

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年1月2日(02.01.2020)



(10) 国際公開番号

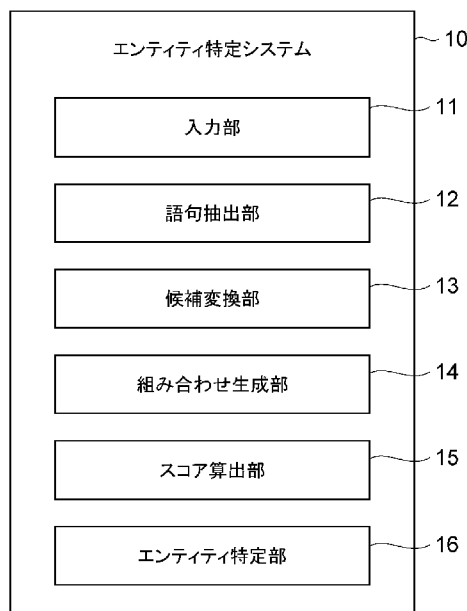
WO 2020/003928 A1

- (51) 国際特許分類:
G06F 17/27 (2006.01) *G06F 16/383* (2019.01)
- (21) 国際出願番号: PCT/JP2019/022218
- (22) 国際出願日: 2019年6月4日(04.06.2019)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2018-123451 2018年6月28日(28.06.2018) JP
- (71) 出願人:株式会社 N T T ドコモ(NTT DOCOMO, INC.) [JP/JP]; 〒1006150 東京都千代田区永田町二丁目 1 1 番 1 号 Tokyo (JP).
- (72) 発明者: 白水 優太郎 (SHIRAMIZU Yutaro); 〒1006150 東京都千代田区永田町二丁目 1 1 番 1 号 山王パークタワー 株式会社 N T T ドコモ 知的財産部内 Tokyo (JP).
- (74) 代理人: 長谷川 芳樹, 外 (HASEGAWA Yoshiki et al.); 〒1000005 東京都千代田区丸の内二丁目 1 番 1 号 丸の内 M Y P L A Z A (明治安田生命ビル) 9 階 創英国際特許法律事務所 Tokyo (JP).
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

(54) Title: ENTITY IDENTIFICATION SYSTEM

(54) 発明の名称: エンティティ特定システム

[図1]



10... ENTITY IDENTIFICATION SYSTEM
11... INPUT UNIT
12... PHRASE EXTRACTION UNIT
13... CANDIDATE CONVERSION UNIT
14... COMBINATION GENERATION UNIT
15... SCORE CALCULATION UNIT
16... ENTITY IDENTIFICATION UNIT

(57) Abstract: In the present invention, an entity appropriate to the context of a sentence is identified. An entity identification system 10 comprises: an input unit 11 to which a sentence is input; a phrase extraction unit 12 that extracts one or more phrases from the input sentence; a candidate conversion unit 13 that converts at least one of the extracted phrases to one or more candidate phrases of an entity linked to the phrase; a combination generation unit 14 that generates at least one phrase combination corresponding to the sentence and including each of the converted one or more phrases; a

HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH,
KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

- 一 国際調査報告 (条約第21条(3))

score calculation unit 15 that, for each generated combination, calculates a score on the basis of a similarity score between phrases included in the combination; and an entity identification unit 16 that, on the basis of the calculated combination score, identifies an entity phrase to be linked, from the one or more candidate phrases.

(57) 要約：文章の文脈に適したエンティティを特定する。エンティティ特定システム10は、文章を入力する入力部11と、入力された文章から1つ以上の語句を抽出する語句抽出部12と、抽出された語句のうち少なくとも何れかについて、当該語句にリンクされるエンティティの1つ以上の候補の語句に変換する候補変換部13と、変換された1つ以上の語句の何れかをそれぞれ含む、文章に対応する語句の組み合わせを1つ以上生成する組み合わせ生成部14と、生成された各組み合わせについて、組み合わせに含まれる語句同士の類似性のスコアに基づいてスコアを算出するスコア算出部15と、算出された組み合わせのスコアに基づいて、1つ以上の候補の語句から、リンクされるエンティティの語句を特定するエンティティ特定部16とを備える。

明 細 書

発明の名称： エンティティ特定システム

技術分野

[0001] 本発明は、文章中の語句にリンクされるエンティティを特定するエンティティ特定システムに関する。

背景技術

[0002] 文章中の語句（キーワード）と、当該語句に対応するエンティティとを対応付けるエンティティリンクングが知られている。エンティティは、文章中における当該語句の概念（文章中において当該語句が示すもの）である。例えば、特許文献1には、インターネット上のデータベースから集められた人名情報を含むWebページ中の文書を解析し、有名人の別表現（愛称等）を抽出することが示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開2008-130034号公報

発明の概要

発明が解決しようとする課題

[0004] 従来のエンティティリンクングでは、文脈及びリンク確率等に基づいて、語句にリンクされるエンティティが特定されていた。しかしながら、従来の方法では、エンティティ候補から適切なエンティティを特定することが困難な場合があった。

[0005] 本発明の一実施形態は、上記に鑑みてなされたものであり、文章の文脈に適したエンティティを特定することができるエンティティ特定システムを提供することを目的とする。

課題を解決するための手段

[0006] 上記の目的を達成するために、本発明の一実施形態に係るエンティティ特定システムは、文章を入力する入力部と、入力部によって入力された文章か

ら1つ以上の語句を抽出する語句抽出部と、語句抽出部によって抽出された語句のうち少なくとも何れかについて、当該語句にリンクされるエンティティの1つ以上の候補の語句に変換する候補変換部と、候補変換部によって変換された1つ以上の語句の何れかをそれぞれ含む、文章に対応する語句の組み合わせを1つ以上生成する組み合わせ生成部と、組み合わせ生成部によって生成された各組み合わせについて、組み合わせに含まれる語句同士の類似性のスコアに基づいてスコアを算出するスコア算出部と、スコア算出部によって算出された組み合わせのスコアに基づいて、1つ以上の候補の語句から、リンクされるエンティティの語句を特定するエンティティ特定部と、を備える。

[0007] 本発明の一実施形態に係るエンティティ特定システムでは、文章に対応する語句同士の類似性に基づいて、文章に含まれる語句にリンクされるエンティティの語句が特定される。従って、本発明の一実施形態に係るエンティティ特定システムによれば、文章の文脈に適したエンティティを特定することができる。

発明の効果

[0008] 本発明の一実施形態によれば、文章に対応する語句同士の類似性に基づいて、文章に含まれる語句にリンクされるエンティティの語句が特定されるため、文章の文脈に適したエンティティを特定することができる。

図面の簡単な説明

[0009] [図1]本発明の実施形態に係るエンティティ特定システムの構成を示す図である。

[図2]文章から抽出される語句の例を示す図である。

[図3]文章中の語句から変換されるエンティティの候補の語句の例を示す図である。

[図4]語句の組み合わせの例を示す図である。

[図5]本発明の実施形態に係るエンティティ特定システムで実行される処理を示すフローチャートである。

[図6]本発明の実施形態に係るエンティティ特定システムのハードウェア構成を示す図である。

発明を実施するための形態

[0010] 以下、図面と共に本発明に係るエンティティ特定システムの実施形態について詳細に説明する。なお、図面の説明においては同一要素には同一符号を付し、重複する説明を省略する。

[0011] 図1に本実施形態に係るエンティティ特定システム10を示す。エンティティ特定システム10は、文章（テキスト、文字列）を入力して、入力した文章に含まれる語句にリンクされるエンティティを特定する装置（システム）である。即ち、エンティティ特定システム10は、エンティティリンクを行う装置である。なお、本実施形態では、日本語の文章を例として説明する。但し、日本語以外の文章であっても、同様にエンティティを特定することができる。例えば、文章中に「連邦裁判所」との語句が含まれていた場合に、エンティティ特定システム10は、当該文章中における「連邦裁判所」が、「アメリカ合衆国連邦裁判所」「連邦裁判所（ドイツ）」「連邦裁判所（スイス）」「オーストラリア連邦裁判所」の何れのエンティティを指しているかを特定する。

[0012] エンティティ特定システム10によるエンティティの特定は、例えば、文章から固有表現を抽出する前処理として行われてもよいし、文章中の語句の語義曖昧性解消のために行われてもよい。また、上記以外の目的でエンティティの特定が行われてもよい。エンティティ特定システム10は、例えば、サーバ装置によって実現される。また、エンティティ特定システム10は、何らかのクライアントーサーバ型システム（例えば、対話システム）の一部であってもよいし、単体の装置であってもよい。

[0013] 引き続いて、本実施形態に係るエンティティ特定システム10の機能を説明する。図1に示すようにエンティティ特定システム10は、入力部11と、語句抽出部12と、候補変換部13と、組み合わせ生成部14と、スコア算出部15と、エンティティ特定部16とを備えて構成される。

[0014] 入力部 11 は、エンティティの特定対象の語句を含む文章を入力する機能部である。入力部 11 は、例えば、端末からエンティティ特定システム 10 に対して送信される文章を受信して入力する。あるいは、入力部 11 は、端末から音声を受信して、受信した音声を音声認識して、音声認識の結果である文章を取得して入力してもよい（即ち、音声データでの入力）。この場合、入力部 11 は、従来の任意の音声認識方法を用いて音声認識を行うことができる。また、入力部 11 は、予め設定された生成ルールに基づいて、ユーザの指示に応じて自動的に文章を音声データ又はテキストデータの形式で生成して入力することとしてもよい。また、入力部 11 は、上記以外の任意の方法で文章を入力することができる。入力部 11 は、入力した文章を語句抽出部 12 に出力する。

[0015] 語句抽出部 12 は、入力部 11 によって入力された文章から 1 つ以上の語句を抽出する機能部である。語句抽出部 12 によって抽出される語句は、エンティティがリンクされる対象となる語句を含む。また、語句抽出部 12 によって抽出される語句は、エンティティがリンクされる対象とならない語句を含んでいてもよい。後述するようにエンティティがリンクされる対象とならない語句も、エンティティの特定に用いられ得る。抽出される語句は、単語単位でもよいし、複数の単語からなる語句であってもよく、任意の単位の文字列でよい。抽出される語句は、1 つ以上であってもよいし、複数であってもよい。語句抽出部 12 は、例えば、以下のように語句を抽出する。

[0016] 語句抽出部 12 は、入力部 11 から文章を入力する。例えば、語句抽出部 12 は、形態素解析を用いて語句を抽出する。この場合、語句抽出部 12 は、入力した文章を形態素解析によって形態素に分割する。形態素解析自体は、従来の方法によって行うことができる。語句抽出部 12 は、文章を分割して得られた形態素全てを語句として抽出してもよい。あるいは、形態素のうちの一部を語句として抽出してもよい。具体的には、語句抽出部 12 は、形態素解析によって各形態素に付与された品詞に基づいて形態素を語句として抽出してもよい。例えば、語句として抽出する品詞（例えば、名詞）あるい

は語句として抽出しない品詞を予め設定しておいてもよい。

[0017] また、語句抽出部 12 は、コーパスを入力して、入力したコーパスに基づいて文章から語句を抽出することとしてもよい。コーパスとしては、例えば、オンライン百科事典（例えば、ウィキペディア）又はオンライン辞書等を用いることができる。コーパスの入力は、例えば、エンティティ特定システム 10 の管理者の操作によって行われる。具体的には、語句抽出部 12 は、コーパスに出現する語句の出現頻度を算出して、語句の出現頻度に基づいて語句を抽出してもよい。例えば、形態素解析によって得られた語句のうち、予め設定された出現頻度以上の語句を、一般的な語句であるとして抽出する語句から除外することとしてもよい。

[0018] また、語句抽出部 12 は、形態素解析にかえて、あるいは加えて予め記憶した語句抽出用の辞書を用いて語句を抽出してもよい。語句抽出用の辞書は、抽出すべき語句をリスト化したものである。語句抽出用の辞書は、エンティティ特定システム 10 の管理者等によって人工的に作成されたものであってもよい。あるいは、語句抽出用の辞書は、上述したコーパスに基づいて生成されたものであってもよい。例えば、コーパスに出現する語句のうち、予め設定された出現頻度未満の語句のリストを語句抽出用の辞書としてもよい。語句抽出部 12 は、語句抽出用の辞書に含まれる各語句と入力した文章とを比較して、文字列のマッチングを行い、文章に含まれる語句を抽出する。語句抽出部 12 は、抽出した語句を候補変換部 13 に出力する。

[0019] 図 2（a）に「合衆国最高裁判所は米政府の連邦裁判所を統括する」との文章から形態素解析によって抽出された語句の例を示す。図 2（b）に当該文章から語句抽出用の辞書によって抽出された語句の例を示す。例えば、「合衆国最高裁判所」は、形態素解析を用いた場合では、「合衆国」「最高」「裁判所」の 3 語の語句に分割されるが、語句抽出用の辞書を用いた場合では、辞書に「合衆国最高裁判所」との語句が含まれていれば、「合衆国最高裁判所」の 1 語の語句となる。以下では、語句抽出用の辞書を用いた場合の語句の例を用いて説明する。

- [0020] 候補変換部 13 は、語句抽出部 12 によって抽出された語句のうち少なくとも何れかについて、当該語句にリンクされるエンティティの 1 つ以上の候補の語句に変換する機能部である。候補変換部 13 は、例えば、以下のように語句をエンティティの候補の語句に変換する。
- [0021] 候補変換部 13 は、予め文章中出现し得る語句と当該語句にリンクされ得るエンティティを示す語句とを対応付けて記憶しておく。記憶されるエンティティを示す語句は、文章中の語句の変換候補、即ち、文章中出现し得る語句にリンクされるエンティティの候補の語句である。例えば、候補変換部 13 は、図 3 に示すように「連邦裁判所」との文章中出现し得る語句に対して、「アメリカ合衆国連邦裁判所」「連邦裁判所（ドイツ）」「連邦裁判所（スイス）」「オーストラリア連邦裁判所」等のエンティティを示す語句を対応付けて予め記憶しておく。文章中出现し得る語句 1 つに対して、エンティティの候補の語句は 1 つであってもよいし、複数であってもよい。
- [0022] 上記の情報は、エンティティ特定システム 10 の管理者等によって人工的に作成されたものであってもよい。あるいは、上記の情報は、上述したコーパスに基づいて生成されたものであってもよい。例えば、コーパスに含まれるアンカーテキストに基づいて生成されたものであってもよい。あるいは、コーパスに基づいて決定された語句間の文字列距離（例えば、後述するコサイン距離）に基づいて生成されたものであってもよい。
- [0023] 候補変換部 13 は、語句抽出部 12 から語句を入力する。候補変換部 13 は、語句抽出部 12 から入力した語句毎に、予め記憶した上記の情報に当該語句が含まれているか否かを確認する。候補変換部 13 は、予め記憶した情報に含まれている語句を、当該情報において当該語句に対応付けられたエンティティを示す語句に変換する。候補変換部 13 は、語句抽出部 12 によって抽出された語句毎に変換した後のエンティティの候補の語句を組み合わせ生成部 14 に出力する。また、語句抽出部 12 から入力した語句のうち、予め記憶した情報に含まれていないものについても、候補変換部 13 は当該（変換がされない）語句を組み合わせ生成部 14 に出力してもよい。記憶した

情報に含まれていない語句は、エンティティの特定の対象とならない語句である。

[0024] 組み合わせ生成部 14 は、候補変換部 13 によって変換された 1 つ以上の語句の何れかをそれぞれ含む、文章に対応する語句の組み合わせを 1 つ以上生成する機能部である。

[0025] 組み合わせ生成部 14 は、候補変換部 13 から語句を入力する。組み合わせ生成部 14 は、入力部 11 によって入力された文章毎、即ち、エンティティの特定対象の語句を含む文章毎に語句の組み合わせを生成する。組み合わせ生成部 14 は、1 つの組み合わせに対して、語句抽出部 12 によって抽出された語句毎に、候補変換部 13 によって変換されたエンティティの候補の語句の何れか 1 つを含める。組み合わせ生成部 14 は、全てのエンティティの候補の語句の組み合わせを生成する。これによって、変換後のエンティティの候補の語句の数の積の組み合わせが生成される。何れかの語句に対して複数のエンティティの候補の語句の数があれば、上記の組み合わせも複数になる。組み合わせの例を図 4 に示す。

[0026] 組み合わせ生成部 14 は、候補変換部 13 から入力された変換後のエンティティの候補の語句のうち、一部の語句のみを組み合わせの生成に用いることとしてもよい。具体的には、組み合わせ生成部 14 は、エンティティの候補の語句の文字列長又はコーパス中の当該語句の出現頻度によって、当該語句をフィルタリングし、フィルタリングした語句を組み合わせの生成に用いることとしてもよい。例えば、組み合わせ生成部 14 は、エンティティの候補の語句の文字列長が予め設定した範囲内である場合、あるいは、コーパス中の当該語句の出現頻度が予め設定した値以上、又は変換された語句のうち予め設定した順位以上である場合に語句を組み合わせの生成に用いることとしてもよい。文字列長をフィルタリングに用いるのは、例えば、機械的に抽出された文字列長が極端に短い又は長い候補の語句は、エンティティを示す語句として適切ではない場合があるためである。また、エンティティの候補の語句の文字列長及びコーパス中の当該語句の出現頻度の両方を用いてフィ

ルタリングを行ってもよい。このフィルタリングによって、例えば、「連邦裁判所」という語句から変換された複数の候補のうち、コーパス中の出現頻度に基づいて、「アメリカ合衆国連邦裁判所」と「連邦裁判所（ドイツ）」との2つだけが組み合わせの生成に用いられてもよい。

[0027] フィルタリングによって語句の候補の数を減らし、それによって語句の組み合わせの数を減らすことで計算量を削減することができる。例えば、文章中から3つの語句が抽出でき、それらの語句に対する候補の語句の数がそれぞれ3つ、5つ、3つであるとする、生成される組み合わせの数は $3 \times 5 \times 3 = 45$ 通りとなる。語句の候補をそれぞれ1つずつフィルタリングによって除外すれば、生成される組み合わせの数は $2 \times 4 \times 2 = 16$ 通りとなり、計算量を半分以上にすることができる。

[0028] 候補の語句のフィルタリングは、フィルタリングを行わない場合の語句の組み合わせの数に応じて行われてもよい。例えば、フィルタリングを行わない場合の語句の組み合わせの数が、予め設定した閾値以上になる場合に行うこととしてもよい。これによって、計算量の削減が必要であると考えられる場合に適切にフィルタリングを行うことができる。また、語句の候補のフィルタリングは、候補変換部13によって行われてもよい。また、候補変換部13は、フィルタリング後の候補の語句を、変換用の語句として予め記憶しておいてもよい。

[0029] 組み合わせ生成部14は、語句抽出部12によって抽出された全ての語句のエンティティの候補の語句を組み合わせの生成に用いてもよいし、一部の語句のエンティティの候補の語句を組み合わせの生成に用いてもよい。具体的には、組み合わせ生成部14は、語句抽出部12によって抽出された語句の品詞、又はコーパスに出現する語句の出現頻度に基づいて、組み合わせの生成に用いる語句を決定してもよい。例えば、語句抽出部12による語句の抽出の際と同様に品詞が用いられてもよい。あるいは、コーパス中の当該語句の出現頻度が予め設定した値以上、又は抽出された語句のうち予め設定した順位以上である場合に、当該語句のエンティティの候補の語句を組み合わせ

せの生成に用いることとしてもよい。また、語句の品詞、及びコーパスに出現する語句の出現頻度の両方に基づいて、組み合わせの生成に用いる語句を決定してもよい。これによって、例えば、「合衆国最高裁判所」、「米政府」及び「連邦裁判所」の3つの語句に対する候補のうち、「合衆国最高裁判所」及び「連邦裁判所」の2つの語句に対する候補だけが組み合わせの生成に用いられてもよい。上記のように語句の組み合わせの数を減らすことで、上述したフィルタリングと同様に計算量を削減することができる。なお、組み合わせの生成に用いる語句の決定（語句抽出部12による語句の抽出に相当）は、語句抽出部12及び組み合わせ生成部14の何れか一方のみで一律の基準で行われることとしてもよい。

[0030] 語句抽出部12によって抽出された語句の一部のみを用いた組み合わせの生成は、全ての語句を組み合わせの生成に用いる場合の語句の組み合わせの数に応じて行われてもよい。例えば、全ての語句を組み合わせの生成に用いる場合の語句の組み合わせの数が、予め設定した閾値以上になる場合に一部の語句のみを組み合わせの生成に用いることとしてもよい。これによって、計算量の削減が必要であると考えられる場合に適切に語句の削減を行うことができる。なお、この場合、組み合わせ生成部14による語句の削減を意義のあるものとするため、語句抽出部12による語句の抽出は、語句の品詞、又はコーパスに出現する語句の出現頻度を用いずに行うか、用いたとしても組み合わせ生成部14による語句の削減とは異なる（ゆるい）基準で行うこととしてもよい。

[0031] 候補変換部13から入力された語句に、エンティティの候補の語句に変換されていない語句が含まれている場合には、組み合わせ生成部14は、当該語句を含めて組み合わせを生成してもよい。組み合わせ生成部14は、生成した組み合わせを示す情報をスコア算出部15に出力する。

[0032] スコア算出部15は、組み合わせ生成部14によって生成された各組み合わせについて、組み合わせに含まれる語句同士の類似性のスコアに基づいてスコアを算出する機能部である。スコア算出部15は、コーパスを入力して

、入力したコーパスに基づいて語句同士の類似性のスコアを算出することとしてもよい。スコア算出部15は、例えば、以下のように各組み合わせについてスコアを算出する。

[0033] スコア算出部15は、組み合わせ生成部14から組み合わせを示す情報を入力する。スコア算出部15は、組み合わせに含まれる2つの語句同士の類似性のスコアを特定する。語句の類似性のスコアは、例えば、以下のように算出される。スコア算出部15は、コーパスを入力して、コーパスに基づいて2つの語句同士の類似性のスコアを算出する。コーパスに基づく語句同士の類似性のスコアの算出は、例えば、Word2Vec等の機械学習によって語句の解析を行う手法によって行うことができる。Word2Vecを用いる場合には、語句の特徴を示す単語ベクトル同士のコサイン距離を類似度とすることができる。あるいは、語句間の共起確率に基づいて、類似度が算出されてもよい。なお、コーパスに基づく類似度は、全ての語句の組み合わせについて予め算出されてスコア算出部15に記憶されていてもよい。また、語句同士の類似度は、上記以外の方法で算出されてもよく、あるいは、予め他の装置によって又は人工的に生成されたものが用いられてもよい。

[0034] スコア算出部15は、組み合わせに含まれる全ての2つの語句同士の類似性のスコアを算出する。スコア算出部15は、それらの類似性のスコアから、組み合わせ全体に対するスコアを算出する。例えば、スコア算出部15は、組み合わせに含まれる全ての2つの語句同士の類似性のスコアを足し合わせて、組み合わせ全体に対するスコアを算出する。スコア算出部15は、全ての組み合わせについてスコアを算出する。スコア算出部15は、組み合わせを示す情報及び算出したスコアをエンティティ特定部16に出力する。

[0035] エンティティ特定部16は、スコア算出部15によって算出された組み合わせのスコアに基づいて、1つ以上の候補の語句から、リンクされるエンティティの語句を特定する機能部である。

[0036] エンティティ特定部16は、スコア算出部15から組み合わせを示す情報及びスコアを入力する。スコアは、組み合わせに含まれるエンティティの候

補の語句の文章に対する妥当性を示すものである。例えば、上記の２つの語句同士の類似性のスコアの値が高い程、類似性が高いものであった場合、組み合わせのスコアが高い程、組み合わせに含まれるエンティティの候補の語句の文章に対する妥当性が高いことを示している。

[0037] エンティティ特定部１６は、各組み合わせのうち、スコアが、上記の妥当性が最も高いことを示すものである（例えば、スコアが最も高い）組み合わせに含まれるエンティティの候補の語句を、対応する語句にリンクされるエンティティの語句として特定する。また、エンティティ特定部１６は、スコアと予め設定された閾値とを比較して、スコアが閾値以上である場合にエンティティを特定することとしてもよい。スコアが閾値未満である場合、エンティティ特定部１６は、語句にリンクされるエンティティが（候補の中には）ないとしてもよい。上記のようにエンティティ特定部１６は、文章に含まれる語句一つ一つに対してエンティティを特定するのではなく、スコア（組み合わせの整合性）に基づいて文章に含まれる全ての語句に対して、リンクされるエンティティの語句を一度に特定する。

[0038] エンティティ特定部１６は、特定したエンティティの語句を、当該語句が用いられるシステム又はモジュール等へ出力する。なお、特定したエンティティの語句の出力は、任意の方法で行われてもよい。以上が、本実施形態に係るエンティティ特定システム１０の機能である。

[0039] 引き続いて、図５のフローチャートを用いて、本実施形態に係るエンティティ特定システム１０で実行される処理（エンティティ特定システム１０が行う動作方法）を説明する。本処理では、入力部１１によって、エンティティの特定対象の語句を含む文章が入力される（Ｓ０１）。続いて、語句抽出部１２によって、文章から語句が抽出される（Ｓ０２）。続いて、候補変換部１３によって、文章中の語句が、当該語句にリンクされるエンティティの候補の語句に変換される（Ｓ０３）。続いて、組み合わせ生成部１４によって、変換された語句をそれぞれ含む、文章に対応する語句の組み合わせが生成される（Ｓ０４）。続いて、スコア算出部１５によって、各組み合わせに

ついて、組み合わせに含まれる語句同士の類似性のスコアに基づいてスコアが算出される（S05）。続いて、エンティティ特定部16によって、組み合わせのスコアに基づいて、候補の語句からリンクされるエンティティの語句が特定されて出力される（S06）。以上が、本実施形態に係るエンティティ特定システム10で実行される処理である。

[0040] 本実施形態では、文章に対応する語句同士の類似性に基づいて、文章に含まれる語句にリンクされるエンティティの語句が特定される。従って、本実施形態によれば、文章の文脈に適したエンティティを特定することができる。また、語句同士の類似性を予め算出しておけば、従来のエンティティの特定と比べて比較的簡易な処理でエンティティを特定することができる。即ち、本実施形態によれば、エンティティの特定における処理負荷を低減することができる。

[0041] また、上述したようにコーパスに基づいて文章から語句を抽出することとしてもよい。この構成によれば、エンティティの特定対象となる語句を適切に抽出することができる。但し、語句の抽出には、必ずしもコーパスを用いる必要はない。

[0042] また、上述したようにコーパスに基づいて語句同士の類似性を算出することとしてもよい。この構成によれば、適切かつ確実に語句同士の類似性を算出することができ、その結果、適切かつ確実に文章の文脈に適したエンティティを特定することができる。但し、語句同士の類似性は、必ずしもコーパスに基づいたものとしなくてもよい。

[0043] なお、上記実施形態の説明に用いたブロック図は、機能単位のブロックを示している。これらの機能ブロック（構成部）は、ハードウェア及びソフトウェアの少なくとも一方の任意の組み合わせによって実現される。また、各機能ブロックの実現方法は特に限定されない。すなわち、各機能ブロックは、物理的又は論理的に結合した1つの装置を用いて実現されてもよいし、物理的又は論理的に分離した2つ以上の装置を直接的又は間接的に（例えば、有線、無線などを用いて）接続し、これら複数の装置を用いて実現されても

よい。機能ブロックは、上記１つの装置又は上記複数の装置にソフトウェアを組み合わせて実現されてもよい。

[0044] 機能には、判断、決定、判定、計算、算出、処理、導出、調査、探索、確認、受信、送信、出力、アクセス、解決、選択、選定、確立、比較、想定、期待、見做し、報知（broadcasting）、通知（notifying）、通信（communicating）、転送（forwarding）、構成（configuring）、再構成（reconfiguring）、割り当て（allocating、mapping）、割り振り（assigning）などがあるが、これらに限られない。たとえば、送信を機能させる機能ブロック（構成部）は、送信部（transmitting unit）又は送信機（transmitter）と呼称される。いずれも、上述したとおり、実現方法は特に限定されない。

[0045] 例えば、本開示の一実施の形態におけるエンティティ特定システム１０は、本開示の情報処理を行うコンピュータとして機能してもよい。図６は、本開示の一実施の形態に係るエンティティ特定システム１０のハードウェア構成の一例を示す図である。上述のエンティティ特定システム１０は、物理的には、プロセッサ１００１、メモリ１００２、ストレージ１００３、通信装置１００４、入力装置１００５、出力装置１００６、バス１００７などを含むコンピュータ装置として構成されてもよい。

[0046] なお、以下の説明では、「装置」という文言は、回路、デバイス、ユニットなどに読み替えることができる。エンティティ特定システム１０のハードウェア構成は、図に示した各装置を１つ又は複数含むように構成されてもよいし、一部の装置を含まずに構成されてもよい。

[0047] エンティティ特定システム１０における各機能は、プロセッサ１００１、メモリ１００２などのハードウェア上に所定のソフトウェア（プログラム）を読み込ませることによって、プロセッサ１００１が演算を行い、通信装置１００４による通信を制御したり、メモリ１００２及びストレージ１００３におけるデータの読み出し及び書き込みの少なくとも一方を制御したりすることによって実現される。

[0048] プロセッサ１００１は、例えば、オペレーティングシステムを動作させて

コンピュータ全体を制御する。プロセッサ１００１は、周辺装置とのインターフェース、制御装置、演算装置、レジスタなどを含む中央処理装置（CPU: Central Processing Unit）によって構成されてもよい。例えば、上述のエンティティ特定システム１０における各機能は、プロセッサ１００１によって実現されてもよい。

[0049] また、プロセッサ１００１は、プログラム（プログラムコード）、ソフトウェアモジュール、データなどを、ストレージ１００３及び通信装置１００４の少なくとも一方からメモリ１００２に読み出し、これらに従って各種の処理を実行する。プログラムとしては、上述の実施の形態において説明した動作の少なくとも一部をコンピュータに実行させるプログラムが用いられる。例えば、エンティティ特定システム１０における各機能は、メモリ１００２に格納され、プロセッサ１００１において動作する制御プログラムによって実現されてもよい。上述の各種処理は、１つのプロセッサ１００１によって実行される旨を説明してきたが、２以上のプロセッサ１００１により同時又は逐次に実行されてもよい。プロセッサ１００１は、１以上のチップによって実装されてもよい。なお、プログラムは、電気通信回線を介してネットワークから送信されても良い。

[0050] メモリ１００２は、コンピュータ読み取り可能な記録媒体であり、例えば、ROM (Read Only Memory)、EPROM (Erasable Programmable ROM)、EEPROM (Electrically Erasable Programmable ROM)、RAM (Random Access Memory) などの少なくとも１つによって構成されてもよい。メモリ１００２は、レジスタ、キャッシュ、メインメモリ（主記憶装置）などと呼ばれてもよい。メモリ１００２は、本開示の一実施の形態に係る情報処理を実施するために実行可能なプログラム（プログラムコード）、ソフトウェアモジュールなどを保存することができる。

[0051] ストレージ１００３は、コンピュータ読み取り可能な記録媒体であり、例えば、CD-ROM (Compact Disc ROM) などの光ディスク、ハードディスクドライブ、フレキシブルディスク、光磁気ディスク(例えば、コンパク

トディスク、デジタル多用途ディスク、Blu-ray（登録商標）ディスク）、スマートカード、フラッシュメモリ（例えば、カード、スティック、キードライブ）、フロッピー（登録商標）ディスク、磁気ストリップなどの少なくとも1つによって構成されてもよい。ストレージ1003は、補助記憶装置と呼ばれてもよい。エンティティ特定システム10が備える記憶媒体は、例えば、メモリ1002及びストレージ1003の少なくとも一方を含むデータベース、サーバその他の適切な媒体であってもよい。

[0052] 通信装置1004は、有線ネットワーク及び無線ネットワークの少なくとも一方を介してコンピュータ間の通信を行うためのハードウェア（送受信デバイス）であり、例えばネットワークデバイス、ネットワークコントローラ、ネットワークカード、通信モジュールなどともいう。

[0053] 入力装置1005は、外部からの入力を受け付ける入力デバイス（例えば、キーボード、マウス、マイクロフォン、スイッチ、ボタン、センサなど）である。出力装置1006は、外部への出力を実施する出力デバイス（例えば、ディスプレイ、スピーカー、LEDランプなど）である。なお、入力装置1005及び出力装置1006は、一体となった構成（例えば、タッチパネル）であってもよい。

[0054] また、プロセッサ1001、メモリ1002などの各装置は、情報を通信するためのバス1007によって接続される。バス1007は、単一のバスを用いて構成されてもよいし、装置間ごとに異なるバスを用いて構成されてもよい。

[0055] また、エンティティ特定システム10は、マイクロプロセッサ、デジタル信号プロセッサ（DSP: Digital Signal Processor）、ASIC（Application Specific Integrated Circuit）、PLD（Programmable Logic Device）、FPGA（Field Programmable Gate Array）などのハードウェアを含んで構成されてもよく、当該ハードウェアにより、各機能ブロックの一部又は全てが実現されてもよい。例えば、プロセッサ1001は、これらのハードウェアの少なくとも1つを用いて実装されてもよい。

- [0056] 本開示において説明した各態様／実施形態の処理手順、シーケンス、フローチャートなどは、矛盾の無い限り、順序を入れ替えてもよい。例えば、本開示において説明した方法については、例示的な順序を用いて様々なステップの要素を提示しており、提示した特定の順序に限定されない。
- [0057] 入出力された情報等は特定の場所（例えば、メモリ）に保存されてもよいし、管理テーブルを用いて管理してもよい。入出力される情報等は、上書き、更新、又は追記され得る。出力された情報等は削除されてもよい。入力された情報等は他の装置へ送信されてもよい。
- [0058] 判定は、1ビットで表される値（0か1か）によって行われてもよいし、真偽値（Boolean：true又はfalse）によって行われてもよいし、数値の比較（例えば、所定の値との比較）によって行われてもよい。
- [0059] 本開示において説明した各態様／実施形態は単独で用いてもよいし、組み合わせで用いてもよいし、実行に伴って切り替えて用いてもよい。また、所定の情報の通知（例えば、「Xであること」の通知）は、明示的に行うものに限られず、暗黙的（例えば、当該所定の情報の通知を行わない）ことによって行われてもよい。
- [0060] 以上、本開示について詳細に説明したが、当業者にとっては、本開示が本開示中に説明した実施形態に限定されるものではないということは明らかである。本開示は、請求の範囲の記載により定まる本開示の趣旨及び範囲を逸脱することなく修正及び変更態様として実施することができる。したがって、本開示の記載は、例示説明を目的とするものであり、本開示に対して何ら制限的な意味を有するものではない。
- [0061] ソフトウェアは、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語と呼ばれるか、他の名称で呼ばれるかを問わず、命令、命令セット、コード、コードセグメント、プログラムコード、プログラム、サブプログラム、ソフトウェアモジュール、アプリケーション、ソフトウェアアプリケーション、ソフトウェアパッケージ、ルーチン、サブルーチン、オブジェクト、実行可能ファイル、実行スレッド、手順、機能

などを意味するよう広く解釈されるべきである。

- [0062] また、ソフトウェア、命令、情報などは、伝送媒体を介して送受信されてもよい。例えば、ソフトウェアが、有線技術（同軸ケーブル、光ファイバケーブル、ツイストペア、デジタル加入者回線（DSL: Digital Subscriber Line）など）及び無線技術（赤外線、マイクロ波など）の少なくとも一方を使用してウェブサイト、サーバ、又は他のリモートソースから送信される場合、これらの有線技術及び無線技術の少なくとも一方は、伝送媒体の定義内に含まれる。
- [0063] 本開示において使用する「システム」及び「ネットワーク」という用語は、互換的に使用される。
- [0064] また、本開示において説明した情報、パラメータなどは、絶対値を用いて表されてもよいし、所定の値からの相対値を用いて表されてもよいし、対応する別の情報を用いて表されてもよい。
- [0065] サーバ及びクライアントの少なくとも一方は、送信装置、受信装置、通信装置などと呼ばれてもよい。なお、サーバ及びクライアントの少なくとも一方は、移動体に搭載されたデバイス、移動体自体などであってもよい。当該移動体は、乗り物（例えば、車、飛行機など）であってもよいし、無人で動く移動体（例えば、ドローン、自動運転車など）であってもよいし、ロボット（有人型又は無人型）であってもよい。なお、サーバ及びクライアントの少なくとも一方は、必ずしも通信動作時に移動しない装置も含む。例えば、基地局及び移動局の少なくとも一方は、センサなどのIoT (Internet of Things) 機器であってもよい。
- [0066] また、本開示におけるサーバは、クライアント端末で読み替えてもよい。例えば、サーバ及びクライアント端末間の通信を、複数のユーザ端末間の通信（例えば、D2D (Device-to-Device)、V2X (Vehicle-to-Everything) などと呼ばれてもよい）に置き換えた構成について、本開示の各態様／実施形態を適用してもよい。この場合、上述のサーバが有する機能をクライアント端末が有する構成としてもよい。

[0067] 同様に、本開示におけるクライアント端末は、サーバで読み替えてもよい。この場合、上述のクライアント端末が有する機能をサーバが有する構成としてもよい。

[0068] 本開示で使用する「判断(determining)」、「決定(determining)」という用語は、多種多様な動作を包含する場合がある。「判断」、「決定」は、例えば、判定(judging)、計算(calculating)、算出(computing)、処理(processing)、導出(deriving)、調査(investigating)、探索(looking up、search、inquiry) (例えば、テーブル、データベース又は別のデータ構造での探索)、確認(ascertaining)した事を「判断」「決定」したとみなす事などを含み得る。また、「判断」、「決定」は、受信(receiving) (例えば、情報を受信すること)、送信(transmitting) (例えば、情報を送信すること)、入力(input)、出力(output)、アクセス(accessing) (例えば、メモリ中のデータにアクセスすること)した事を「判断」「決定」したとみなす事などを含み得る。また、「判断」、「決定」は、解決(resolving)、選択(selecting)、選定(choosing)、確立(establishing)、比較(comparing)などした事を「判断」「決定」したとみなす事を含み得る。つまり、「判断」「決定」は、何らかの動作を「判断」「決定」したとみなす事を含み得る。また、「判断(決定)」は、「想定する(assuming)」、「期待する(expecting)」、「みなす(considering)」などで読み替えられてもよい。

[0069] 「接続された(connected)」、「結合された(coupled)」という用語、又はこれらのあらゆる変形は、2又はそれ以上の要素間の直接的又は間接的なあらゆる接続又は結合を意味し、互いに「接続」又は「結合」された2つの要素間に1又はそれ以上の中間要素が存在することを含むことができる。要素間の結合又は接続は、物理的なものであっても、論理的なものであっても、或いはこれらの組み合わせであってもよい。例えば、「接続」は「アクセス」で読み替えられてもよい。本開示で使用する場合、2つの要素は、1又はそれ以上の電線、ケーブル及びプリント電気接続の少なくとも一つを用いて、並びにいくつかの非限定的かつ非包括的な例として、無線周波数領域、マ

マイクロ波領域及び光（可視及び不可視の両方）領域の波長を有する電磁エネルギーなどを用いて、互いに「接続」又は「結合」され则认为することができ

[0070] 本開示において使用する「に基づいて」という記載は、別段に明記されていない限り、「のみに基づいて」を意味しない。言い換えれば、「に基づいて」という記載は、「のみに基づいて」と「に少なくとも基づいて」の両方を意味する。

本開示において、「含む（include）」、「含んでいる（including）」及びそれらの変形が使用されている場合、これらの用語は、用語「備える（comprising）」と同様に、包括的であることが意図される。さらに、本開示において使用されている用語「又は（or）」は、排他的論理和ではないことが意図される。

[0071] 本開示において、例えば、英語でのa, an及びtheのように、翻訳により冠詞が追加された場合、本開示は、これらの冠詞の後に続く名詞が複数形であることを含んでもよい。

[0072] 本開示において、「AとBが異なる」という用語は、「AとBが互いに異なる」ことを意味してもよい。なお、当該用語は、「AとBがそれぞれCと異なる」ことを意味してもよい。「離れる」、「結合される」などの用語も、「異なる」と同様に解釈されてもよい。

符号の説明

[0073] 10…エンティティ特定システム、11…入力部、12…語句抽出部、13…候補変換部、14…組み合わせ生成部、15…スコア算出部、16…エンティティ特定部、1001…プロセッサ、1002…メモリ、1003…ストレージ、1004…通信装置、1005…入力装置、1006…出力装置、1007…バス。

請求の範囲

[請求項1]

文章を入力する入力部と、

前記入力部によって入力された文章から1つ以上の語句を抽出する語句抽出部と、

前記語句抽出部によって抽出された語句のうち少なくとも何れかについて、当該語句にリンクされるエンティティの1つ以上の候補の語句に変換する候補変換部と、

前記候補変換部によって変換された1つ以上の語句の何れかをそれぞれ含む、前記文章に対応する語句の組み合わせを1つ以上生成する組み合わせ生成部と、

前記組み合わせ生成部によって生成された各組み合わせについて、組み合わせに含まれる語句同士の類似性のスコアに基づいてスコアを算出するスコア算出部と、

前記スコア算出部によって算出された組み合わせのスコアに基づいて、前記1つ以上の候補の語句から、前記リンクされるエンティティの語句を特定するエンティティ特定部と、

を備えるエンティティ特定システム。

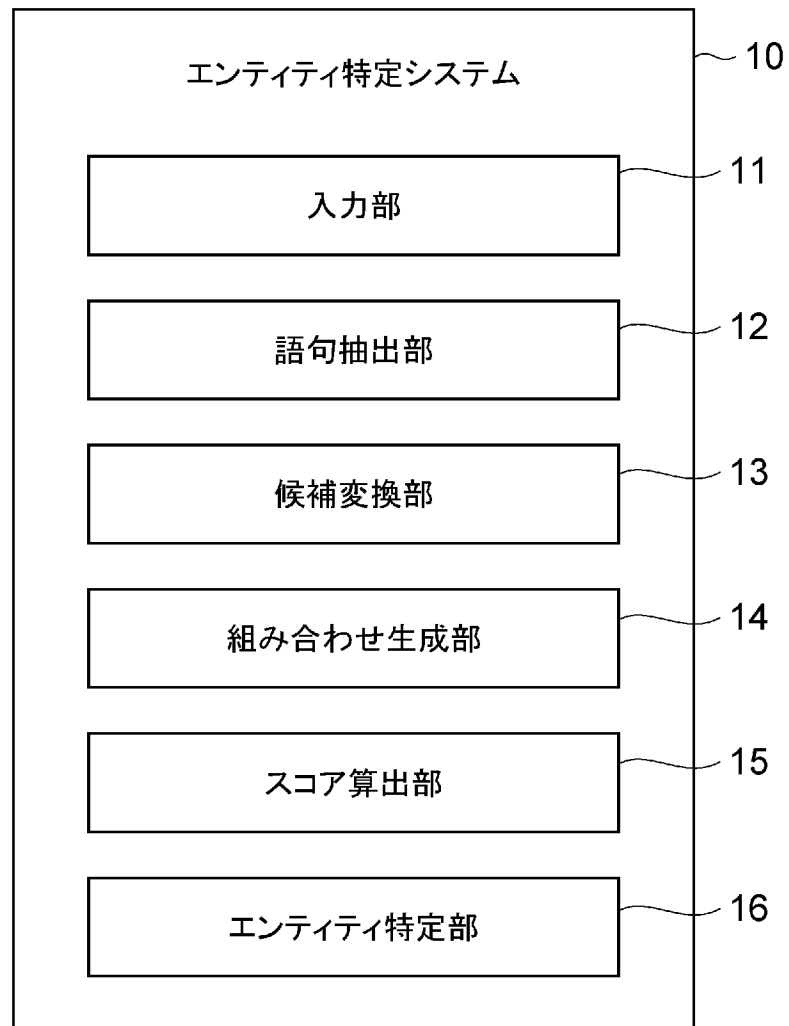
[請求項2]

前記語句抽出部は、コーパスを入力して、入力したコーパスに基づいて文章から語句を抽出する請求項1に記載のエンティティ特定システム。

[請求項3]

前記語句抽出部は、コーパスを入力して、入力したコーパスに基づいて語句同士の類似性のスコアを算出する請求項1又は2に記載のエンティティ特定システム。

[図1]



[図2]

(a)

合衆国最高裁判所は米政府の連邦裁判所を統括する



合衆国 / 最高 / 裁判所 / 米 / 政府 / 連邦 / 裁判所

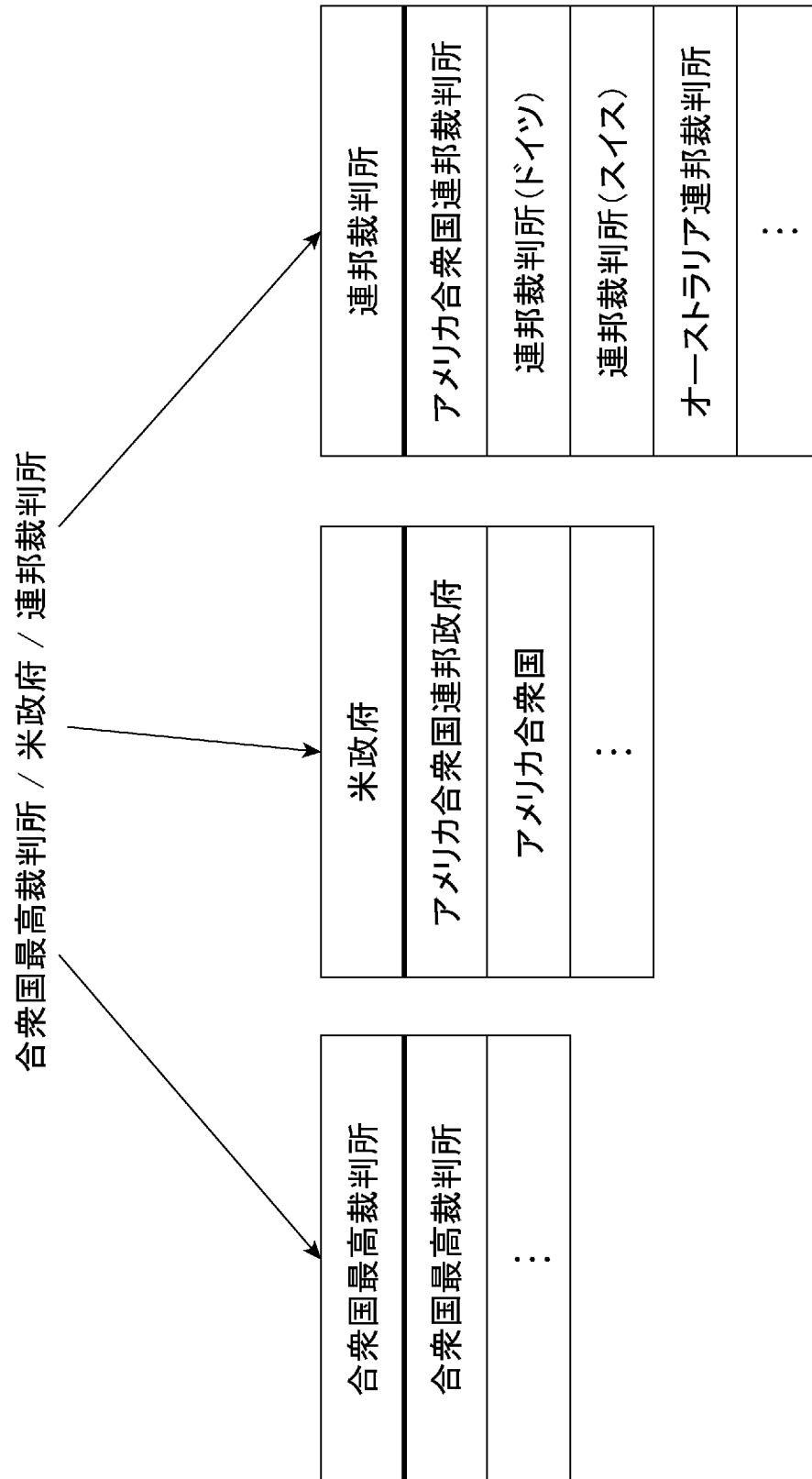
(b)

合衆国最高裁判所は米政府の連邦裁判所を統括する



合衆国最高裁判所 / 米政府 / 連邦裁判所

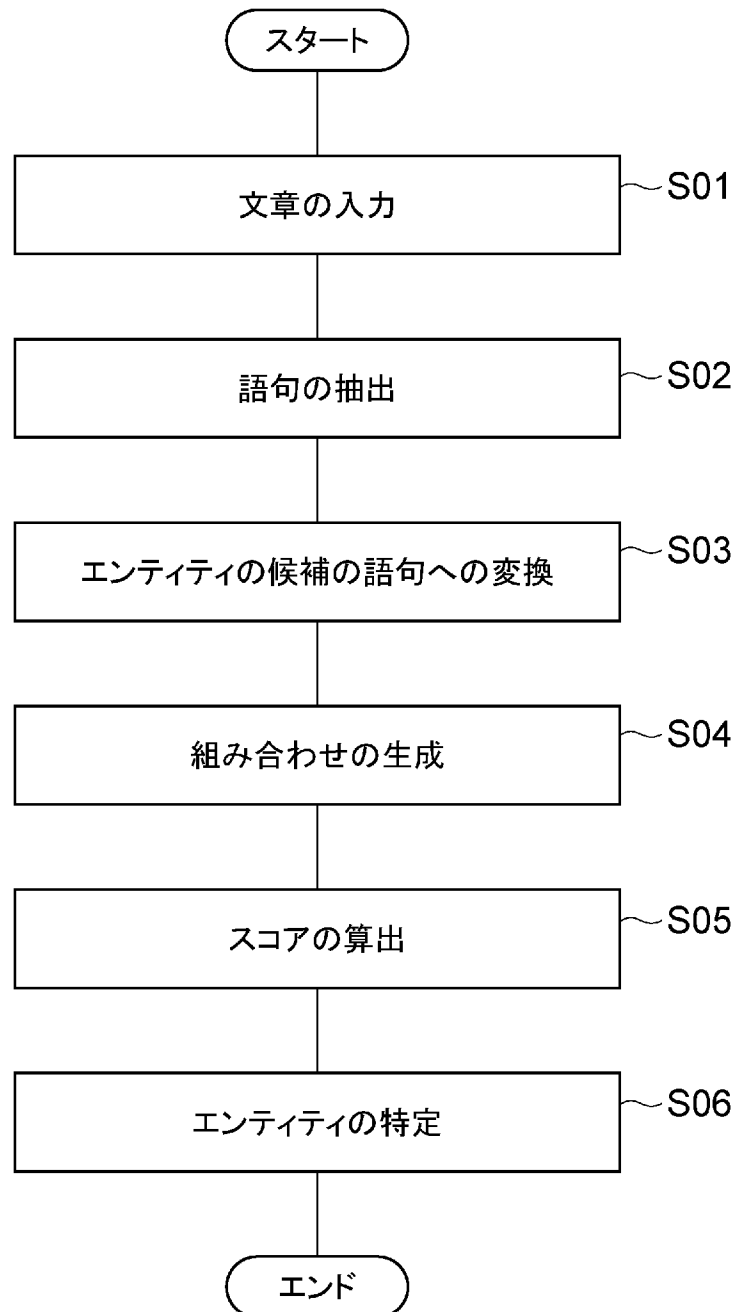
[図3]



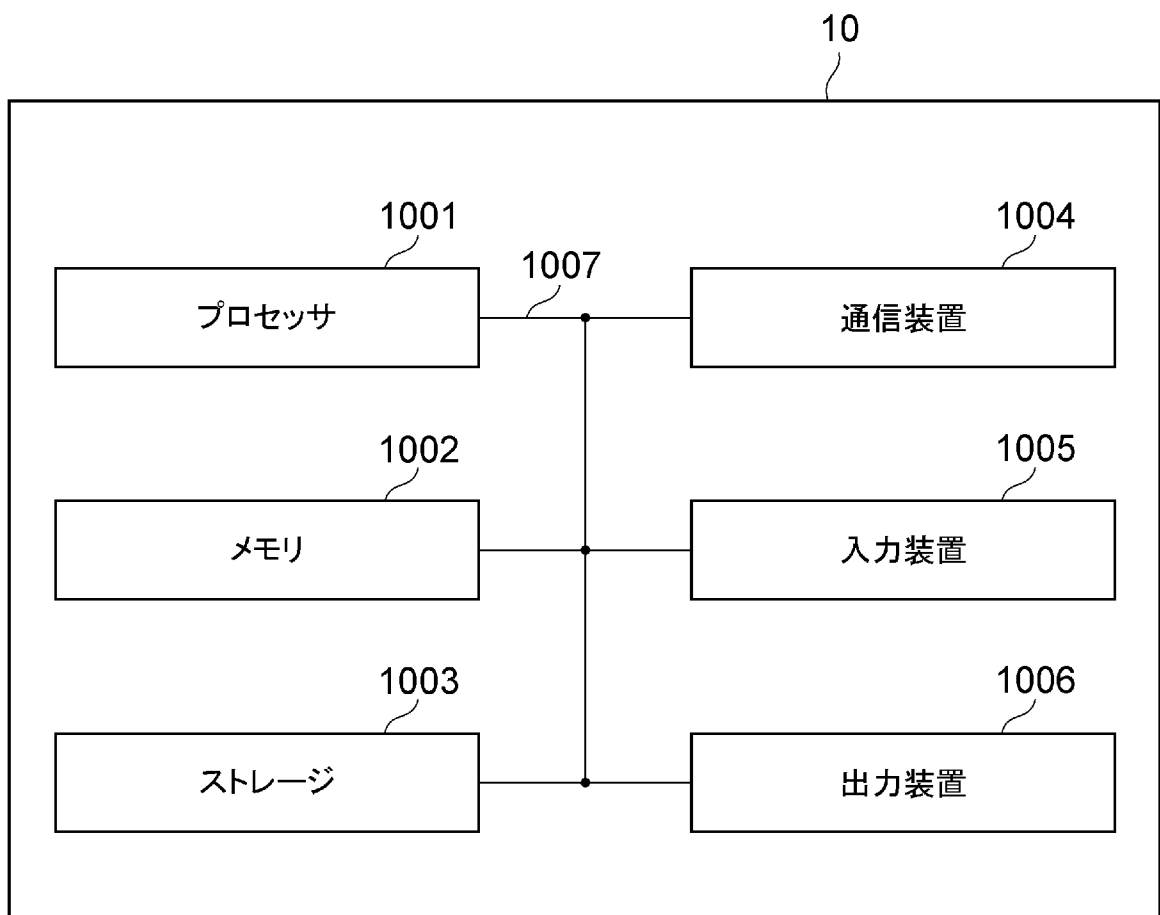
[図4]

エンティティ候補の 組み合わせ	合衆国最高裁判所	米政府	連邦裁判所	スコア
組み合わせ1	合衆国最高裁判所	アメリカ合衆国連邦政府	アメリカ合衆国連邦裁判所	0.XXX
組み合わせ2	合衆国最高裁判所	アメリカ合衆国連邦政府	連邦裁判所(ドイツ)	0.XXX
⋮	⋮	⋮	⋮	⋮
組み合わせN	合衆国最高裁判所	アメリカ合衆国	連邦裁判所(スイス)	0.XXX

[図5]



[図6]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/022218

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06F17/27 (2006.01) i, G06F16/383 (2019.01) 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)

Int.Cl. G06F17/27, G06F16/383

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2019
Registered utility model specifications of Japan	1996-2019
Published registered utility model applications of Japan	1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2017-199139 A (JAPAN BROADCASTING CORPORATION) 02 November 2017, entire text, all drawings (Family: none)	1-3
A	US 2016/0124939 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 05 May 2016, entire text, all drawings & CN 105630763 A	1-3

☐ 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 another 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
08 August 2019 (08.08.2019)

Date of mailing of the international search report
20 August 2019 (20.08.2019)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl. G06F17/27(2006.01)i, G06F16/383(2019.01)i

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int.Cl. G06F17/27, G06F16/383

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2017-199139 A (日本放送協会) 2017.11.02, 全文全図 (ファミリーなし)	1-3
A	US 2016/0124939 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016.05.05, 全文全図 & CN 105630763 A	1-3

C欄の続きにも文献が列挙されている。

パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの

「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)

「O」 口頭による開示、使用、展示等に言及する文献

「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」 同一パテントファミリー文献

国際調査を完了した日

08.08.2019

国際調査報告の発送日

20.08.2019

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

郵便番号 100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

長 由紀子

電話番号 03-3581-1101 内線 3586

5 N

4 2 3 3