

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7550756号
(P7550756)

(45)発行日 令和6年9月13日(2024.9.13)

(24)登録日 令和6年9月5日(2024.9.5)

(51)国際特許分類

F I

G 0 6 F 16/33 (2019.01)

G 0 6 F 40/247 (2020.01)

G 1 6 H 15/00 (2018.01)

G 0 6 F 16/33

G 0 6 F 40/247

G 1 6 H 15/00

請求項の数 19 (全32頁)

(21)出願番号	特願2021-529772(P2021-529772)	(73)特許権者	521227148
(86)(22)出願日	令和1年11月22日(2019.11.22)		アルゴテック システムズ リミテッド
(65)公表番号	特表2022-509199(P2022-509199 A)		イスラエル国 4 3 2 1 5 ラーナナ ハ
(43)公表日	令和4年1月20日(2022.1.20)		プニナ ストリート 2 ケフェア - サバ
(86)国際出願番号	PCT/EP2019/082288	(74)代理人	110001690
(87)国際公開番号	WO2020/109177		弁理士法人M & S パートナーズ
(87)国際公開日	令和2年6月4日(2020.6.4)	(72)発明者	アルメル オル
審査請求日	令和4年11月18日(2022.11.18)		オランダ国 5 6 5 6 アーエー アイン
(31)優先権主張番号	62/771,308		ドーフェン ハイ テック キャンパス 5
(32)優先日	平成30年11月26日(2018.11.26)	審査官	齊藤 貴孝
(33)優先権主張国・地域又は機関	米国(US)		

最終頁に続く

(54)【発明の名称】 放射線学レポート内の医学概念をマッチングさせるためのシステム及び方法

(57)【特許請求の範囲】

【請求項 1】

コンピュータ上で命令を実行することによって、医学概念のセットのうちのどの概念が入力テキストに関連しているかを自動的に決定する方法であって、前記方法は、

a) 2つ以上の単語を有する各概念について、1つ以上のクエリのセットを作成するステップであって、各クエリは、前記概念内の前記単語と同じ順序の、前記概念内の前記単語のサブストリングである2つの単語のストリングである、作成するステップと、

b) 2つ以上の単語を有する前記概念の選択されたサブセットにおける各概念について、各クエリに対して、前記入力テキストに当該クエリのすべての単語が含まれているか否かを決定し、含まれている場合は、当該クエリと前記入力テキストとの間のマッチングの程度を示すサブスコアを計算するステップと、

c) 最小数よりも多い前記クエリが、基準に従って十分に互いに近い前記入力テキストにおけるそれらの単語のすべてを有する前記選択されたサブセットにおける各概念について、前記概念についての前記クエリの前記サブスコアに依存して、前記概念と前記入力テキストとの間のマッチングの程度を示すスコアを計算するステップと、

d) スコアが計算された、前記選択されたサブセットにおける前記概念のうちのどれが前記入力テキストに関連し、どれが関連しないかを決定するために、1つ以上の規則を適用するステップであって、前記規則のうちの少なくともいくつかは、前記概念の前記スコアに依存している、適用するステップと、

を含む、方法。

【請求項 2】

少なくとも、最小数よりも多い単語が前記入力テキスト内にあり、かつ、最小数よりも多いクエリが、基準に従って十分に互いに近い前記入力テキスト内のそれらの単語のすべてを有する、2つ以上の単語を有する前記選択されたサブセットにおける各概念について、マッチングスコアを計算するステップをさらに含み、前記マッチングスコアは、バッグ・オブ・ワーズ方法に従った前記概念と前記入力テキストとの間のマッチングの程度を示し、各概念について前記スコアを計算するステップは、前記概念の前記マッチングスコアに依存して、及び、前記概念の前記クエリの前記サブスコアに依存して、前記スコアを計算するステップを含む、請求項 1 に記載の方法。

【請求項 3】

各概念について前記スコアを計算するステップは、前記マッチングスコアと前記クエリの前記サブスコアとの加重和を計算するステップを含む、請求項 2 に記載の方法。

【請求項 4】

単語の前記最小数は、2つの単語を有する概念については2であり、3つの単語を有する概念については2又は3であり、4つの単語を有する概念については2、3又は4であり、5つの単語を有する概念については3又は4であり、6つの単語を有する概念については3、4又は5である、請求項 2 に記載の方法。

【請求項 5】

1つの単語のみを有する概念に、前記1つの単語が前記入力テキスト内に見つかけられるときに、スコアを割り当てるステップをさらに含み、前記概念のスコアに依存する前記規則は、前記1つの単語のみを有する概念と、2つ以上の単語を有する概念との両方に適用される、請求項 2 に記載の方法。

【請求項 6】

スコアが計算された2つの概念がそれらの単語において十分に大きい重複を有するときに、前記1つ以上の規則は、より低いスコアを有する概念が前記入力テキストに関連しておらず、また、当該概念が、その単語において十分に大きい重複を有する他の任意の概念について計算されたスコアよりも高い計算されたスコアを有する場合に、前記概念は前記入力テキストに関連していると規定する、請求項 1 又は 5 に記載の方法。

【請求項 7】

第1の最小数は、前記概念について作成されたクエリの数の35%～65%である、請求項 1 に記載の方法。

【請求項 8】

前記スコアを計算するステップは、前記概念においてより遠く離れている単語を有するクエリにより低い重みを与えて、前記クエリの前記サブスコアの加重和を計算するステップを含む、請求項 1 に記載の方法。

【請求項 9】

前記概念の選択されたサブセットは、少なくとも、その概念の必須単語として定義された1つ以上の単語が前記入力テキスト内に見つからない概念を除外している、請求項 1 に記載の方法。

【請求項 10】

前記概念のセットにおけるすべての単語の中で最もまれな前記概念内の単語が、必須単語として定義される、請求項 9 に記載の方法。

【請求項 11】

病名である前記概念内の如何なる単数形の単語も、必須単語として定義される、請求項 9 に記載の方法。

【請求項 12】

前記概念の選択されたサブセットは、少なくとも、身体部分の単語、及び前記身体部分の場所又は向きを記述する単語を含み、前記入力テキストにおいて、前記身体部分の前記場所又は向きを記述する単語が、前記身体部分の単語から1単語よりも多く離れている概念を除外している、請求項 1 に記載の方法。

10

20

30

40

50

【請求項 1 3】

前記入力テキスト内で十分に互いに近い前記クエリ内の単語についての前記基準は、前記単語間の最大距離を規定し、前記最大距離は、前記入力テキストにおいて前記クエリ内にあるのと同じ順序ではない単語についてはより小さく、前記概念において互により近くに離間している単語についてよりも前記概念においてより遠くに離間している単語についてはより大きい、請求項 1 に記載の方法。

【請求項 1 4】

前記概念において隣接し、前記クエリ及び前記入力テキストにおいて同じ順序である単語についての前記最大距離は、10～25である、請求項 1 3 に記載の方法。

【請求項 1 5】

前記概念のセットを準備するステップと、前記入力テキストを前処理するステップと、をさらに含み、

a) 医学概念のデータベースからの概念の初期セットを提供するステップと、
b) 単語「椎骨」及び「脊柱」を含むように椎骨の文字 - 番号指定を拡張し、前記文字 - 番号指定の文字を、それが表す身体領域、つまり、頸部、胸部、又は腰部によって置き換えることによって、前記概念の初期セットを修正するステップと、

c) 単語「椎骨」及び「脊柱」を含むように椎骨の文字 - 番号指定を拡張し、前記文字 - 番号指定の文字を、それが表す身体領域、つまり、頸部、胸部、又は腰部によって置き換えることによって、前記入力テキストを前処理するステップと、

を含む、請求項 1 に記載の方法。

【請求項 1 6】

1 つ以上のコンピュータに請求項 1 に記載の方法を行わせるための命令が記憶された、少なくとも 1 つのコンピュータ記憶媒体。

【請求項 1 7】

コンピュータに請求項 1 に記載の方法を行わせるための命令が記憶されたコンピュータ記憶媒体。

【請求項 1 8】

請求項 1 に記載の方法のステップを行うためのコンピュータ可読媒体。

【請求項 1 9】

医学概念のセットのうちのどの概念が医用レポート内に見つけられるかを自動的に特定するためのシステムであって、

a) 1 つ以上の医用レポートへのアクセスを提供する第 1 のデータベースと、
b) 医用レポートを入力テキストに分割する、前記第 1 のデータベースへのアクセスを有するスプリッタモジュールと、

c) 前記医学概念のセットへのアクセスを提供する第 2 のデータベースと、
d) 前記入力テキスト及び前記第 2 のデータベースへのアクセスを有するプロセッサモジュールであって、各入力テキストについて、

1) 2 つ以上の単語を有する各概念について、それぞれ、前記概念内の前記単語と同じ順序の、前記概念内の前記単語のサブストリングである 2 つの単語のストリングである 1 つ以上のクエリのセットを作成し、

2) 2 つ以上の単語を有する前記概念の選択されたサブセットにおける各概念について、各クエリに対して、前記入力テキストに当該クエリのすべての単語が含まれているか否かを決定し、含まれている場合は、前記クエリと前記入力テキストとの間のマッチングの程度を示すサブスコアを計算し、

3) 最小数よりも多い前記クエリが、基準に従って十分に互いに近い前記入力テキストにおけるそれらの単語のすべてを有する前記選択されたサブセットにおける各概念について、前記概念についての前記クエリの前記サブスコアに依存して、前記概念と前記入力テキストとの間のマッチングの程度を示すスコアを計算し、

4) スコアが計算された、前記選択されたサブセットにおける前記概念のうちのどれが前記入力テキストに関連し、どれが関連しないかを決定するために、そのうちの少なく

10

20

30

40

50

ともいくつかは前記概念の前記スコアに依存している1つ以上の規則を適用する、プロセスサモジュールと、

e) 前記医用レポートの前記入力テキストに関連すると決定された前記概念を出力する出力モジュールと、

を含む、システム。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願

本願は、Or Almerの名義で、2018年11月26日に仮出願された「SYSTEM AND METHOD FOR MATCHING MEDICAL CONCEPTS IN RADIOLOGICAL REPORTS」なる名称の米国仮出願第62/771,308号の利益を主張し、その全体が本明細書に組み込まれる。

【0002】

本開示は、概して、自然言語処理の分野に関し、特に、事前に定義された医学概念のセットからどの医学概念が医用レポート内に見つけられるかを特定することに関する。より具体的には、しかし、排他的ではないが、本開示は、放射線学レポートを用いてこれを行うための方法に関する。

【背景技術】

【0003】

所与の概念が所与の入力テキスト内に見つけられるか否かを自動的に決定するための多くの検索アルゴリズムが存在する。例えば、概念は、検索用語、つまり、検索エンジンに入力された単語のストリングであり、入力テキストは、検索エンジンが検索する多くのウェブページのうちの1つに見つけられるテキストである。いくつかの検索アルゴリズムでは、検索アルゴリズムが良い結果を返すために、検索用語が入力テキスト内にまったく同じ形式で見つけられる必要はなく、検索用語の一部の単語が入力テキストにないか、又は検索用語の単語が入力テキスト内に異なる順序で見つけられるか、及び/又はそれらの間に他の単語があって入力テキスト内で見つけられる可能性がある。この場合、検索アルゴリズムは、マッチングの程度を示す当該検索用語及び入力テキストのスコアを計算し、検索エンジンは、検索しているウェブページの検索を終了すると、当該検索用語について良い結果が見つかったウェブページをスコア順にランク付けしたリストを提供する。

【0004】

例えば、ElasticSearchは、<https://www(dot)elastic(dot)co/products/ElasticSearch>に説明される市販の検索エンジンであり、<https://www(dot)elastic(dot)co/guide/en/elasticsearch/guide/current/index(dot)html>にユーザガイドがある。ElasticSearchは、文書のタイトルフィールドなどのテキストフィールドを、様々なタイプのクエリを使用して解析して、文書の検索用語との関連性スコアを計算する。「マッチング」クエリでは、検索用語内の最小数よりも多い単語又は最小割合よりも多い単語がテキストフィールド内に見つかった場合、テキストフィールドは検索用語とのマッチングであると見なされ、テキストフィールド内に見つけられる検索用語内の単語が多いほど、「マッチング」関連性スコアは高くなる。ElasticSearchの「マッチング」クエリは、入力テキストフィールドにおける語順は重要でない「バッグ・オブ・ワーズ(bag of words)」方法の一例である。

【0005】

ElasticSearchの「マッチングフレーズ」クエリでは、検索用語のすべての単語がテキストフィールド内に見つけられ、かつ、検索用語内の単語の相対位置がテキストフィールド内の同じ単語の相対位置から離れすぎしていない場合に、テキストフィールドは検索用語とのマッチングであると見なされる。単語の位置がどれだけ離れているかは

10

20

30

40

50

、検索用語内の単語がテキストフィールド内でそれらの単語が持つ位置と同じ位置を持つようにするために、いずれかの方向に1だけ単語の位置を変更して、検索用語内の単語の位置を何回変更しなければいけないのかを意味する。例えば、検索用語が2つの単語からなり、これらの2つの単語が、テキストフィールド内で互いに隣接しているが逆の順序で見つけられる場合、位置の変更回数は2である。これは、2つの単語の各々が1回の位置変更を経なければならないからである。2つの単語が、テキストフィールド内で検索用語におけるのと同じ順序であるが、それらの間にn個の他の単語があって見つけられる場合、必要な位置変更回数はnである。テキストフィールドが検索用語とのマッチングであると思なされるためには、必要な位置変更回数は、最大回数(「スロップ(slop)」と呼ばれるパラメータ)以下でなければならない。「マッチングフレーズ」クエリは、必要とされる位置変更回数が少なければ、高い関連性スコアを有する。「マッチング」関連性スコア及び「マッチングフレーズ」関連性スコアは足し合わされて、検索用語とのテキストフィールドの全体的な関連性スコアが得られる。

10

【0006】

検索アルゴリズムが、テキストフィールド及び/又は検索用語内の単語をそれらの語幹に置き換えることが知られている。例えば、Elastic Searchは、これを英語のテキストに対して行う英語語幹化特徴を有する。

【0007】

Savovaらの「Mayo clinical Text Analysis and Knowledge Extraction System(cTAKES): architecture, component evaluation and applications」、J. Am. Med. Inform. Assoc. 2010:17:507-513 (doi:10.1136/jamia.2009.001560)は、電子医療記録臨床フリーテキストから情報を抽出するためのcTAKESシステムについて説明し、この目的を有する他のシステムの歴史を提供している。

20

【0008】

<https://browser(dot)ihtsdotools(dot)org>にあるSNOMED CT医学辞書は、身体部分、疾患、及び疾患を診断するために使用され得る症状を含む多くの医学概念の構造化されたリストを有する。各概念は、1つ以上の単語で構成されるストリングによって定義されている。場合によっては、複数の概念が同じものを指すことがあり、この場合、好ましい用語及び代替の用語が与えられる。概念は、より一般的な概念を指す親、及びより具体的な概念を指す子供を有し得る。例えば、肺(lung)は、肺部(lung part)の親であり、肺部(lung part)は、肺葉(lobe of lung)の親である。疾患概念は、身体部分の概念との所見部位関係を有し、例えば、肺癌(lung cancer)は、肺(lung)との所見部位関係を有する。

30

【発明の概要】

【0009】

本発明のいくつかの実施形態の一態様は、医学概念のセットのうちのどれが医用レポートからの入力テキストに関連するかを決定するための方法に関し、概念は、概念内の十分に多い数の単語のペアが両方とも、十分に互いに近接して入力テキスト内に見つけられる場合に、入力テキストとの可能なマッチングとみなされる。

40

【0010】

本開示の一態様によれば、コンピュータ上で命令を実行することによって、医学概念のセットのうちのどの概念が入力テキストに関連しているかを自動的に決定する方法が提供される。方法は、

a) 2つ以上の単語を有する各概念について、1つ以上のクエリのセットを作成するステップであって、各クエリは、概念内の単語と同じ順序の、概念内の単語のサブストリングである2つの単語のストリングである、作成するステップと、

b) 2つ以上の単語を有する概念の選択されたサブセットにおける各概念について、各

50

クエリに対して、入力テキストにクエリのすべての単語が含まれているか否かを決定し、含まれている場合は、クエリと入力テキストとの間のマッチングの程度を示すサブスコアを計算するステップと、

c) 最小数よりも多いクエリが、基準に従って十分に互いに近い入力テキストにおけるそれらの単語のすべてを有する選択されたサブセットにおける各概念について、概念についてのクエリのサブスコアに依存して、概念と入力テキストとの間のマッチングの程度を示すスコアを計算するステップと、

d) スコアが計算された、選択されたサブセットにおける概念のうちのどれが入力テキストに関連し、どれが関連しないかを決定するために、1つ以上の規則を適用するステップであって、規則のうちの少なくともいくつかは、概念のスコアに依存している、適用するステップとを含む。

10

【0011】

任意選択で、方法はまた、少なくとも、最小数よりも多い単語が入力テキスト内にあり、かつ、最小数よりも多いクエリが、基準に従って十分に互いに近い入力テキスト内のそれらの単語のすべてを有する、2つ以上の単語を有する選択されたサブセットにおける各概念について、マッチングスコアを計算するステップを含み、マッチングスコアは、バッグ・オブ・ワーズ方法に従って概念と入力テキストとの間のマッチングの程度を示し、各概念についてスコアを計算するステップは、概念のマッチングスコアに依存して、並びに、概念のクエリのサブスコアに依存してスコアを計算するステップを含む。

【0012】

20

任意選択で、各概念についてスコアを計算するステップは、マッチングスコアとクエリのサブスコアとの加重和を計算するステップを含む。

【0013】

任意選択で、単語の最小数は、2つの単語を有する概念については2であり、3つの単語を有する概念については2又は3であり、4つの単語を有する概念については2、3又は4であり、5つの単語を有する概念については3又は4であり、6つの単語を有する概念については3、4又は5である。

【0014】

任意選択で、方法はまた、1つの単語のみを有する概念に、当該1つの単語が入力テキスト内に見つかけられるときに、スコアを割り当てるステップをさらに含み、概念のスコアに依存する規則は、1つの単語のみを有する概念と、2つ以上の単語を有する概念との両方に適用される。

30

【0015】

任意選択で、スコアが計算された2つの概念がそれらの単語において十分に大きい重複を有するときに、1つ以上の規則は、より低いスコアを有する概念は入力テキストに関連しておらず、また、概念が、その単語において十分に大きい重複を有する他の任意の概念について計算されたスコアよりも高い計算されたスコアを有する場合に、当該概念は入力テキストに関連していると規定する。

【0016】

任意選択で、第1の最小数は、その概念について作成されたクエリの数の35%~65%である。

40

【0017】

任意選択で、スコアを計算するステップは、概念に対してより遠く離れている単語を有するクエリにより低い重みを与えて、クエリのサブスコアの加重和を計算するステップを含む。

【0018】

任意選択で、概念の選択されたサブセットは、少なくとも、その概念の必須単語として定義された1つ以上の単語が入力テキスト内に見つからない概念を除外している。

【0019】

任意選択で、概念のセットにおけるすべての単語の中で最もまれな概念内の単語が、必

50

須単語として定義される。

【0020】

任意選択で、病名である概念内の任意の単数形の単語が、必須単語として定義される。

【0021】

任意選択で、概念の選択されたサブセットは、少なくとも、身体部分の単語、及び身体部分の場所又は向きを記述する単語を含み、入力テキストにおいて、身体部分の場所又は向きを記述する単語が身体部分の単語から1単語よりも多く離れている概念を除外している。

【0022】

任意選択で、入力テキスト内で十分に互いに近いクエリ内の単語についての基準は、単語間の最大距離を規定し、最大距離を規定は、入力テキストにおいてクエリ内にあるのと同じ順序ではない単語についてはより小さく、概念において互いにより近くに離間している単語についてよりも概念においてより遠くに離間している単語についてはより大きい。

【0023】

任意選択で、概念において隣接し、クエリ及び入力テキストにおいて同じ順序である単語についての最大距離は、10～25である。

【0024】

任意選択で、方法はまた、概念のセットを準備するステップと、入力テキストを前処理するステップとを含み、

a) 医学概念のデータベースからの概念の初期セットを提供するステップと、

b) 単語「vertebra (椎骨)」及び「spine (脊柱)」を含むように椎骨の文字 - 番号指定を拡張し、文字 - 番号指定の文字を、それが表す身体領域、つまり、頸部 (cervical)、胸部 (thoracic)、又は腰部 (lumbar) によって置き換えることによって、概念の初期セットを修正するステップと、

c) 単語「vertebra (椎骨)」及び「spine (脊柱)」を含むように椎骨の文字 - 番号指定を拡張し、文字 - 番号指定の文字を、それが表す身体領域、つまり、頸部 (cervical)、胸部 (thoracic)、又は腰部 (lumbar) によって置き換えることによって、入力テキストを前処理するステップとを含む。

【0025】

本開示の別の態様によれば、1つ以上のコンピュータに本発明の例示的な方法を行わせるための命令が記憶された少なくとも1つのコンピュータ記憶媒体を有するコンピュータ記憶製品が提供される。

【0026】

本開示の別の態様によれば、コンピュータに本発明の例示的な方法を行わせるための命令が記憶されたコンピュータ記憶媒体が提供される。

【0027】

本開示の別の態様によれば、本発明の例示的な方法のステップを行うためのコンピュータ可読媒体に具現化されたコンピュータ製品が提供される。

【0028】

本開示の別の態様によれば、医学概念のセットのうちのどの概念が医用レポート内に見つけられるかを自動的に特定するためのシステムが提供される。システムは、

a) 1つ以上の医用レポートへのアクセスを提供する第1のデータベースと、

b) 医用レポートを入力テキストに分割する、第1のデータベースへのアクセスを有するスプリッタモジュールと、

c) 医学概念のセットへのアクセスを提供する第2のデータベースと、

d) 入力テキスト及び第2のデータベースへのアクセスを有するプロセッサモジュールであって、各入力テキストについて、

1) 2つ以上の単語を有する各概念について、それぞれ、概念内の単語と同じ順序の、概念内の単語のサブストリングである2つの単語のストリングである1つ以上のクエリのセットを作成し、

10

20

30

40

50

2) 2つ以上の単語を有する概念の選択されたサブセットにおける各概念について、各クエリについて、入力テキストにクエリのすべての単語が含まれているか否かを決定し、含まれている場合は、クエリと入力テキストとの間のマッチングの程度を示すサブスコアを計算し、

3) 最小数よりも多いクエリが、基準に従って十分に互いに近い入力テキストにおけるそれらの単語のすべてを有する選択されたサブセットにおける各概念について、概念についてのクエリのサブスコアに依存して、概念と入力テキストとの間のマッチングの程度を示すスコアを計算し、

4) スコアが計算された、選択されたサブセットにおける概念のうちのどれが入力テキストに関連し、どれが関連しないかを決定するために、そのうちの少なくともいくつかは概念のスコアに依存している1つ以上の規則を適用する、プロセッサモジュールと、

e) 医用レポートの入力テキストに関連すると決定された概念を出力する出力モジュールとを含む。

【0029】

本開示の別の態様によれば、コンピュータ上で命令を実行することによって、医学概念のセットのうちのどの概念が入力テキストに関連しているかを自動的に決定する方法が提供される。方法は、

a) セット内の各概念について、概念が入力テキストに対する可能なマッチングであるか否かを決定するための1つ以上の基準を適用するステップであって、基準は、概念の少なくともいくつかについて、概念の単語のすべてが入力テキスト内に見つけられることを必要としないが、概念の必須単語として規定された1つ以上の単語が入力テキスト内に見つけられることを必要とする、適用するステップと、

b) 入力テキストに対する可能なマッチングである概念のうちのどれが入力テキストに関連しているか、どれが関連していないかを決定するために1つ以上の規則を適用するステップとを含む。

【0030】

任意選択で、方法はまた、入力テキストとの可能なマッチングである各概念について、概念と入力テキストとの間のマッチングの程度を示すスコアを計算するステップを含み、1つ以上の規則のうちの少なくともいくつかは、概念のスコアと、入力テキストとの可能なマッチングである任意の他の概念のスコアとに依存する。

【0031】

任意選択で、セット内のすべての概念の中で最も一般的でない概念の単語が、その概念の必須単語であるように規定される。

【0032】

任意選択で、疾患の1単語用語である概念の単語が、その概念の必須単語であるように規定される。

【0033】

任意選択で、身体部分の単語、及び身体部分の場所又は向きを記述する単語を含む概念について、両方の単語は、その概念の必須単語であるように規定され、基準はさらに、単語が入力テキストにおいて互いに隣接することを必要とする。

【0034】

本開示の別の態様によれば、コンピュータ上で命令を実行することによって、医学概念のセットのうちのどの概念が入力テキストに関連しているかを自動的に決定する方法が提供される。方法は、

a) セット内の各概念について、概念が入力テキストに対する可能なマッチングであるか否かを決定するための基準を適用するステップと、

b) 入力テキストとの可能なマッチングである各概念について、概念と入力テキストとの間のマッチングの程度を示すスコアを計算するステップと、

c) スコアが計算された2つの概念がそれらの単語において十分に大きい重複を有するときに、より低いスコアを有する概念は入力テキストに関連しておらず、また、概念が、

10

20

30

40

50

その単語において十分に大きい重複を有する他の如何なる概念について計算されたスコアよりも高い計算されたスコアを有する場合に、概念が入力テキストに関連していると決定するステップとを含む。

【 0 0 3 5 】

任意選択で、セット内の概念の少なくともいくつかについて、スコアを計算するステップは、マッチングスコアを計算するステップと、マッチングフレーズスコアを計算するステップと、マッチングスコアとマッチングフレーズスコアとを組み合わせ、スコアを取得するステップとを含み、マッチングスコアは、入力テキスト内に見つけられるが、入力テキスト内の単語の順序にはない概念内の単語の数に依存し、マッチングフレーズスコアは、入力テキスト内に見つけられ、かつ、入力テキスト内の単語の順序にある概念内の単語の数に依存する。

10

【 0 0 3 6 】

本開示の別の態様によれば、医学概念のセットのうちのどの概念が医用レポート内に見つけられるかを自動的に特定するためのシステムが提供される。システムは、

- a) 1 つ以上の医用レポートへのアクセスを提供する第 1 のデータベースと、
- b) 医用レポートを入力テキストに分割する、第 1 のデータベースへのアクセスを有するスプリッタモジュールと、
- c) 医学概念のセットへのアクセスを提供する第 2 のデータベースと、
- d) 入力テキスト及び第 2 のデータベースへのアクセスを有するプロセッサモジュールであって、各入力テキストについて、

20

1) セット内の各概念について、概念が入力テキストに対する可能なマッチングであるか否かを決定するための 1 つ以上の基準を適用し、

2) 入力テキストとの可能なマッチングである概念のうちのどれが入力テキストに関連し、どれが関連しないかを決定するために、1 つ以上の規則を適用し、

基準は、概念のうちの少なくともいくつかについて、概念の単語のすべてが入力テキスト内に見つけられることを必要としないが、概念の必須単語であると規定された 1 つ以上の単語が入力テキスト内に見つけられることを必要とする、プロセッサモジュールと、

e) 医用レポートの入力テキストに関連すると決定された概念を出力する出力モジュールとを含む。

【 0 0 3 7 】

30

本開示の別の態様によれば、医学概念のセットのうちのどの概念が医用レポート内に見つけられるかを自動的に特定するためのシステムが提供される。システムは、

- a) 1 つ以上の医用レポートへのアクセスを提供する第 1 のデータベースと、
- b) 医用レポートを入力テキストに分割する、第 1 のデータベースへのアクセスを有するスプリッタモジュールと、
- c) 医学概念のセットへのアクセスを提供する第 2 のデータベースと、
- d) 入力テキスト及び第 2 のデータベースへのアクセスを有するプロセッサモジュールであって、各入力テキストについて、

1) セット内の各概念について、概念が入力テキストに対する可能なマッチングであるか否かを決定するための基準を適用し、

40

2) 入力テキストとの可能なマッチングである各概念について、概念と入力テキストとの間のマッチングの程度を示すスコアを計算し、

3) スコアが計算された 2 つの概念がそれらの単語において十分に大きい重複を有しているときに、より低いスコアを有する概念は入力テキストに関連しておらず、また、概念が、その単語において十分に大きい重複を有する他の如何なる概念について計算されたスコアよりも高い計算されたスコアを有する場合に、概念は入力テキストに関連していると決定する、プロセッサモジュールと、

e) 医用レポートの入力テキストに関連すると決定された概念を出力する出力モジュールとを含む。

【 0 0 3 8 】

50

別段の定義がない限り、本明細書で使用されるすべての技術用語及び／又は科学用語は、本発明が関連する技術分野の当業者によって一般に理解されるものと同じ意味を有する。本明細書で説明されるものと類似又は同等の方法及び材料を、本発明の実施形態の実施又は実験で使うことができるが、例示的な方法及び／又は材料を以下に説明する。矛盾する場合には、定義を含む特許明細書が優先される。さらに、材料、方法、及び実施例は、例示にすぎず、必ずしも限定することを意図するものではない。

【 0 0 3 9 】

本発明の実施形態の方法及び／又はシステムの実施は、選択されたタスクを手動で、自動的に、又はそれらの組合せで実行又は完了することを含み得る。さらに、本発明の方法及び／又はシステムの実施形態の実際の器具類及び機器によれば、いくつかの選択されたタスクは、ハードウェア、ソフトウェア、若しくはファームウェア、又はオペレーティングシステムを使用するそれらの組合せによって実施できる。

10

【 0 0 4 0 】

例えば、本発明の実施形態による選択されたタスクを実行するためのハードウェアは、チップ又は回路として実装できる。ソフトウェアとして、本発明の実施形態による選択されたタスクは、任意の適切なオペレーティングシステムを使用してコンピュータによって実行される複数のソフトウェア命令として実装できる。本発明の例示的な実施形態では、本明細書で説明する方法及び／又はシステムの例示的な実施形態による１つ以上のタスクは、複数の命令を実行するためのコンピューティングプラットフォームなどのデータプロセッサによって実行される。任意選択で、データプロセッサは命令及び／又はデータを記憶するための揮発性メモリ、及び／又は命令及び／又はデータを記憶するための、磁気ハードディスク及び／又はリムーバブルメディアなどの不揮発性記憶装置を含む。任意選択で、ネットワーク接続も同様に提供される。ディスプレイや、キーボードやマウスなどのユーザ入力デバイスも、任意選択で提供される。

20

【図面の簡単な説明】

【 0 0 4 1 】

本発明の上記及び他の目的、特徴、及び利点は、添付の図面に示される本発明の実施形態の以下のより具体的な説明から明らかになるであろう。図面の要素は、必ずしも互いに相対的な縮尺であるとは限らない。ここで、図面を詳細に特に参照すると、示される詳細は、例として、かつ、本発明の実施形態の例示的な議論の目的のためであることが強調される。この点に関して、図面を参照した説明は、本発明の実施形態がどのように実施され得るかを当業者に明らかにする。

30

【 0 0 4 2 】

【図 1 A】図 1 A は、本発明の例示的な実施形態による、医用レポート内の医学概念を特定するためのシステムのブロック図を示す。

【図 1 B】図 1 B は、本発明の例示的な実施形態による、図 1 A のシステムのプロセッサモジュールによって使用される方法の高レベルフローチャートを示す。

【図 2】図 2 は、本発明の例示的な実施形態による、図 1 A のシステムのプロセッサモジュールによって使用される方法のより詳細なフローチャートを示す。

【図 3】図 3 は、本発明の異なる例示的な実施形態による、図 1 A のシステムのプロセッサモジュールによって使用される方法の高レベルフローチャートを示す。

40

【図 4】図 4 は、本発明の異なる例示的な実施形態による、図 1 A のシステムのプロセッサモジュールによって使用される方法の高レベルフローチャートを示す。

【図 5】図 5 は、本発明の異なる例示的な実施形態による、図 1 A のシステムのプロセッサモジュールによって使用される方法の高レベルフローチャートを示す。

【発明を実施するための形態】

【 0 0 4 3 】

以下は好ましい実施形態の詳細な説明であり、同じ参照番号は、いくつかの図面の各々における構造の同じ要素を特定する図面を参照する。

【 0 0 4 4 】

50

本開示は、一般に、自然言語処理の分野に関し、特に、事前に定義された医学概念のセットからどの医学概念が医用レポート内に見つけられるかを特定することに関する。より具体的には、しかし、排他的ではないが、本開示は、放射線学レポートを用いてこれを行うための方法に関する。

【 0 0 4 5 】

本明細書で説明されるシステム（医用レポートを処理するためのコンピュータシステムを含む）及び／又は方法（例えば、１つ以上のプロセッサによって実行されるコード命令）は、医用レポート内の医学概念を自動的に特定する従来技術の方法よりも見逃す概念が少なく、かつ偽陽性が少なく、患者に関する医用レポート内で言及される医学概念を自動的に特定する技術的問題に対処する。本システム及び方法は、１）語順が重要である「マッ

10

【 0 0 4 6 】

本明細書で説明されるシステム（医用レポートを処理するためのコンピュータシステムを含む）及び／又は方法（例えば、１つ以上のプロセッサによって実行されるコード命令）は、自然言語処理の技術分野内の基礎となる技術プロセスを向上させる。本明細書で説明されるシステム及び／又は方法は、１）概念を可能なマッチングと見なすために、概念のすべての単語が入力テキスト内に存在することを必要としないこと、２）特定の必須単語が存在することを必要とすること、及び、３）可能なマッチングごとにスコアを計算し、より高いスコアの可能な概念と有意に重複するときに、可能なマッチングを除外すること、のうちの１つ以上によって、医用レポート内の医学概念を特定するプロセスを向上させる。

20

【 0 0 4 7 】

本明細書で説明されるシステム（医用レポートを処理するためのコンピュータシステムを含む）及び／又は方法（例えば、１つ以上のプロセッサによって実行されるコード命令）は、医用レポート内の医学概念を特定するコード命令を実行する計算ユニットのパフォーマンスを向上させる。パフォーマンスの向上によって、成功裏に特定される医用レポートの各センテンス（文）に実際に存在する医学概念の数を増加させる一方で、偽陽性、すなわち、医用レポートのセンテンスに見つけられたと誤って特定される医学概念の数を減少させる。パフォーマンスの向上は、概念のすべての単語が入力テキスト内に存在することを必要としないこと（これは、成功裏に見つけられる概念の数を増加させる）、概念の特定の必須単語が存在することを必要とすること（これは、欠落した単語の適切でない推測による偽陽性を減少させる）、可能な概念ごとにスコアを見つけ、他のより高いスコアの可能な概念と有意に重複する可能な概念を除外すること（これは、より狭い概念のみが存在するときに、より広い概念を特定することによる偽陽性を減少させる）によって少なくとも部分的に達成される。

30

40

【 0 0 4 8 】

本明細書で説明されるシステム（医用レポートを処理するためのコンピュータシステムを含む）及び／又は方法（例えば、１つ以上のプロセッサによって実行されるコード命令）は、物理的な実生活構成要素に関連付けられている。これは、それらがデータ記憶媒体に記憶された医用レポートを処理するからであり、また、医用レポートは、医用診断機器、例えば、ＣＴ走査デバイス及びＭＲＩデバイスなどの医用イメージング機器を使用して、実際の患者に対して行われた医学的検査を記述するために、医師によって書かれているからである。本システム及び方法は、医学概念が医用レポートから人間の読者によって手動で特定しなければならない場合よりもはるかに早く、かつ、医用レポート内の医学概念

50

を自動的に特定する従来技術のコンピュータシステムよりも正確に、医用レポートから医療情報を取り出すことを可能にすることによって、患者のケアを向上させる。

【 0 0 4 9 】

本明細書で説明されるシステム（医用レポートを処理するためのコンピュータシステムを含む）及び／又は方法（例えば、１つ以上のプロセッサによって実行されるコード命令）は、医用レポート内の医学概念を特定する独自の注目すべき高度な技法を提供する。

【 0 0 5 0 】

したがって、本明細書で説明されるシステム及び／又は方法は、医用レポートの自然言語処理において生じる実際の技術的問題を克服するために、コンピュータ技術及び物理的ハードウェアに不可分に関連付けられている。

10

【 0 0 5 1 】

本発明の少なくとも１つの実施形態を詳細に説明する前に、本発明は、その用途において、以下の説明に記載され、並びに／又は図面及び／若しくは実施例に示される構成要素及び／又は方法の構成及び配置の詳細に必ずしも限定されないことを理解されたい。本発明は、他の実施形態が可能であるか、又は様々なやり方で実施若しくは実行されることが可能である。

【 0 0 5 2 】

本発明は、システム、方法、及び／又はコンピュータプログラム製品とすることができる。コンピュータプログラム製品は、プロセッサに本発明の態様を実行させるためのコンピュータ可読プログラム命令を有するコンピュータ可読記憶媒体を含み得る。

20

【 0 0 5 3 】

コンピュータ可読記憶媒体は、命令実行デバイスによって使用される命令を保持し記憶することができる有形デバイスであってもよい。コンピュータ可読記憶媒体は、例えば、電子記憶デバイス、磁気記憶デバイス、光記憶デバイス、電磁記憶デバイス、半導体記憶デバイス、又は前述の任意の適切な組合せとすることができるが、これらに限定されない。

【 0 0 5 4 】

コンピュータ可読記憶媒体のより具体的な例の非網羅的なリストには、ポータブルコンピュータディスク、ハードディスク、ランダムアクセスメモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、消去可能プログラマブル読み出し専用メモリ（ＥＰＲＯＭ又はフラッシュメモリ）、スタティックランダムアクセスメモリ（ＳＲＡＭ）、ポータブルコンパクトディスク読み出し専用メモリ（ＣＤ－ＲＯＭ）、デジタル汎用ディスク（ＤＶＤ）、メモリスティック、フロッピー（登録商標）ディスク、及び前述の任意の適切な組み合わせが含まれる。本明細書で使用されるコンピュータ可読記憶媒体は、電波又は他の自由に伝播する電磁波、導波管又は他の伝送媒体を通して伝播する電磁波（例えば、光ファイバケーブルを通過する光パルス）、又は有線を通して伝送される電気信号など、それ自体が一時的な信号であると解釈されるべきではない。

30

【 0 0 5 5 】

本明細書に説明されるコンピュータ可読プログラム命令は、コンピュータ可読記憶媒体から対応する計算／処理デバイスに、又は、ネットワーク、例えば、インターネット、ローカルエリアネットワーク、広域ネットワーク及び／又は無線ネットワークを介して、外部コンピュータ又は外部記憶デバイスにダウンロードできる。ネットワークは、銅線伝送ケーブル、光伝送ファイバ、無線伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイコンピュータ、及び／又はエッジサーバを含み得る。各計算／処理デバイス内のネットワークアダプタカード又はネットワークインターフェースは、ネットワークからコンピュータ可読プログラム命令を受信し、対応する計算／処理デバイス内のコンピュータ可読記憶媒体に記憶するために、コンピュータ可読プログラム命令を転送する。

40

【 0 0 5 6 】

本発明の動作を実行するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令セットアーキテクチャ（ＩＳＡ）命令、機械命令、機械依存命令、マイクロコード、ファームウェア命令、状態設定データ、又は、Ｓｍａｌｌ ｔａｌｋ（登録商標）、Ｃ＋＋

50

などのオブジェクト指向プログラミング言語、及び「C」プログラミング言語又は類似のプログラミング言語などの従来の手続き型プログラミング言語を含む、1つ以上のプログラミング言語の任意の組合せで書かれたソースコード又はオブジェクトコードのいずれかであり得る。

【0057】

コンピュータ可読プログラム命令は、ユーザのコンピュータ上で全体的に、ユーザのコンピュータ上で部分的に、スタンドアロンのソフトウェアパッケージとして、ユーザのコンピュータ上で部分的に、リモートコンピュータ上で部分的に、又はリモートコンピュータ又はサーバ上で全体的に実行できる。後者のシナリオでは、リモートコンピュータは、ローカルエリアネットワーク（LAN）又は広域ネットワーク（WAN）を含む任意のタイプのネットワークを介してユーザのコンピュータに接続されるか、又は（例えば、インターネットサービスプロバイダを使用してインターネットを介して）外部コンピュータに接続されてもよい。いくつかの実施形態では、例えば、プログラマブル論理回路、フィールドプログラマブルゲートアレイ（FPGA）、又はプログラマブル論理アレイ（PLA）を含む電子回路は、本発明の態様を行うために、電子回路をパーソナル化するためにコンピュータ可読プログラム命令の状態情報を利用することによって、コンピュータ可読プログラム命令を実行できる。

10

【0058】

本発明の態様は、本発明の実施形態による方法、装置（システム）、及びコンピュータプログラム製品のフローチャート図及び／又はブロック図を参照して本明細書で説明される。フローチャート図及び／又はブロック図の各ブロック、並びにフローチャート図及び／又はブロック図のブロックの組合せは、コンピュータ可読プログラム命令によって実装できることを理解されたい。

20

【0059】

これらのコンピュータ可読プログラム命令は、汎用コンピュータ、専用コンピュータ、又は他のプログラマブルデータ処理装置のプロセッサに提供されて、コンピュータ又は他のプログラマブルデータ処理装置のプロセッサを介して実行される命令がフローチャート及び／又はブロック図の1つ以上のブロックで指定された機能／動作を実装するための手段を作成するように、機械を生成できる。これらのコンピュータ可読プログラム命令は、コンピュータ、プログラマブルデータ処理装置、及び／又は他のデバイスに特定のやり方で機能するように指示できるコンピュータ可読記憶媒体に格納されてもよく、その結果、命令が格納されたコンピュータ可読記憶媒体は、フローチャート及び／又はブロック図の1つ以上のブロックで指定された機能／動作の態様を実施する命令を含む製造品を含む。

30

【0060】

コンピュータ可読プログラム命令はまた、コンピュータ、他のプログラマブルデータ処理装置、又は他のデバイス上にロードされて、コンピュータ、他のプログラマブル装置、又は他のデバイス上で一連の動作ステップを行わせてコンピュータ実施プロセスを生成し、その結果、コンピュータ、他のプログラマブル装置、又は他のデバイス上で実行される命令が、フローチャート及び／又はブロック図の1つ以上のブロックで指定された機能／動作を実施する。

40

【0061】

図中のフローチャート及びブロック図は、本発明の様々な実施形態による、システム、方法、及びコンピュータプログラム製品の可能な実装のアーキテクチャ、機能、及び動作を示す。この点に関して、フローチャート又はブロック図の各ブロックは、指定された論理機能を実装するための1つ以上の実行可能命令を含むモジュール、セグメント、又は命令の一部を表すことができる。

【0062】

いくつかの代替実装形態では、ブロックに示される機能が、図に示される順序以外で発生する場合がある。例えば、連続して示される2つのブロックは、実際には実質的に同時に実行されるか、又は、ブロックは、関与する機能に応じて、時には逆の順序で実行され

50

てもよい。また、ブロック図及び／又はフローチャート図の各ブロック、並びにブロック図及び／又はフローチャート図のブロックの組合せは、指定された機能又は動作を行うか、又は特殊目的ハードウェア及びコンピュータ命令の組合せを実行する特殊目的ハードウェアベースのシステムによって実装され得ることに留意されたい。

【0063】

本開示のいくつかの実施形態は、自然言語処理を行うコンピュータシステムを使用して、医学概念の包括的なデータベースから選択される、患者を検査した後に医師が書くレポート内で言及された医学概念を自動的に特定することを目的とする。いくつかの実施形態では、本コンピュータは、上記の目的のための既存の方法を使用するコンピュータよりも、レポート内に言及された概念をより高い割合で特定すること、及びレポートにない多くの医学概念を誤って特定することの両方において、優れたパフォーマンスを有する。医師は、レポートを読む別の医師が暗示されているものと理解するであろうと考える概念の単語を省略することが多いので、両方のこれらの手段によってパフォーマンスを同時に向上させることは難しい。これらの概念を見つけるためには、省略された単語を推測する必要があるが、単語を推測することはまた、レポートには実際にはないより多くの概念を誤って特定することにもなり得る。これらの目的を達成する際の有用な成功の尺度は、パラメータF1であり、これは、特定された存在する概念の割合と、実際に存在する特定された概念の割合との調和平均である。

10

【0064】

本開示のいくつかの実施形態は、例えば、F1によって測定される上記パフォーマンスの向上を達成する一方で、医用レポートを十分迅速に処理し、その結果、本方法が、放射線学レポートのためのより大きな自然言語処理システムの第1段階として機能できるようにすることを別の目的とする。例えば、いくつかの実施形態では、コンピュータは、4つのコア及び8つの論理プロセッサを有する3.40GHzのIntel（登録商標）Core（商標）i7-6700CPUを使用して、毎秒少なくとも100のレポートを処理することができる。

20

【0065】

本開示のいくつかの実施形態は、アルゴリズムが各概念の深層AIのような理解を有することなく、上記パフォーマンスの向上を達成することを別の目的とする。各概念のこのような深層AIのような理解を提供することは、アルゴリズムの開発を非常に労働集約的にする。

30

【0066】

これらの目的は、例示的な例としてのみ与えられたものであり、このような目的は、本発明の1つ以上の実施形態の例示であり得る。本質的に本発明によって達成される他の望ましい目的及び利点は、当業者に想起され得るか、又は明らかになり得る。本発明は、添付の特許請求の範囲によって定義される。

【0067】

コンピュータシステムの有用な特徴

本発明のいくつかの実施形態の一態様は、本発明者が偽陽性を低減することと、実際に存在する概念をあまり多く見逃さないこととにおいてしばしば有効であることを見出した3つの特徴のうちの1つ以上を用いて、コンピュータ上で命令を実行することによって、医用レポート、例えば、放射線学レポート内に、医学概念のセットのうちのどれが見つけられるかを自動的に決定するコンピュータシステムに関する。これらの特徴は、コンピュータを使用してレポート内の概念を自動的に特定するという問題に対する2つのアプローチに関して説明することができる。Elastic Searchのマッチングテストによって例示されるように、バッグ・オブ・ワーズ方法は、概念のすべての単語、又はほとんどすべての単語が、語順に関連なく、入力テキスト（例えば、医用レポート内の1つのセンテンス）内に見つけられた場合に、当該概念が入力テキストにマッチングすると見なす。この方法は、存在する概念を見つけるのにかなり効率的であるが、多くの偽陽性を生じる、つまり、実際には存在しない概念を見つける傾向がある。cTAKESアルゴリズム

40

50

によって使用され、Elastic Searchのマッチングフレーズテストによって例示される別の方法は、概念のすべての単語が、概念におけるのと同じ順序又はほぼ同じ順序で入力テキスト内に見つけられた場合に、当該概念が入力テキストにマッチングすると見なす。この方法は、偽陽性を除外する点でバッグ・オブ・ワーズ方法よりもいくぶん優れているが、存在する概念を見つけることにおいていくぶん劣る傾向がある。本発明者は、本発明のいくつかの実施形態において、以下の特徴のうちの任意の1つ若しくは2つ、又はより良くは以下の特徴のうちの3つすべてを使用する方法が、偽陽性を除外することと、実際に存在する概念を見つけることとの両方において、より良い結果を与えることができることを見出した：

1) 語順が重要であるマッチングフレーズテストであっても、すべての単語が存在することを必要としないこと。これにより、レポートを書いた医師が、レポートの人間の読者にとって自明であろうと考えた概念から単語を省略したときに、追加の概念を見つけることが可能になる。

10

2) すべての単語が存在しなくてもよい場合であっても、マッチングテスト又はマッチングフレーズテストのいずれにおいても、特定の必須単語が存在することを必要とすること。これにより、レポートを書いた医師がよって一般に省略されることのない重要な単語を推測することから生じる偽陽性の数が減少する。

3) マッチングテスト若しくはマッチングフレーズテストから、又はより良くは、両方のテストの組み合わせからスコアを計算し、スコアを使用して、ある概念に有意に重複し、より高いスコアを有する別の概念が存在する場合に、より低いスコアを有する当該概念を除外すること。これはより、狭い概念のみが存在する場合に、関連するより広い概念を除外することによって、また、単語が欠落していない、又は、欠落している単語が少ない類似の概念が存在する場合に、単語が欠落している概念を除外することによって、偽陽性が減少する。

20

【0068】

本発明者は、いくつかの場合において、これらの特徴の3つすべてを含むアルゴリズムが、より少ない偽陽性をもたらす、かつ、マッチングテスト又はマッチングフレーズテストのいずれか単独よりも、実際に存在する概念をより大きな割合で見つけることを見出した。

【0069】

30

マッチングフレーズテストにおいて、概念のすべての単語が存在することを必要としない最初の特徴は、任意選択で次のように実装される。各概念は、単語のストリングである。2つ以上の単語で構成される各概念について、1つ以上のクエリが生成される。各クエリは、任意選択で、概念におけるのと同じ順序にある、概念の単語の2単語サブストリングからなる。クエリは、概念の単語の正確なサブストリングである必要はないが、2単語概念の場合、クエリは、概念と同一である可能性がある。少なくとも最小数のクエリが、両方のそれらの単語を入力テキストに有し、互いに十分に近い場合、概念は、入力テキストとの可能なマッチングであると任意選択で見なされる。任意選択で、最小数は、その概念のクエリの総数に依存する。例えば、最小数は、クエリの総数の割合であり、例えば、クエリの35%、40%、45%、50%、55%、60%、又は65%である。任意選択で、概念を入力テキストとの可能なマッチングであると見なすためには、例えば、必須単語が存在することを必要とすることである、満たされなければならない他の基準がまたある。概念が入力テキストとの可能なマッチングであると見なされる場合、各クエリに対してクエリサブスコアが見つけられる。クエリサブスコアは、クエリの2つの単語が入力テキスト内で互いにより近くに見つかった場合、及び、当該2つの単語が、入力テキスト内で概念におけるのと同じ順序にある場合により高くなる。クエリサブスコアに基づいた概念の全体的なマッチングフレーズスコア（例えば、クエリサブスコアの加重平均）が見つけられ、概念は、それが高いスコアを有する別の概念と十分に重複している場合、入力テキストに関連しているとは見なされない。任意選択で、加重平均は、2つの単語が概念内で遠くに離れているクエリに、より低い重みを与える。

40

50

【 0 0 7 0 】

特定の必須単語が入力テキスト内に存在することを必要とすることである 2 番目の特徴は、任意選択で次のように実装される。任意選択で、概念のセット内のすべての概念の単語の中で、概念内の単語のうちで最も一般的でない概念内の単語が、入力テキスト内に見つけられなければならない必須単語である。任意選択で、病名である任意の単数形の単語が、入力テキスト内に見つけられなければならない必須単語である。任意選択で、身体部分の向き又は場所を指す任意の単語は、概念を入力テキストとの可能なマッチングと見なすためには、入力テキスト内で、それが指す身体部分の単語に隣接していなければならない。

【 0 0 7 1 】

任意選択で、どのクエリが入力テキスト内に見つけられるか、及びそれらのクエリサブスコアは何であるかを見ることによって、どの概念が入力テキストとの可能なマッチングであるかを見つけることに加えて、概念はまた、マッチングテストと呼ばれるバッグ・オブ・ワーズ方法に従って、入力テキストとの可能なマッチングであるかどうかを確認するためにテストされる。マッチングテストでは、十分な数の概念内の単語が入力テキスト内に見つけられた場合、当該概念は可能なマッチングと見なされ、単語の最小数は概念内の単語の数に依存し得る。概念がこの基準に従って入力テキストとの可能なマッチングと見なされる場合、任意選択で、入力テキスト内の語順に関係なく、任意選択で、入力テキスト内に存在する概念の単語の数にのみ依存して、概念のマッチングスコアが見つけられる。

【 0 0 7 2 】

任意選択で、マッチングスコアは、上述のマッチングフレーズスコアと組み合わせられ、例えば、マッチングスコアとマッチングフレーズスコアとの加重和が見つけられ、概念の合計スコアが見つけられる。2つの概念が十分な重複を有するときに、そのうちのどちらを保持するかを決定するためにこの合計スコアが使用される。任意選択で、概念は、十分なクエリが入力テキストにマッチングするという基準（「マッチングフレーズテスト」）と、十分な概念の単語が入力テキスト内に存在するという基準（「マッチングテスト」）との両方に従って、可能なマッチングと見なされる場合にのみ、入力テキストとの可能なマッチングであると見なされる。概念がマッチングテストとマッチングフレーズテストとの両方に合格することを必要とすることは、入力テキスト内に概念の一部の単語が見つからない場合でも、概念が入力テキストに関連している可能性を高くする。或いは、マッチングテストのみが行われても、マッチングフレーズテストのみが行われても、又は両方のテストが行われるが、入力テキストとの可能なマッチングであると見なされるために、概念は一方のテストに合格するだけでもよい。

【 0 0 7 3 】

任意選択で、マッチングテストは、1単語概念にも適用される。これには、マッチングフレーズテストは適用されず、マッチングフレーズスコアは計算されない。これは、1単語概念には2単語クエリがないからである。1単語概念は、当該1つの単語が入力テキスト内に見つけられる場合にマッチングテストに合格し、任意選択で、そのマッチングスコア及びその合計スコアは常に同じである。任意選択で、1単語概念では、マッチングフレーズスコアはゼロである。任意選択で、1単語概念のスコアが、1単語概念と重複する他の複数単語概念、例えば、当該単語も有し、マッチングテスト及びマッチングフレーズテストに合格する他の概念と比較される。任意選択で、この場合、複数単語概念のスコアは常に高くなり、1単語概念は検討から除外される。

【 0 0 7 4 】

任意選択で、概念のセットは、単語を語幹化することによって、すなわち、単語をそれらの語幹によって置き換えることによって前処理される。任意選択で、前処理において、ストップワードが概念から削除される。任意選択で、いくつかの医学用語が同じ意味を有する場合、概念内の単語は、好ましい同義語に置き換えられる。例えば、「h e p a t i c（肝臓の）」などの形容詞は、対応する名詞、この場合は「l i v e r（肝臓）」に置き換えられる。任意選択で、概念内の頭字語は展開される。任意選択で、文字 - 番号指定

10

20

30

40

50

によって言及される椎骨には、単語「vertebra（椎骨）」及び「spine（脊柱）」が追加され、文字は、それが指している脊柱の領域によって、「cervical（頸部）」、「thoracic（胸部）」、又は「lumbar（腰部）」のいずれかで置き換えられる。更に又は或いは、概念について本明細書で説明した前処理手順のいずれか又はすべてが入力テキストに適用される。

【0075】

本明細書で説明され、特許請求される方法は、抽象的なアイデアを対象としておらず、自然言語処理のためのコンピュータ技術を向上させる方法を対象としており、特許請求の範囲は、医用レポート内の医学概念を特定するすべての方法を先取りするものではなく、その結果を達成する特定の方法に限定されていることに留意されたい。例えば、請求項の少なくともいくつかは、概念から2単語クエリが作成される方法に限定され、入力テキストとのマッチングである2単語クエリの数を見ることによって、各概念に対してマッチングフレーズテストを行う。請求項の少なくともいくつかは、概念が入力テキストに対する可能なマッチングと見なされるために、概念の特定の必須単語が入力テキストに存在しなければならない方法に限定されてもよい。本方法のこれらの要素は、先行技術に対して新規かつ発明的であると考えられており、医用レポートに見られる医学概念を特定するためにコンピュータ技術を使用するのに特に適しており、人間の脳の自然言語処理能力を使用して人がその目的のために使用する方法とは非常に異なる。

【0076】

システムの説明

ここで図面を参照すると、図1Aは、本発明の例示的な実施形態による方法を実施するシステムのブロック図100を示す。図100の各ブロックは、モジュールを実施するようにプログラムされたコンピュータに格納されたソフトウェアモジュール、又は例えばデジタル形式でデータを格納するか若しくはデータへのアクセスを提供するサーバを表す。レポートデータベース102が、システムが処理する医用レポートを格納するか又は医用レポートへのアクセスを提供する。スプリッタモジュール104が、各レポートを複数の入力テキスト106に分割する。各入力テキストは、それが包含する概念を見つけるために個別に処理される。通常、レポートの各センテンスは、別個の入力テキストである。任意選択で、長いセンテンス、特に複合センテンスは、例えば、文法規則を使用して複合センテンスの各構成要素の境界を決定することを試みるパーサによって、2つ以上の入力テキストに分割される。本発明者によって行われたテストでは、長いセンテンスは、PythonのNatural Language Tool Kitによって提供されるスプリッタによって分割されているが、当技術分野で知られている任意の他のセンテンススプリッタを使用してもよい。任意選択で、センテンスは、単独のコンマ、例えば、それらの15文字以内に他のコンマがないコンマにおいて分割されるが、15文字以内に別のコンマがあるコンマでは分割されない。これは、これらのコンマは、1つの概念を橋渡しする「lung, liver, or throat cancer」などにおけるように、リストの一部になる可能性が高いためである。

【0077】

概念データベース108が、医学概念のリスト、例えば、SNOMEDデータベースにリストされる医学概念、又はSNOMEDデータベースにリストされた概念のセクションを有する。任意選択で、概念データベース108は、SNOMED又は同様のデータベースにリストされた概念のすべてを含むわけではなく、障害及び形態学的異常を記述する概念に限定され、また、最大単語数よりも長くない概念に限定される。例えば、本発明者らは、一般に、前処理の後に、6単語以下を有する概念に限定された概念のデータベースを使用した。これは、本方法が、より多くの単語を有する概念について成功率が低くなる傾向があり、このような概念は、医用レポートにおいてあまり頻繁には見られないからである。概念データベースの本実装形態では、障害及び形態学的異常を記述する概念のみが含まれている。全体で107,453個の概念があり、前処理後には、4,606個の1単語概念、24,399個の2単語概念、25,838個の3単語概念、22,815個

10

20

30

40

50

の4単語概念、18,329個の5単語概念、及び11,466個の6単語概念を含む。

【0078】

プロセッサモジュール110が、入力テキスト106と医学概念リスト108との両方へのアクセスを有し、図2で説明されるように、各入力テキスト内に見つけられる医学概念を特定するためにアルゴリズムを使用する。一般に、このプロセスはいくつかの誤差を生成する。このプロセスの結果、つまり、各入力テキスト内に見つけられた医学概念のリストは、任意選択で出力モジュール112に送られる。出力モジュール112は、ユーザが読み取れる形式で医学概念のリストを出力する。任意選択で、出力モジュールは、各入力テキスト内に見つけられた医学的概念を、出力において個別にリストするのではなく、入力テキストのセットにおいて、例えば、医用レポート全体において見つけられた医学的概念をまとめてリストするだけである。

10

【0079】

方法の概要

図1のプロセッサモジュール110の機能は、図2のフローチャート200によってより詳細に説明する。図1Bのフローチャート114に、アルゴリズムの様々な部分を高レベルで示すフローチャート200の概要が示されている。

【0080】

図1Bの116において、概念及び入力テキストの前処理が行われる。概念の前処理は、多くの入力テキストを処理する前に1回だけ行われればよく、任意選択で、概念のセットが変更された場合にのみ、再度行われる。入力テキストの前処理は、処理されている新しい入力テキストの各々に対して任意選択で行われるが、入力テキストのセットに対して、例えば、医用レポートのすべての入力テキストに対して、それらの何れかを処理する前に行われてもよい。概念及び入力テキストの両方の前処理には、例えば、語幹化、ストップワードの削除、頭字語の拡張、及び同義語による単語の置換が含まれる。概念の前処理にはまた、例えば、2つ以上の単語で構成される概念についてのクエリの生成、各概念の必須単語を見つけること、及び、特定の概念、例えば、最大数よりも多い単語を有する概念の除外が含まれ得る。

20

【0081】

118において、最初の入力テキストが検討される。

【0082】

120において、検討される入力テキストについて、概念のループオーバーが行われる。ループ120、及び図1Bに示される他のループの各々は、直列若しくは並列に行われるか、又は直列及び並列処理の組合せを使用して行われてよい。図2に示される詳細なフローチャートでは、これらのループはすべて直列に処理されるものとして示されている。ループ120は、各概念に対して、1)122において、入力テキストを必須単語についてチェックすること、2)124において、マッチングテストを行って、マッチングスコアを見つけること、及び、3)126において、マッチングフレーズテストを行って、マッチングフレーズスコアを見つけることを含む。これらの3つのテストは、直列又は並列で行われてもよく、直列に行われる場合、任意の順序で行われてよい。図2に示される詳細なフローチャートでは、これらのテストは、最初に必須単語のチェック、次にマッチングテスト、次にマッチングフレーズテストのように直列で行われるものとして示されている。テストをこの順序で行うことは、必須単語のチェックが、一般にマッチングテストよりも高速であり(マッチングテストは一般にマッチングフレーズテストよりも高速である)、概念が高速テストのうちの1つに合格しない場合には、低速テストをスキップして、コンピュータ時間を節約するという潜在的な利点を有する。マッチングフレーズテストを行い、マッチングフレーズスコアを見つけることには、その概念についてのクエリをループオーバーすること(128)が含まれる。ループ120の終わりに、130において、例えば、マッチングスコアとマッチングフレーズスコアとを組み合わせることによって、概念の合計スコアが見つかる。任意選択で、入力テキストに対する可能なマッチングと見なされるためには、概念は、3つのテスト、すなわち、122での必須単語のチェック、

30

40

50

1 2 4 でのマッチングテスト、及び 1 2 6 でのマッチングフレーズテストをすべて合格しなければならない。テストは、任意の順序で又は並列に行われてよいが、テストが直列に行われる場合は、概念がテストのうちの 1 つを合格しない場合、ループは、任意選択でその概念について終了し、次の概念の検討に進む。図 2 に示される詳細なフローチャートでは、必須単語のチェックは最初に行われ、その後マッチングテストが続くものとして示され、マッチングフレーズテストは最後に行われるものとして示されている。どの概念がすべてのテストに合格し、したがって、その入力テキストに対して可能なマッチングがあるかの記録と、可能なマッチングである各概念の合計スコアの記録とが保持される。

【 0 0 8 3 】

ループ 1 2 0 が、その入力テキストに対して可能なマッチングであるすべての概念を見つけた後、1 3 2 において、後処理が行われる。後処理は、マッチングした可能性のある概念のループオーバー 1 3 4 を含む。ループ 1 3 4 では、マッチングした可能性のある概念は、例えば、別のマッチングした可能性のある概念の単語がより多く入力テキスト内に見つけられる、及び / 又は、入力テキスト内の語順が当該別の概念内の語順により近いためにより高いスコアを有する当該別のマッチングした可能性のある概念と有意な重複を有する場合、検討から除外される。

【 0 0 8 4 】

すべてのマッチングした可能性のある概念について後処理 1 3 2 が完了した後、1 3 6 において、除外されていない残りのマッチングした概念が出力される。これらが、アルゴリズムが入力テキスト内に見つけられたと特定した概念である。任意選択で、各入力テキスト内で特定された概念を個別にリストするのではなく、いくつかの入力テキストにおいて、例えば、レポート内のすべての入力テキストにおいて特定されたすべての概念のリストを、それらの入力テキストが処理された後に出力として提供する。

【 0 0 8 5 】

1 3 8 において、処理されていない入力テキストが残っている場合、次の入力テキストが検討される。

【 0 0 8 6 】

入力テキスト及び概念の前処理

2 0 2 において、入力テキストが前処理される。前処理は、任意選択で、以下のうちの 1 つ以上を含む：

1) 冠詞、接続詞、前置詞、及び / 又は、所与の概念が入力テキスト内に見つけられるかどうかに関する情報を提供する可能性が低い他の一般的な単語を含むストップワードを削除すること。

2) 誤差を引き起こす可能性のある 2 つ以上の意味を持つ曖昧な単語を削除すること。例えば、単語「*exposure*」は、時には、医学概念である、風邪に曝されている患者 (*patient being exposed to the cold*) を指すことがあるが、この単語は、画像の暴露パラメータ (*exposure parameter*) を指すために、ほぼすべての放射線学レポートにおいても使用されている。本発明者らは、これを除外すると、アルゴリズムの F 1 スコアが良くなることを見出した。アルゴリズムの本実装形態では削除されていないが、時には誤差を引き起こす曖昧な単語の別の例は「*sinus*」であり、これは、身体部分を指すことも、疾患を指すこともできる。

3) 単語をそれらの語幹で置き換えること。これは「語幹化」と呼ばれる。

4) 同等の意味を有する単語のグループ内のすべての単語を、単一の好ましい形式で置き換えること。例えば、「*hepatic* (肝臓の)」のような形容詞は、対応する名詞「*liver* (肝臓)」によって置き換えられる。

5) 頭字語を、それが表す単語に展開する。ソフトウェアは、任意選択で、医用レポートで使用される頭字語のリスト及びそれらの拡張形式を含む。このようなリストは、例えば手動で作成できる。時々、頭字語のリストは、例えば、医用レポート内に現れ始めた新しい頭字語を追加することによって、又は曖昧であることが分かっている及び / またははや一般的に使用されていない頭字語を削除することによって、任意選択で修正される。

10

20

30

40

50

6) 椎骨の文字 - 番号指定では、単語「vertebra (椎骨)」及び「spine」を加え、Cの文字を「cervical (頸部)」、Tの文字を「thoracic (胸部)」、Lの文字を「lumbar (腰部)」で置き換えること。

【0087】

204において、概念は、任意選択で、入力テキストについて説明したのと同じ手順を使用して前処理される。任意選択で、前処理後に、最大数よりも多い単語を有する概念は、概念のリストから削除される。アルゴリズムの本実装形態では、6つよりも多い単語を有する概念は除外されている。或いは、5つよりも多い単語、7つよりも多い単語、又は8つよりも多い単語を有する概念が除外される。本発明者が行ったテストにより、概念に多くの単語があるほど、その概念が入力テキスト内に存在するかどうかを決定することが難しいことが示されており、さらに、非常に多くの単語を有する概念は範囲が狭く、あまり頻繁には発生しない可能性が高いため、実際には、それらを検討することはあまり重要ではない。206において、各概念のクエリが生成される。任意選択で、各クエリは、(前処理された)概念内にある単語のペアであって、概念内にある単語と同じ順序の単語のペアで構成され、そのような単語のペアごとにクエリが生成される。Elastic Searchの「Percolator」特徴が、クエリを生成するための便利なやり方を提供する。クエリにおける語順が重要であるのは、入力テキストの単語が同じ順序であることが必要とされるからではなく、入力テキストにおける単語の順序がばらばらである概念は一般に、入力テキストにおける単語の順序が概念における順序と同じ順序である場合よりもスコアが低くなるからであることを理解されるべきである。任意選択で、3つ以上の単語を有するクエリ、例えば、最大3つの単語を有するクエリも生成される。しかし、本発明者は、3つ以上の単語で構成されるクエリを使用することによって得られるものはほとんどないことを見出している。

10

20

【0088】

208において、各概念が調べられ、それが包含する必須単語が見つけられる。概念の必須単語は、その概念を入力テキストに対するマッチングと見なすためには、入力テキスト内に現れなければならない単語である。必須単語には、任意選択で以下のうちのいずれかが含まれる：

1) 概念のセットにおけるすべての概念の中で最も一般的でない、その概念におけるすべての単語の中の単語。

30

2) 疾患についての任意の1つの単語。

3) 例えば、概念「left lung (左肺)」における単語「left (左)」である、身体部分を記述する身体の向き又は体側を示す任意の単語。概念が入力テキストとのマッチングであると見なされるためには、向きを示す単語は、それが記述する身体部分から1単語以内で入力テキスト内に見つけられなければならない。

【0089】

さらに、他の必須単語が、例えば、それらが有用であることが分かった場合に、いくつかの概念に対して手動で、任意選択で指定される。

【0090】

手順204、206、及び208は、任意の医用レポートを処理する前に1回行われてよく、これらは、使用されているアルゴリズムのバージョンの一部となることを理解されたい。これらは、処理される新しいレポートごとに再度行われる必要はない。アルゴリズムの新しいバージョンが制定されるとき、例えば、概念から削除されるべき単語に対して異なる選択がある場合、又は、概念のリストを作るために拡張されたデータベースが使用される場合、概念のリストになんらかの変更がある場合があり、この場合、手順204、206、及び208は、少なくとも、新しいか又は変更された概念に対して、再びやり直すことができる。

40

【0091】

入力テキストのループオーバー及び概念のループオーバー

210において、最初の入力テキストが検討される。212において、最初の概念が検

50

討され、それがその入力テキストに対するマッチングであるかどうかを決定する。214において、その概念の必須単語が入力テキスト内にあるかどうか、及び、必要とされる位置がある身体部分を記述する向きの単語といった任意の必須用語の場合には、必要とされる位置にあるかどうか決定される。

【0092】

すべての必須単語が存在する場合、216において、マッチングテストに合格したかどうか決定される。マッチングテストは、概念内のすべての単語のうち、特定の最小数の単語が入力テキスト内に存在することを必要とするが、マッチングテストは「バッグ・オブ・ワーズ」テストであるので、それらの語順は問題ではない。入力テキスト内に存在する単語の数及び概念内の単語の総数の両方は、任意選択で、概念及び入力テキストが、例えば、ストップワードが削除され、語幹化される前処理の後に参照されることを理解されたい。本発明者は、マッチングテストに合格するために入力テキスト内に存在しなければならない概念内の単語の最小数が、概念内のすべての単語の2/3よりも多いが、概念内のすべての単語の3/4以下であるとき、最良の結果を見出した。しかし、概念内のすべての単語の1/2よりも多く、5/6以下の任意の値に最小値を設定することによって、合理的に良好な結果が得られた。例えば、2単語概念の場合、任意選択で、両方の単語が入力テキスト内になければならない。3単語概念の場合、任意選択で、2つの単語が入力テキスト内になければならないか、或いは、3つの単語すべてがなければならぬ。4単語概念の場合、入力テキスト内になければならない単語の最小数は、任意選択で、2、3、又は4である。5単語概念の場合、単語の最小数は、任意選択で、3又は4である。6単語概念の場合、単語の最小数は、任意選択で、3、4、又は5である。任意選択で、単語の最小数は、概念内の単語の総数の50%、60%、70%、又は80%である。

【0093】

概念がマッチングテストに合格した場合、218において、マッチングスコアが見つけれられ、記録される。マッチングスコアは、入力テキスト内にある概念内の単語の数に依存する。一般に、入力テキスト内にある概念内の単語の数が多いほど、入力テキスト内にある概念の単語の割合が大きくなっても、概念が入力テキスト内にある可能性が高くなる。任意選択で、マッチングスコアは、概念内の単語の数に関連なく、入力テキスト内にある概念の単語の数にのみ依存する。或いは、マッチングスコアを計算する際に、概念内の単語の総数も考慮される。本方法の例示的な実装形態では、マッチングスコアは、入力テキスト内にある概念の単語の数の平方根である。

【0094】

マッチングフレーズテストのクエリのループオーバー

220において、概念がマッチングフレーズテストに合格するかどうかを決定するために、及びマッチングフレーズスコアを計算するために、その概念のすべてのクエリを調べるループの開始時に、その概念の最初のクエリが検討される。各クエリは概念からの2つの単語を包含するため、1単語概念の場合、クエリはなく、したがって、マッチングフレーズテストを行うことができず、マッチングフレーズスコアを計算することができないことを理解されたい。この場合、フローチャート200には明示的に示されていないが、制御は、次の概念が試みられる238に進む。

【0095】

222において、検討されているクエリが入力テキスト内にあるかどうか決定される。つまり、必ずしもクエリの両方の単語がクエリ内にある順序とは同じ順序（つまり、概念における順序と同じ順序）でなくてもよいが、それらが入力テキスト内にあるかどうか決定される。3つ以上の単語を有するクエリが使用される場合、クエリのすべての単語が入力テキスト内にあるかどうか決定される。しかしながら、入力テキスト内のこれらの単語の位置が、クエリ内のそれらの位置と異なりすぎている場合、クエリは、入力テキスト内に見なされる。

【0096】

単語の位置がどのくらい異なるかは、任意選択で、クエリ内の単語の位置を入力テキス

10

20

30

40

50

ト内の単語の位置とマッチングさせるの必要となる位置の1ステップ変更の回数によって定義される。例えば、クエリに2つの単語があり、入力テキスト内の2つの単語がクエリにおける順序と同じ順序で互いに隣接している場合、位置の差 s はゼロである。2つの単語が、入力テキストにおいて互いに隣接しているが、クエリ内の単語の順序とは逆である場合、 $s = 2$ である。これは、クエリ内の各単語が、入力テキストにおける位置と同じ位置になるためには、1つの位置だけ動く必要があるからである。入力テキスト内の単語が、クエリ内の単語と同じ順序であるが、それらの間に n 個の他の単語がある場合、 $s = n$ になる。そして、入力テキスト内の単語が逆の順序であり、それらの間に n 個の他の単語がある場合、 $s = n + 2$ である。

【0097】

クエリが入力テキスト内にあると見なすために許可される最大 s を規定するパラメータは、「スロップ (slop)」と呼ばれる。スロップ値は、任意選択で、クエリの単語が概念内でどれだけ離れているかに依存し、クエリの単語が概念内で遠く離れているほど、スロップ値は高い。例えば、 d がクエリ内の2つの単語間の概念における距離であって、概念内で隣接している単語については $d = 0$ とすると、「スロップ」は、任意選択で、 $d + \text{定数}$ 、例えば $d + 15$ に等しい。本発明者は、「スロップ」が約 $d + 10 \sim d + 25$ であるときに、アルゴリズムのパフォーマンスが最良であることを見出した。実際には、医用レポート内のほとんどのセンテンスが15単語未満であり、これらの入力テキストについて、位置の差 s は、それが $d + 15$ と定義される場合、「スロップ」を超えることが決してなく、クエリの単語が入力テキスト内にある場合、すべてのクエリは入力テキスト内にあるものとしてカウントされる。しかしながら、医用レポート内には、いくつかの概念があることが多い異常に長いセンテンスがいくつかあり、 $d + 10 \sim d + 25$ の「スロップ」値を使用することにより、アルゴリズムが1つの概念から1つの単語を取り、別の概念から1つの単語を取って、それらを単一のクエリに不適切にマッチングさせることを阻止することができる。アルゴリズムのパフォーマンスが「スロップ」の選択された値に依存するのはこれらの比較的少ない異常に長いセンテンスのためである。

【0098】

224において、クエリが入力テキスト内にある場合、クエリサブスコアが計算されて記録される。クエリサブスコアは、任意選択で、入力スコアにおけるクエリの単語の位置が、クエリにおけるそれらの位置とどのくらい異なるかに基づいており、単語の位置が異なるほど、クエリサブスコアは低い。任意選択で、位置の差 s は、上述のように定義される。例えば、クエリサブスコアは、 $1 / (s + 1)$ に比例し、 s は上述のように定義される。例えば、クエリサブスコアは、 $\sqrt{2} / (s + 1)$ である。

【0099】

226において、クエリが入力テキスト内にあるか否かが記録される。228において、その概念についてのさらなるクエリがあるか否かが決定される。さらなるクエリがある場合、230において、次のクエリが調べられて、制御は再び222に進む。その概念において調べるクエリが残っていない場合、制御は232に進み、232において、その概念がマッチングフレーズテストに合格するか否かが決定される。概念は、その概念内の少なくとも最小数のクエリが入力テキスト内にある場合に、マッチングフレーズテストに合格する。例えば、その概念についてのクエリの少なくとも最小の割合が入力テキスト内にある。例えば、概念の少なくとも35%、少なくとも40%、少なくとも45%、少なくとも50%、少なくとも55%、少なくとも60%、又は少なくとも65%が入力テキスト内にある。

【0100】

マッチングフレーズスコア

概念がマッチングフレーズテストに合格する場合、234において、その概念のマッチングフレーズスコアが計算される。任意選択で、マッチングフレーズスコアは、入力テキスト内にあるその概念のクエリのすべてのクエリサブスコアの加重和を、その概念のクエリの数 N で除算したものである。任意選択で、その単語が概念内でより遠くに離れている

クエリには、低い重みを使用される。例えば、クエリが、概念内の i 番目の単語と j 番目の単語とからなる場合、そのクエリに使用される重みは、 $1.2 - 0.15(i - j)$ である。或いは、この式の定数 1.2 は、例えば $0.9 \sim 1.5$ の別の定数に置き換えられるか、及び/又は、定数 0.15 は、 $0.10 \sim 0.20$ の別の定数に置き換えられる。概念が 6 以下の単語を持つ場合にのみ概念のセットに含まれている場合、 $i - j$ は常に $1 \sim 5$ であり、重み $1.2 - 0.15(i - j)$ は、 $0.45 \sim 1.05$ の範囲にあることに留意されたい。任意選択で、重みの式は、概念内の単語の最大数を所与として、重みが決して負にならないように選択される。

【0101】

概念の合計スコア

概念がマッチングフレーズテストに合格する場合、236において、概念は、入力テキストに対する可能なマッチングとして記録され、概念の合計スコアが見つけれられ、記録される。概念の合計スコアは、マッチングスコアとマッチングフレーズスコアとの両方に依存する。例えば、合計スコアは、マッチングスコアとマッチングフレーズスコアとの加重和である。例えば、マッチングスコアには、マッチングフレーズスコアの重みの 1.5 倍の重みが与えられ、マッチングスコアは、入力テキスト内で見つかった概念内の単語の数の平方根として定義され、マッチングフレーズスコアは、次のように定義される：

【数1】

$$\frac{1}{N} \sum_{i \leq j} [1.2 - 0.15(j - 1)] MP(i, j)$$

ここで、 $MP(i, j)$ は、 $\sqrt{2} / (s + 1)$ として定義される、概念内の i 番目の単語と j 番目の単語とからなるクエリのクエリサブスコアであり、マッチングフレーズスコアの式中の和は、その概念のすべてのクエリにわたり、 N は、概念内のクエリの数である。マッチングスコアが係数 x だけ異なるように再定義され、マッチングフレーズスコアが係数 y だけ異なるように定義され、マッチングスコア及びマッチングフレーズスコアの相対重みが係数 y / x だけ異なるように再定義される場合、合計スコアの定義は変わらないことを理解されたい。また、合計スコアがマッチングスコアとマッチングフレーズスコアとの加重和である場合、マッチングスコア及びマッチングフレーズスコアの相対重みのみが重要であることを理解されたい。これは、スコアは、所与の入力テキストに対する異なるマッチングした概念のスコアを比較するためだけに使用され、すべてのスコアが同じ係数で変更されている場合は差が生じないからである。

【0102】

概念が入力テキストに対する可能なマッチングとして記録され、その合計スコアが記録されると、制御は238に進み、そこで、その入力テキストに対して検討されるべき概念がそれ以上残っているかどうか決定される。214において、概念のすべての必須単語が入力テキスト内で、必要な位置に見つからない場合、216において、概念がマッチングテストに合格しない場合、又は、232において、概念がマッチングフレーズテストに合格しない場合、その概念は、その入力テキストに対する可能なマッチングとしては記録されず、制御は238に直接進む。

【0103】

1 単語概念

概念に単語が1つしかない場合、それに関連付けられたクエリはない。これは、各クエリには概念からの2つの単語があるからである。この場合、上述したようにマッチングフレーズテストを行ったり、マッチングフレーズスコアを計算したりすることができない。任意選択で、この場合、概念は、マッチングテストに合格する限り、マッチングであると見なされる。これは、概念の1つの単語が入力テキスト内にあることを意味する。任意選択で、1単語概念では、合計スコアは、マッチングスコアのみ依存し、マッチングスコアは、マッチングスコアが入力テキスト内に見つけれられる概念の単語の数のみに依存する

場合、1単語概念では常に同じである。マッチングスコアが入力テキスト内に見つけられる概念の単語数の平方根として定義される場合、マッチングスコアは、1単語概念に対して常に1である。スコアが、マッチングスコアの重みが1.5として、マッチングスコアとマッチングフレーズスコアとの加重和として定義される場合、また、マッチングスコアがゼロに設定されて、同じ定義が1単語概念に使用される場合、1単語概念では、概念がその入力テキストに対するマッチングである場合、合計スコアは常に1.5になる。本発明者は、場合によっては、1単語概念をこのように扱うことが、アルゴリズムの良好なパフォーマンスをもたらし、より多くの実際に存在する概念を見つけ、偽陽性を低減することを見出した。或いは、例えば1若しくは2、中間値、又は1未満若しくは2を超える値である異なる低数値が、1つの単語が入力テキスト内に見つけられるすべての1単語概念の合計スコアに使用される。

10

【0104】

238において、概念のセットにおいて検討すべき概念がさらにあると決定された場合、240において、次の概念が検討され、制御は再び214に進む。検討すべき概念が残っていない場合、242において、その入力テキストについて、2つ以上のマッチングする可能性のある概念が記録されているかどうか決定される。2つ以上のマッチングする可能性のある概念がある場合、244において、マッチングする可能性のある概念がスコアによってソートされる。

【0105】

マッチングする可能性のある概念の後処理

20

246において、2番目にスコアの高いマッチングする可能性のある概念が検討される。248において、検討されている概念のスコアが、それよりも高いスコアを有する概念の各々のスコアと比較される。2番目にスコアの高い概念でこれが最初に行われるとき、それが比較されなければならない、より高いスコアの概念は1つしか存在しない。検討されている概念が、より高いスコアを有する概念と有意な重複を有する場合、検討されている概念は、この入力テキストに対するマッチングする可能性のある概念のリストから削除される。任意選択で、「有意な重複」は、これらの3つの条件のいずれかが真であるとして定義される：

1) 検討されている概念におけるすべての単語が、より高いスコアの概念にある。

2) より高いスコアの概念におけるすべての単語が、検討されている概念にある。

30

3) 2つの概念のうちの一方にあるが他方の概念にはない単語の数が、検討されている概念内の単語の数に比べて十分に少ない。例えば、一方の概念にあるが他方にはない単語の数を1.7で除算したものが、検討されている概念内の単語の数の0.5倍以下である最大整数よりも少ない。

【0106】

本発明者は、「有意な重複」のこの定義が、アルゴリズムの良好なパフォーマンスを与えることを見出した。或いは、異なる定義が使用される。例えば、1.7及び/又は0.5の代わりに、異なる定数が使用される。例えば、1.7の代わりに1.3~2.1の定数が使用され、0.5の代わりに0.35~0.65の定数が使用される。

【0107】

40

250において、調べるべきマッチングする可能性のある概念がさらにあるかどうか決定される。ある場合、252において、次にスコアの高い概念が検討され、制御は248に戻る。検討すべきマッチングする可能性のある概念がそれ以上ない場合、254において、アルゴリズムの結果、つまり、この入力テキストに対する残りのマッチングする概念が出力内にリストされる。これらの残りのマッチングする概念は、アルゴリズムが入力テキスト内に存在することが分かった概念である。

【0108】

任意選択で、概念のセットは、同じ意味を有すると考えられる2つ以上の概念のいくつかのグループを含み、各グループ内の概念のうちの1つは、より好ましい形態として指定され、各グループ内の他の概念は、あまり好ましくない形態として指定される。これらの

50

概念は、入力テキスト内の概念が特定されるまで、アルゴリズムのプロセス全体を通して別個の概念として扱われる。次に、特定された概念が出力内にリストされると、それほど好ましくない形態は、対応するより好ましい形態に置き換えられる。

【 0 1 0 9 】

入力テキストのループオーバーの終了

2 5 6において、処理すべき入力テキストがそれ以上あるか否かが決定される。ある場合、次の入力テキストが2 5 8において調べられ、制御は2 1 2に戻って、次の入力テキスト内に存在する概念を探す。処理すべき入力テキストがそれ以上ない場合、アルゴリズムは2 6 0において終了する。

【 0 1 1 0 】

一部の特徴のみの使用

上述したように、本発明者は、従来技術よりも結果を向上させて、コンピュータシステムが、従来技術と比較して、存在する概念をより多く見つけること、偽陽性を少なくすること、又はその両方を可能にする3つの特徴を見出している。これらの特徴は：

1) 語順が重要であるマッチングフレーズテストでも、概念のすべての単語が入力テキスト内に存在することを必要としないこと、

2) 概念の特定の必須単語が入力テキスト内に存在することを必要とすること、及び

3) マッチングする可能性のある各概念のスコアを計算し、2つのマッチングする可能性のある概念が有意な重複を有する場合、スコアが低い方の概念を除外することである。本発明者は、これらの特徴の3つすべてが使用される場合に最良の結果が得られることを見出したが、これらの特徴のうちの1つだけ又は2つを使用することも可能である。図3、図4、及び図5は、それぞれ、第1の特徴のみが使用される場合、第2の特徴のみが使用される場合、及び第3の特徴のみが使用される場合について、図1Bと同様のフローチャートを示す。3つの場合すべてにおいて、フローチャートは、図1Bについて上述したように、前処理1 1 6、最初の入力テキストの取得1 1 8、及び概念のループオーバーの実行1 2 0から始まる。そして、3つの場合すべてにおいて、フローチャートは、図1Bについて上述したように、その入力テキストに対する結果の出力1 3 6、次に、次の入力テキストを調べること1 3 8で終了する。フローチャートは、概念のループオーバーに含まれるものと、後処理を含むか否かの点において異なる。

【 0 1 1 1 】

図3において、フローチャート3 0 0は、医用レポート内の医学概念を特定する方法を示す。ここでは、各入力テキストにおいて、概念が入力テキストとの可能なマッチングであると見なされるために、マッチングフレーズテストにおいて、概念のすべての単語が入力テキスト内にある必要はない。概念のループオーバー1 2 0は、マッチングフレーズテスト1 2 8のみを示すが、任意選択で、マッチングテストも行われ、また、任意選択で、必須単語がチェックされる。フローチャート3 0 0には、スコアの計算は示されておらず、また、後処理も示されていない。マッチングフレーズテストには、任意選択で、クエリのループオーバーが含まれ、各クエリは、概念からの単語のペアからなり、すべてのクエリが入力テキスト内にある必要はない。例えば、マッチングフレーズでは、概念がマッチングフレーズテストに合格するためには、特定の数のクエリ、例えば、50%のクエリが入力テキスト内にあればよい。或いは、マッチングフレーズテストは、クエリのループオーバーを含まないが、例えば、マッチングフレーズテストに合格するために、入力テキスト内に、単語の100%未満である概念の特定の数の単語があることを必要とする。この特徴のみを用いても、コンピュータシステムは、マッチングフレーズテストに合格するために、概念内のすべての単語が入力テキスト内に存在しなければならない従来技術においてよりも、入力テキスト内に存在する概念をより多く見つけることができる。

【 0 1 1 2 】

図4において、フローチャート4 0 0は、医用レポート内の医学概念を特定する方法を示す。ここでは、概念内のすべての単語が入力テキスト内にある必要がない場合であっても、概念が入力テキストとの可能なマッチングと見なされるために、各概念について、特

10

20

30

40

50

定の必須単語が入力テキスト内になければならない。概念のループオーバー 120 は、必須単語が存在するというチェック 122 を含む。ループ 120 はまた、テストに合格するために、概念のすべての単語が入力テキスト内に存在することを必ずしも必要としないが、テストに合格するために、概念の特定の数の単語が存在することを必要とする、例えば、マッチングテスト及び/又はマッチングフレーズテストであるテスト 402 を含む。フローチャート 400 では、フローチャート 300 と同様に、スコアの計算は示されておらず、また、後処理ステップは示されていない。

【0113】

図5において、フローチャート500は、医用レポート内の医学概念を特定する方法を示す。ここでは、マッチングする可能性のある各概念に入力テキストとのマッチングの程度を示すスコアが与えられ、低いスコアの場合と有意な重複を有する高いスコアの場合がある場合、低いスコアの場合、マッチングする可能性のあるものから除外される。概念のループオーバー 120 は、マッチングテスト 124 と、マッチングフレーズテスト 126 との両方を示し、概念がテストに合格し、マッチングする可能性がある概念であることがわかった場合に、各テストについてスコアが計算される。マッチングフレーズテストは、クエリのループオーバー 128 を含む。130において、マッチングスコアとマッチングフレーズスコアとが結合されて、例えば、マッチングスコアとマッチングフレーズスコアとの加重和を取得することによって、合計スコアが得られる。或いは、概念のループオーバーにおいて、マッチングテストのみ又はマッチングフレーズテストのみが行われ、スコアは、マッチングスコアのみ、又はマッチングフレーズスコアのみになる。後処理は、132において、マッチングする可能性のある概念に対してループオーバー 134 を行うことによって行われる。マッチングする可能性のある各概念について、各概念は、それに有意に重複する、より高いスコアのマッチングする可能性のある概念が存在する場合にはマッチングする可能性のある概念からは除外される。この後処理は、狭い概念しか見つからない場合に、入力テキスト内により広い概念が特定されないようにする。例えば、左肺 (left lung) のみに言及する入力テキストでは、概念「left lung (左肺)」が特定されるが、より広い概念である「lung (肺)」は特定されない。

【0114】

コンピュータプログラム製品は、1つ以上の記憶媒体、例えば、磁気ディスク（フロッピー（登録商標）ディスクなど）又は磁気テープなどの磁気記憶媒体、光ディスク、光テープ、又は機械可読バーコードなどの光記憶媒体、ランダムアクセスメモリ（RAM）又は読取り専用メモリ（ROM）などのソリッドステート電子記憶デバイス、又は本発明による方法を実施するために1つ以上のコンピュータを制御するための命令を有するコンピュータプログラムを記憶するために使用される任意の他の物理デバイス若しくは媒体を含み得る。

【0115】

本明細書で使用される場合、用語「約」は、±10%を指す。

【0116】

用語「備える」、「含む」、「有する」、及びそれらの同根語は、「～を含むが、これに限定されない」を意味する。

【0117】

用語「からなる」は、「～を含み、これに限定される」を意味する。

【0118】

用語「～から本質的になる」は、組成物、方法、又は構造が、追加の成分、ステップ、及び/又は部品を含んでもよいが、追加の成分、ステップ、及び/又は部品が特許請求される組成物、方法、又は構造の基本的かつ新規な特徴を実質的に変更しない場合に限ることを意味する。

【0119】

本明細書で使用される場合、単数形は、コンテキストが明確に別段の指示をしない限り、複数の参照を含む。例えば、用語「化合物」又は「少なくとも1つの化合物」という用

10

20

30

40

50

語は、それらの混合物を含む、複数の化合物を含み得る。

【0120】

本願全体を通して、本発明の様々な実施形態は、範囲形式で提示され得る。範囲形式での説明は、単に便宜及び簡潔さのために過ぎず、本発明の範囲に対する柔軟性のない限定として解釈されるべきではないことを理解されたい。したがって、範囲の説明は、その範囲内のすべての可能な部分範囲並びに個々の数値を具体的に開示したものとみなされるべきである。例えば、1～6などの範囲の説明は、1～3、1～4、1～5、2～4、2～6、3～6などの部分範囲、並びに例えば、1、2、3、4、5、及び6であるその範囲内の個々の数値を具体的に開示したものと見なされるべきである。これは、範囲の幅に関連なく適用される。

10

【0121】

数値範囲が本明細書で示されるときはいつでも、それは、示された範囲内の任意の引用された数字（分数又は整数）を含むことを意味する。第1の示された数と第2の示された数と「の間の範囲」という語句と、第1の示された数「から」第2の示された数「まで及び」という語句は、本明細書では交換可能に使用され、第1及び第2の示された数、並びにそれらの間のすべての分数及び整数を含むことを意味する。

【0122】

明確にするために、別個の実施形態の文脈で説明される本発明の特定の特徴は、単一の実施形態において組み合わせて提供されてもよいことを理解されたい。逆に、簡潔にするために、単一の実施形態の文脈で説明される本発明の様々な特徴は、別々に、任意の適切な部分組み合わせで、又は本発明の任意の他の説明された実施形態で適切なものとして提供されてもよい。様々な実施形態の文脈で説明される特定の特徴は、実施形態がこれらの要素なしで動作不能でない限り、これらの実施形態の本質的な特徴と見なされるべきではない。

20

【0123】

上述したように、また以下の特許請求の範囲のセクションに説明されているように、本発明の様々な実施形態及び態様は、以下の実施例において実験的支持が見出される。

【0124】

実施例

ここで、以下の実施例を参照するが、これらは上記の説明と共に、本発明のいくつかの実施形態を非限定的な様式で例示する。

30

【0125】

アルゴリズムを試験し、そのパフォーマンスを、医学レポート内の概念を見つける従来技術のアルゴリズムと比較するために、107個の放射線学レポートからテストデータを準備した。各レポートの各センテンスは、手動で調べられて、SNOMED CT医学辞書から取られ、6語以下の概念（ストップワードを削除するために、上述のように前処理した後）に限定され、障害及び医学的異常を記述する概念に限定された、概念のセット内のどの概念が決定された。このようにして、どの概念が実際に各センテンス内に見つかったが決定された。107個のレポートのセンテンスには、合計860個の概念が見つかった。

40

【0126】

アルゴリズムを用いて、セット内のどの概念がレポートの各センテンス内に発生したかを特定した。特定された真の概念（真陽性）の総数、そのセンテンス内に実際には見つからなかった特定された概念（偽陽性）の総数、アルゴリズムが特定することができなかった各センテンス内に実際に存在した概念（偽陰性）の総数、及び、各センテンス内に実際に存在しておらず、アルゴリズムが正しく特定しなかった概念（真陰性）の総数が数えられた。次に、2つの量を計算した。適合率は、真陽性と全陽性との比である。これは、実際に存在したアルゴリズムによって特定されたすべての概念の分数である。再現率は、真陽性と真陽性＋偽陰性との比である。これは、アルゴリズムが特定した実際に存在するすべての概念の分数である。量F1は、適合率及び再現率の調和平均であり、 $2 / (1 / \text{適合}$

50

率 + 1 / 再現率) として定義され得る。

【 0 1 2 7 】

同じテストを、5つの他の方法(そのうちのいくつかは、先行技術において既知である)でも行った。これらのすべてのテストにおいて、より低いスコア of の概念は、有意な重複がある場合には、より高いスコア of の概念が選択されることで、除外され、同じ概念のセットが使用された(前処理後の6語以下のS N O M E D の概念)。

1) 概念内の単語の少なくとも75%が入力テキスト内に存在することを必要とする、マッチングテスト及びマッチングスコアのみ(バッグ・オブ・ワーズモデル)。

2) ゼロ「スロップ」を有するマッチングフレーズを使用し、概念内のすべての単語が、概念内と同じ順序で、かつ間に他の単語がない状態で存在することを必要とするc T A K E S アルゴリズムに類似したアルゴリズム。

3) 概念内のすべての単語が、必ずしも順序通りでないが存在することを必要とするが、概念内の単語が入力テキスト内であまりにも離れているか、又はあまりにも順序がばらばらである場合には、概念をマッチングとしてカウントしない、いくつかの「スロップ」を有するマッチングフレーズ。「スロップ」の値は、10~15であり、結果はその範囲内のスロップの正確な値にあまり依存しなかった。

4) マッチングスコアとマッチングフレーズスコア(いくつかの「スロップ」を有する)との両方をスコアに含むハイブリッドアルゴリズムであって、マッチングフレーズテストは、概念内のすべての単語が入力テキスト内に存在することを必要とする。

5) マッチングスコアとマッチングフレーズスコアとの両方を含むハイブリッドアルゴリズムであって、マッチングフレーズスコア及びマッチングフレーズテストは、概念内のすべての単語とのマッチングを必要とするのではなく、概念内のすべての2単語クエリのセットに基づいているが、必須単語は使用しない。

【 0 1 2 8 】

表1にテスト結果を示す。4つのコア及び8つの論理プロセッサを有する3.40GHzのIntel(登録商標)Core(商標)i7-6700CPUを使用し、また、フルアルゴリズムを使用するコンピュータシステムは、毎秒100を超える放射線学レポートを処理することができた。

【表1】

表1. 異なるアルゴリズムのテスト結果

	マッチのみ	マッチフレーズのみ、ゼロスロップ	マッチフレーズ、いくつかのスロップ	ハイブリッド、マッチフレーズ、ゼロスロップ	ハイブリッド、マッチフレーズ、2単語クエリ	フルアルゴリズム
適合率	61.6%	75.7%	78.3%	86.5%	75.1%	88.3%
再現率	80.8%	70.2%	78.7%	85.9%	86.7%	89.7%
F1	69.9%	72.9%	78.5%	86.2%	80.4%	89.0%

【 0 1 2 9 】

例えば、「マッチング」のみを使用する従来技術の方法のいくつかは、適合率(存在しない概念を拒否する)がかなり良好であるが、再現率(存在する概念のすべてを見つける)はあまり良くない。他の従来技術の方法、例えば、「マッチングフレーズ」は使用するが、「マッチング」は使用しないc T A K E S は、再現率はかなり良好であるが、適合率はあまり良くない。「マッチングフレーズ」にいくつかの「スロップ」を可能にすることは、アルゴリズムが「マッチング」のみを使用するのとはほぼ同じ数の存在する概念を見つ

けることを可能にし、一方、存在しない概念を拒絶することにおいてはるかに良い。マッチングスコアとマッチングフレーズスコアとの両方を組み合わせるハイブリッドアルゴリズムを使用し、また、マッチングフレーズスコア及びテストに2単語クエリを使用することにより、概念内のすべての単語が存在する必要がなく、再現率がさらに向上される一方で、マッチングフレーズテストにおいてゼロスロップのハイブリッドアルゴリズムを使用することにより、概念内のすべての単語が同じ順序で存在することを必要とし、適合率がさらに向上される。マッチングフレーズスコア及びテストのクエリを使用し、必須単語を使用するフルアルゴリズムは、共に89%に近い最高の適合率及び最高の再現率を有する。フルアルゴリズムについて、適合率及び再現率がほぼ同じであるという事実は、アルゴリズムが存在する概念を見逃したときに、実際には存在しない別の概念をほとんど常に見つけ出し、その逆も同様であるという事実を反映する。

10

【0130】

フルアルゴリズムを使用するこのテストデータについて、誤差（偽陽性及び偽陰性）の約27%は、アルゴリズムが、手動で見つけられた概念に密接に関連する概念を見つけることによるものであったが、いずれの概念も許容可能であり、アルゴリズムによって見つけられた概念は、合理的に誤差とは全く見なされ得なかったことに留意されたい。このことの典型的な例は、2つの概念「fracture of radius」及び「fracture of ulna」を見つける代わりに「fracture of radius and/or ulna」を見つけることである。これらの誤差が誤差としてカウントされなかった場合、F1はさらに高くなり、約92%となる。

20

【0131】

アルゴリズムの微調整によって、パフォーマンスのわずかな向上が依然としてなされるかもしれないが、本発明者は、各概念の深層AIのような理解を利用することなく、アルゴリズムのパフォーマンスを実質的に向上させることはできないかもしれないと考えている。

【0132】

本発明は詳細に説明され、適切な又は現在好ましい実施形態を特に参照して説明されたが、本発明の精神及び範囲内で変形及び修正を行うことができることが理解されるであろう。したがって、本開示の実施形態は、すべての点で例示的であり、限定的ではないと見なされる。本発明の範囲は、添付の特許請求の範囲によって示され、その均等物の意味及び範囲内に入るすべての変更はその中に包含されることが意図される。

30

【0133】

本発明をその特定の実施形態に関連して説明したが、多くの代替、修正、及び変形が当業者には明らかであろうことは明らかである。したがって、本発明は、添付の特許請求の範囲の精神及び広い範囲内にある、そのような代替、修正、及び変形のすべてを包含することが意図される。

【0134】

本明細書において言及されるすべての刊行物、特許及び特許出願は、あたかも各個々の刊行物、特許又は特許出願が参照により本明細書に組み込まれるように具体的かつ個別に示されたかのように、その全体が本明細書に参照により組み込まれる。加えて、本願における任意の参照の引用又は特定は、そのような参照が本発明の先行技術として利用可能であることを容認するものとして解釈されるべきではない。セクションの見出しが使用される限り、それらは必ずしも限定するものと解釈されるべきではない。さらに、本願の任意の優先権文書は、その全体が参照により本明細書に組み込まれる。

40

【図面】

【 図 1 A 】

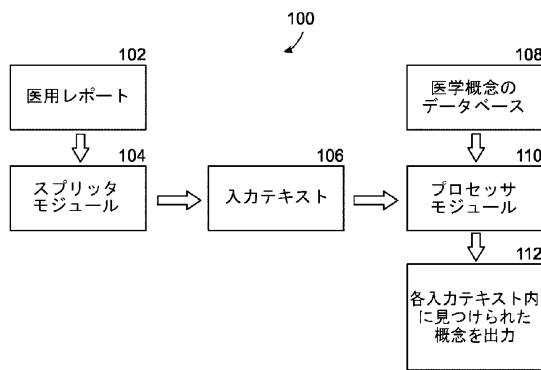


图 1 A

【 図 1 B 】

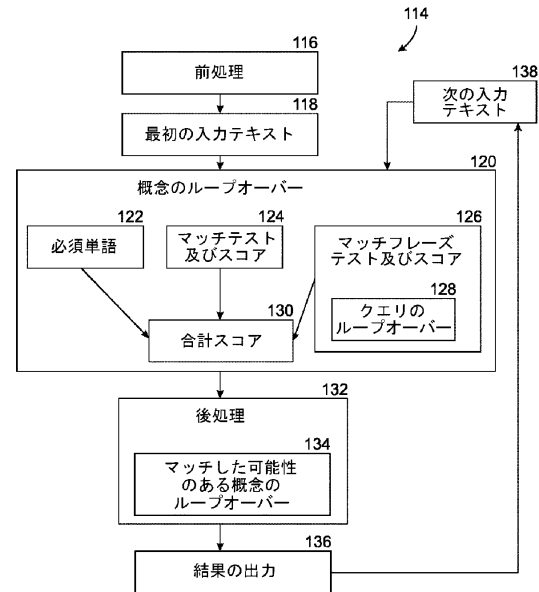
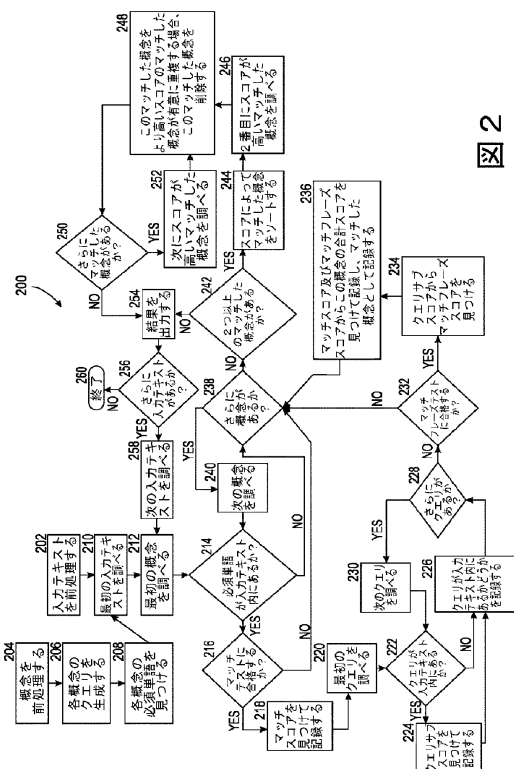


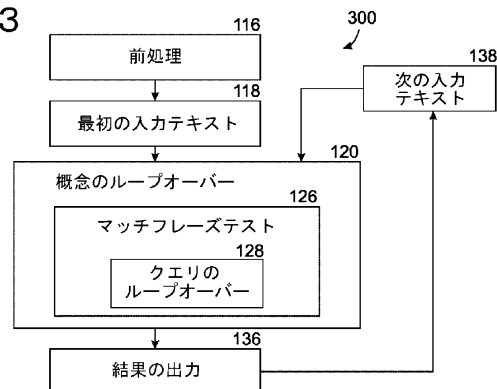
図 1 B

【圖 2】



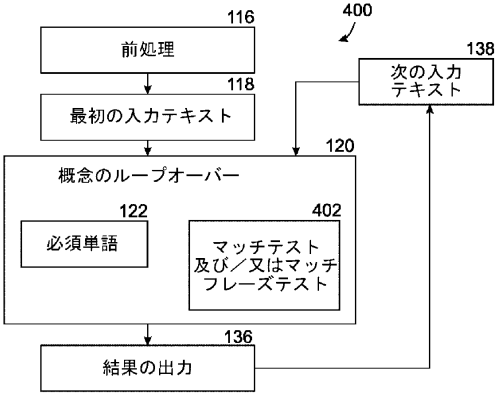
【 図 3 】

图 3



【 図 4 】

図 4



【 図 5 】

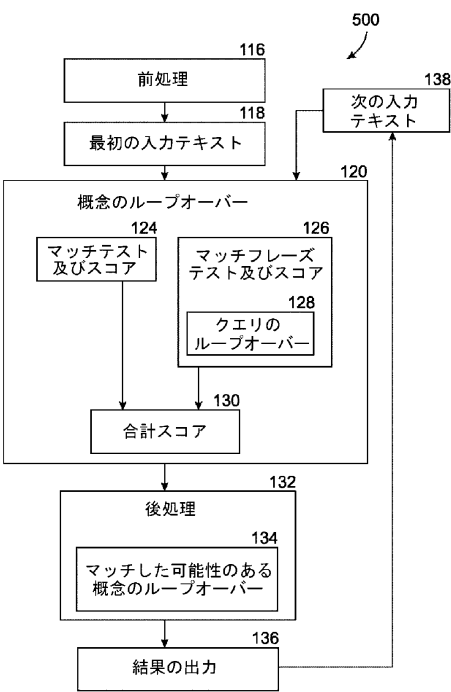


図 5

フロントページの続き

(56)参考文献 米国特許出願公開第 2 0 1 8 / 0 2 9 3 2 2 7 (U S , A 1)
 米国特許出願公開第 2 0 1 6 / 0 2 0 3 2 8 1 (U S , A 1)
 特開 2 0 1 2 - 0 3 8 0 6 4 (J P , A)
 Jian-Yun Nie , Flexible Concept Matching for Medical Information Retrieval , 2015 IEEE International Conference on Systems , 米国 , IEEE , 2016年01月12日 , P1901-P1906 , <https://ieeexplore.ieee.org/document/7379464>
 Eugene Tseythin , NOBLE-Flexible concept recognition for large-scale biomedical natural language processing , BMC Bioinformatics , 米国 , Biomed Central Ltd , 2016年01月14日 , P1-P15 , <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0871-y>
(58)調査した分野 (Int.Cl. , D B 名)
 G 0 6 F 1 6 / 0 0 - 1 6 / 9 5 8
 G 0 6 F 4 0 / 0 0 - 4 0 / 5 8
 G 1 6 H 1 5 / 0 0