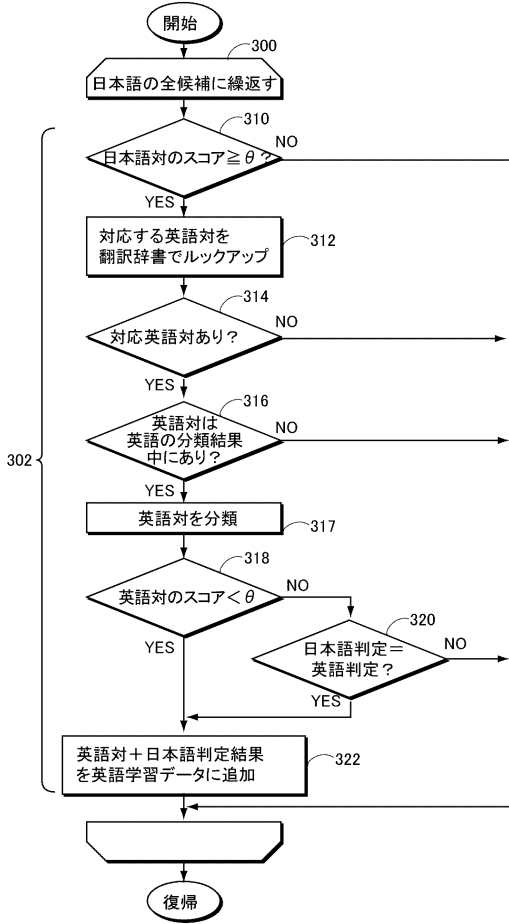
【図6】



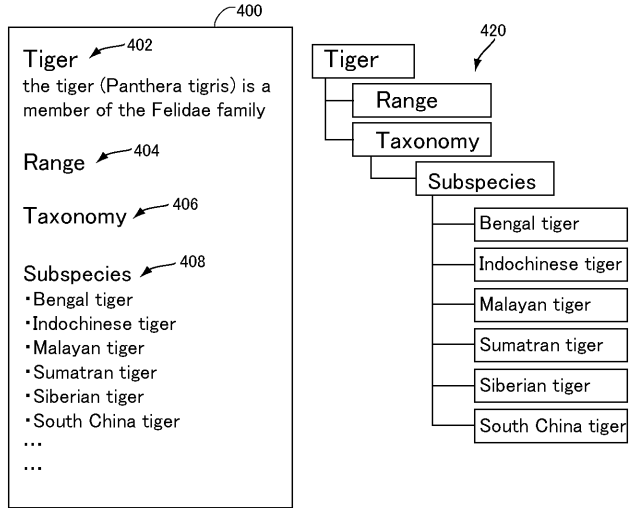
【図7】



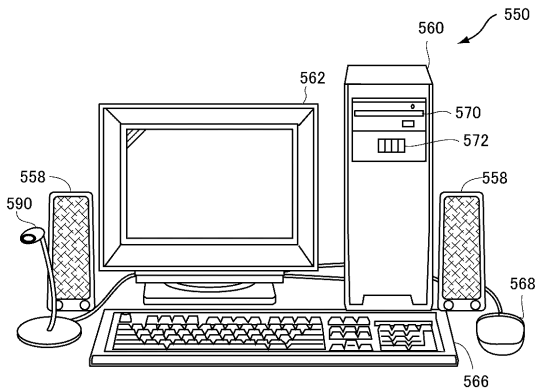
【 図 8 】



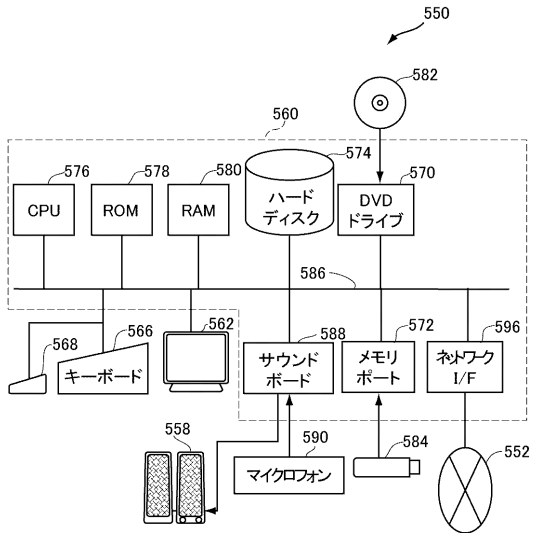
【 図 9 】



【 図 1 0 】



【 図 1 1 】



【 図 1 2 】

