

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2024-524060
(P2024-524060A)

(43)公表日 令和6年7月5日(2024.7.5)

(51)国際特許分類
G 0 6 F 16/383(2019.01)

F I
G 0 6 F 16/383

テーマコード (参考)
5 B 1 7 5

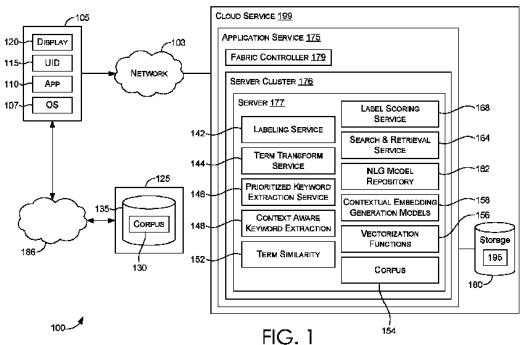
審査請求 未請求 予備審査請求 未請求 (全52頁)

(21)出願番号	特願2023-576164(P2023-576164)	(71)出願人	314015767
(86)(22)出願日	令和4年5月23日(2022.5.23)		マイクロソフト テクノロジー ライセンシング, エルエルシー
(85)翻訳文提出日	令和6年2月7日(2024.2.7)		アメリカ合衆国 ワシントン州 9 8 0 5
(86)国際出願番号	PCT/US2022/030464		2 レッドモンド ワン マイクロソフト
(87)国際公開番号	WO2023/278070		ウェイ
(87)国際公開日	令和5年1月5日(2023.1.5)		
(31)優先権主張番号	202141029147	(74)代理人	100079108
(32)優先日	令和3年6月29日(2021.6.29)		弁理士 稲葉 良幸
(33)優先権主張国・地域又は機関	インド(IN)	(74)代理人	100109346
			弁理士 大貫 敏史
(31)優先権主張番号	17/711,506	(74)代理人	100117189
(32)優先日	令和4年4月1日(2022.4.1)		弁理士 江口 昭彦
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	100134120
			弁理士 内藤 和彦
(81)指定国・地域	AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA	(74)代理人	100108213
	最終頁に続く		最終頁に続く

(54)【発明の名称】 テキストデータの自動ラベル付け

(57)【要約】

本明細書において説明される技術は、要求されたクラスに関しトレーニングされていない可能性のある生成モデルを使用することにより候補テキストが要求クラス内に存在するかどうか、を判断する。本技術は、ラベルクラスの多数の手動ユーザ入力例を必要することなく、教師無しモードで主としてトレーニングされたモデルを使用し得る。本技術は、候補テキスト及びラベルからラベルテキストの意味論的に豊かなポジティブ例を生成し得る。同様に、本技術は、候補テキスト及びラベルからラベルテキストの意味論的に豊かなネガティブ例を生成し得る。ラベル付けサービスは、生成結果を生成するために、ラベルが候補テキストに正しく適用されるという可能性を推定する生成モデルを利用する。別の態様では、本技術は、候補テキストに似ている意味論的に豊かな例を取得する方法に向けられる。



【特許請求の範囲】**【請求項 1】**

クラスラベルとテキストとの対応を判断する方法であって、
候補テキストを受信することと、
ラベル記述を受信することと、
クエリーを生成するために前記ラベル記述を使用することと、
前記クエリーを検索エンジンへ伝達することと、
前記クエリーに応じたテキストストリングを前記検索エンジンから受信することと、
前記テキストストリング及び前記候補テキストを生成モデルへ入力することと、
生成されたテキストを前記生成モデルから受信することであって、前記生成されたテキストは、複数のトークン及び関連確率を含む、ことと、
前記生成されたテキストに基づきラベル確率推定を判断することと、
前記ラベル確率推定に基づいて、前記候補テキストが前記ラベル記述に対応するかどうかのインジケーションを出力することと
を含む方法。

【請求項 2】

前記ラベル確率推定はラベルに対応する前記生成されたテキストのトークン確率から判断される、請求項 1 に記載の方法。

【請求項 3】

前記ラベルはポジティブラベル又はアンチラベルである、請求項 2 に記載の方法。

【請求項 4】

前記ラベル確率推定は前記ラベル記述のキーワード又はアンチラベルのキーワードに対応する前記生成されたテキストのトークン確率から判断される、請求項 1 に記載の方法。

【請求項 5】

前記検索エンジンの検索エンジン技術は、規則ベース検索、意味論的近さに基づく意味論的検索、又はトランスフォーマモデルを使用する文脈検索から成るグループから選択される、請求項 1 に記載の方法。

【請求項 6】

前記生成されたテキストに基づき前記ラベル確率推定を判断することは、前記生成されたテキストに基づいて第 1 のラベルスコアへ適用される第 1 の重み付けと、前記候補テキストが第 2 の生成モデルへ入力された場合に前記第 2 の生成モデルから受信された第 2 の生成されたテキストに基づいて第 2 のラベルスコアへ適用される第 2 の重み付けとを使用することを含む、請求項 1 に記載の方法。

【請求項 7】

前記第 1 の重み付け及び前記第 2 の重み付けは、前記ラベル記述に似ている異なるラベル記述の一組の格納された重み付けを発見することにより判断される、請求項 6 に記載の方法。

【請求項 8】

コンピューティングデバイスにより実行されると前記コンピューティングデバイスにクラスラベルとテキストとの対応を判断する方法を行わせる命令を含むコンピュータ可読媒体であって、前記方法は、

候補テキストを受信することと、

ラベル記述を受信することと、

前記候補テキストを生成モデルへの入力として使用して、前記生成モデルから候補結果を生成することと、

ポジティブ例テキストを前記生成モデルへの入力として使用して、前記生成モデルからポジティブ例結果を生成することであって、前記ポジティブ例テキストは、前記ラベル記述を具現化する、ことと、

ネガティブ例テキストを生成モデルへの入力として使用して、前記生成モデルからネガティブ例結果を生成することであって、前記ネガティブ例テキストは、前記ラベル記述の

反対の概念を具現化する、ことと、

前記ポジティブ例結果及びネガティブ例結果を含むコーパスに対するクエリーとして前記候補結果を検索エンジンに提出することからの応答に基づいて、前記ポジティブ例結果の第1のランクスコアを判断することと、

前記ネガティブ例結果及び前記ネガティブ例結果を含むコーパスに対する前記クエリーとして前記候補結果を前記検索エンジンに提出することからの応答に基づいて、前記ネガティブ例結果の第2のランクスコアを判断することと、

前記ポジティブ例結果の前記第1のランクスコアと前記ネガティブ例結果の前記第2のランクスコアとを比較することによりラベル確率推定を判断することと、

前記ラベル確率推定に基づいて、前記候補テキストが前記ラベル記述に対応するかどうかのインジケーションを出力することとを含む、媒体。

10

【請求項9】

前記検索エンジンは意味論的検索エンジンである、請求項8に記載の媒体。

【請求項10】

前記生成モデルはzero-shotモードで実行されるGPT3である、請求項8に記載の媒体。

【請求項11】

前記インジケーションは、前記ラベル確率推定と異なる方法により計算された第2のラベル確率推定との重み付け組み合わせに基づく、請求項8に記載の媒体。

20

【請求項12】

前記候補テキストは文書のコーパスである請求項11に記載の媒体。

【発明の詳細な説明】

【背景技術】

【0001】

背景

現在の検索技術は単純性及び使い易さのために大きな進歩をとげた。しかし、これらの変化が有益であったのと同じぐらいこれらは2つの重要なやり方に概して制限される。第1に、これらの方法は、キーワードが使用されそして高い関連性を有する文書を発見することにより主題に関連する一組の結果を受信するために正しい具体的名詞又はキーワードを提供するユーザに依存する。クエリーを書く人のボキャブラリ及び経験が制限されれば、これは、数時間の企てを要し、そしてユーザが始めから使用しているべきだった名詞を見出す前に失敗し得る。検索したい主題のインデックス内に出現するキーワードをユーザが知らなければ、ユーザが、試行錯誤の痛みを伴うプロセス無しに主題に関連する結果を取得する可能性は低い。

30

【0002】

第2に、一般的に利用可能である検索技術は通常、関連文書の表面を引っ掻くだけである。より関連する多くの文書が存在し得る。しかし、これらの文書は異なる術語、異なるボキャブラリ及び表現を使用する。従って、これらの文書はユーザのクエリーに関連した高スコアを得ることがない。

40

【0003】

これらの2つの制限は部分的には、言語的データをラベル付けする過去の企ての失敗の結果である。存在した方法は、ユーザ定義に従わなく、そして、検索を行ったユーザに必ずしも知られていなかった又は意味がなかった名詞の分類法を使用した。手動ラベル付けシステムは、ユーザにかかる大量の処理能力及び費用で開発されており、従って、検索プロセスに又は検索プロセスのインデックスに通常は利用可能ではない。

【発明の概要】

【0004】

概要

本概要は、以下の「発明を実施するための形態」において更に説明される概念のうちの

50

選択されたものを単純化形式で導入するために提供される。この「発明の概要」は請求される主題のキーフィーチャ又は必須フィーチャを特定することを意図していてもいないし、請求される主題の範囲を判断する際の助けとして使用されることも意図されていない。

【0005】

本明細書において説明される技術は、候補テキストが要求クラス内に存在するかどうかを判断する。本技術は、要求クラスに関しトレーニングされたいかなる以前のトレーニングデータ又はモデルも無しにこの分類を行い得る。実際、ユーザはクラスを既存クラスから選択するのではなく自然言語入力として規定し得る。要求されるクラスは階層に従う必要もないし予め規定される必要もない。本技術は、要求されたクラスが名詞ではなくむしろ多様性などの概念である時ですら効果的である。要求されるクラスはラベルとして本明細書において説明され得る。他のテキストラベル付けシステムは、いくつかの所定最小数の手動ユーザ入力例を必要としてきたし、加えて、ラベル分類器の教師有りトレーニングを行うために多くのコンピュータ処理を必要としてきた。本技術は、例えばラベルクラスの多数の手動ユーザ入力例を必要することなく主として教師無しモードにおいてトレーニングされたモデルを利用する一方で良好な性能を提供することにより最新技術を改善する。ラベル付けサービスの入力及びコンピュータトレーニング要件は通常よりはるかに低資源集約的であるので、コンピュータ化システムは、結果を描画するためにコンピュータ処理を余り必要としない技術的改善を提供する。

10

【0006】

本明細書において説明される技術は、候補テキスト及びラベルを受信することによりこの改善された効率を提供するので、候補テキスト及びラベルからラベルテキストの意味論的に豊かなポジティブ例を生成し得る。同様に、ラベル付けサービスは、候補テキスト及びラベルからラベルテキストの意味論的に豊かなネガティブ例を生成し得る。ラベル付けサービスは、生成結果を生成するために、ラベルが候補テキストに正しく当てはまるという可能性を推定する生成モデルを利用する。分類の成功率は、生成モデルから第2の生成結果を取得しそして第2の生成結果を使用することによりラベル確率を推定することにより、この改善された効率を維持する一方で改善され得る。

20

【0007】

別の態様では、本技術は、候補テキストに似ている意味論的に豊かな例を取得する方法に向けられる。この問題に対する他の解決策は、入力データの意味論的に劣悪な表現を提供してきた、又はそうでなければ、トレーニングを提供するために莫大な量の手動データに依存してきた。本技術は、例えばラベルの文脈内の候補テキストの豊さを反映する一組のキーワードを取得することにより、この改善を提供する。一組のキーワードは検索サービスへ提示され、そして、良好な関連ランクを有する検索結果からのテキストスニペットが、抽出されたスニペットのラベルクラス信頼度が高い場合に例を提供するために取得される。

30

【0008】

別の態様では、本技術は、ラベルの文脈内の候補テキストから意味論的に豊かな一組のキーワードを提供する方法に向けられる。他の解決策は表現において意味論的に貧弱であったので、一定数の関連結果を取得するために受信されなければならない検索エンジンからの戻りの数は大きかった。本技術は、例えば意味論的に豊かな一組のキーワードを生成する一方で良好な性能を提供することにより、そして従ってトレーニングのために必要とされるデータの量を低減することにより最先端技術を改善する。一組の候補テキスト優先キーワードが候補テキストから取得される。一組のラベル優先キーワードがラベルから取得される。優先キーワードには、トランスフォーマベースモデルを使用することにより埋め込みベクトルを割り当てられる。文脈意識キーワードは、一組の文脈意識キーワードを取得するために、埋め込みベクトルに基づく優先キーワードの類似度により判断される。この文脈意識組のキーワードは、検索エンジンからラベルの文脈内の候補テキストに意味論的に近い情報を取得することを可能にし、従って、一定数の関連結果を戻すために必要とされる検索処理の量が低減される。

40

50

【0009】

本明細書において説明される技術は、同様な参照符号が同様な要素を示す添付図面において一例として示されるのであって制限として示されていない。

【図面の簡単な説明】

【0010】

図面の簡単な説明

【図1】本開示の実装形態に好適な例示的ラベル付けシステム動作環境のブロック図である。

【図2】本開示の態様を実装するために好適なラベル付けアプリケーションの例示的表示である。

【図3】本明細書において説明される技術の一態様による「ラベルが候補テキストへ正しく割り当てられるだろう」確率の推定に基づき結果を提供する方法のフローチャートを示す。

【図4】本明細書において説明される技術の一態様による候補テキストに基づく候補入力の結果を提供する方法のフローチャートである。

【図5】本明細書において説明される技術の一態様による候補テキストに基づく候補入力の結果を提供する方法の追加実施形態のフローチャートである。

【図6】本明細書において説明される技術の一態様による一組のクラス例の拡張に基づく結果を提供する方法のフローチャートである。

【図7】本明細書において説明される技術の一態様によるラベルの文脈内の優先組のキーワードに基づく一組の文脈意識キーワードを生成する方法のフローチャートである。

【図8】本明細書において説明される技術の態様を実装する際の使用に好適な例示的コンピュータ環境のブロック図である。

【図9】本明細書において説明される技術の一態様による優先組のキーワードを用意する方法のフローチャートである。

【図10】本明細書において説明される技術の一態様による類似度を計算する方法のフローチャートである。

【図11】本明細書において説明される技術の一態様による優先ラベルキーワード構造に関係する優先テキストキーワード構造の代表的な表示である。

【図12】本明細書において説明される技術の一態様による候補テキストに基づく候補入力の結果を提供する方法の追加実施形態のフローチャートである。

【図13】本明細書において説明される技術の一態様によるクラスラベルとテキストとの対応を判断する方法を示すフローチャートである。

【図14】本明細書において説明される技術の一態様によるクラスラベルとテキストとの対応を判断する方法を示すフローチャートである。

【図15】本明細書において説明される技術の一態様による分類器のトレーニングデータを拡張する方法を示すフローチャートである。

【図16】本明細書において説明される技術の一態様による候補テキストに基づく候補入力の結果を提供する方法のフローチャートである。

【発明を実施するための形態】

【0011】

詳細な説明

本明細書において説明される様々な技術は法的要件を満たすために十分な特異性と共に記載される。しかし、本明細書自体は本特許の範囲を制限するようには意図されていない。むしろ、本発明者らは「請求される主題もまた、他の現在又は将来技術に関連する様々な工程又は本明細書において説明されるものと同様な工程の組み合わせを含むために他のやり方で具現化される可能性がある」ということを企図した。更に、用語「工程」及び／又は「ブロック」は採用される方法の様々な要素を暗示するために本明細書において使用され得るが、これらの用語は、個々の工程の順番が明示的に記載されない限り又は個々の工程の順番が明示的に記載される場合を除いて本明細書に開示された様々な工程の中の又

10

20

30

40

50

は工程間のいかなる特定順番も暗示するものと解釈されるべきでない。

【0012】

本明細書において説明される技術は候補テキストが要求クラス内に存在するかどうかを判断する。本技術は、要求クラスに関しトレーニングされたいかなる以前のトレーニングデータ又はモデルも無しにこの分類を行い得る。実際、ユーザは、クラスを既存クラスから選択するのではなくむしろクラスを自然言語入力として規定し得る。要求されるクラスは階層に従う必要がないし事前定義される必要もない。本技術は、要求されたクラスが名詞ではなくむしろ多様性などの概念である場合ですら効果的である。要求されるクラスはラベルとして本明細書では説明され得る。

【0013】

ラベル分類システムは、候補テキストがユーザ定義ラベルに適合する可能性が高い又は適合しないということを示すフィードバックをユーザへ提供し得る。例えば、ビジネス文章作成支援アプリケーションは「顧客に心地良いビジネスライクコミュニケーション」などのユーザ定義クラスを受信する可能性がある。候補テキストはワード処理文書であり得る。この例では、文書の各文章はユーザ定義クラスに属する又は属さないとして評価され得る。出力として、ワード処理アプリケーションは文章が「顧客に心地良いビジネスライクコミュニケーション」でない場合に当該文章をハイライトし得る。

【0014】

他のテキストラベル付けシステムは、いくつかの所定最小数の手動ユーザ入力例を必要としてきたし、加えて、ラベル分類器の教師有りトレーニングを行うために多くのコンピュータ処理を必要としてきた。本技術は、例えばラベルクラスの多数の手動ユーザ入力例を必要することなく主として教師無しモードにおいてトレーニングされたモデルを利用する一方で良好な性能を提供することにより最新技術を改善する。ラベル付けサービスの要件をトレーニングする入力及びコンピュータは通常よりはるかに低資源集約的であるので、コンピュータ化システムは、結果を描画するためにコンピュータ処理を余り必要としない技術的改善を提供する。

【0015】

本明細書において説明される技術は、候補テキスト及びラベルを受信することによりこの改善された効率を提供するので、候補テキスト及びラベルからラベルテキストの意味論的に豊かなポジティブ例を生成し得る。同様に、ラベル付けサービスは、候補テキスト及びラベルからラベルテキストの意味論的に豊かなネガティブ例を生成し得る。ラベル付けサービスは、生成結果を生成するために、ラベルが候補テキストへ正しく適用されるという尤度を推定する生成モデルを利用する。分類の成功率は、生成モデルから第2の生成結果を取得しそして第2の生成結果を使用することによりラベル確率を推定することにより、この改善された効率を維持する一方で改善され得る。

【0016】

別の態様では、本技術は、候補テキストに似ている意味論的に豊かな例を取得する方法に向けられる。この問題に対する他の解決策は、入力データの意味論的に劣悪な表現を提供してきた、又はそうでなければ、トレーニングを提供するために莫大な量の手動データに依存してきた。これら他の解決策のいずれも、分類を行うモデルをトレーニングするために大量のコンピュータ処理を必要としてきた。本技術は、ラベルクラスの多数の手動ユーザ入力例を必要することなく意味論的に豊かな例を生成する一方で良好な性能を提供することにより最新技術を改善する。本明細書において説明されるラベル付けサービスの要件をトレーニングする入力及びコンピュータははるかに低資源集約的であるので、コンピュータ化システムは、結果を描画するためにコンピュータ処理を余り必要としない技術的な改善を提供する。ラベル付けサービスは、例えばラベルの文脈内の候補テキストの豊さを反映する一組のキーワードを取得することにより、この改善を提供する。一組のキーワードは検索サービス164へ提示され、そして、良好な関連ランクを有する検索結果からのテキストスニペットが、抽出されたスニペットのラベルクラス信頼度が高い場合に例を提供するために取得される。

10

20

30

40

50

【0017】

別の態様では、本技術は、ラベルの文脈内の候補テキストから意味論的に豊かな一組のキーワードを提供する方法に向けられる。他の解決策は表現において意味論的に劣悪であったので、一定数の関連結果を取得するために受信されなければならない検索エンジンからの戻りの数は大きかった。この多数回の必要とされる戻りは高コンピュータ処理要件を意味した。本技術は、例えば意味論的に豊かな一組のキーワードを生成する一方で良好な性能を提供することにより、そして従ってトレーニングのために必要とされるデータの量を低減することにより最先端技術を改善する。一組の候補テキスト優先キーワードが候補テキストから取得される。一組のラベル優先キーワードがラベルから取得される。優先キーワードには、トランスフォーマベースモデルを使用することにより埋め込みベクトルが割り当てられる。文脈意識キーワードは、一組の文脈意識キーワードを取得するために、埋め込みベクトルに基づく優先キーワードの類似度により判断される。この文脈意識組のキーワードは、検索エンジンからラベルの文脈内の候補テキストに意味論的に近い情報を取得することを可能にし、従って、一定数の関連結果を戻すために必要とされる検索処理の量が低減される。

10

【0018】

定義

ラベルは通常、単一ワード／用語により又はモデルがトレーニングされるコンテンツ要件の記述により記述されるカテゴリである。ラベルは通常、自然言語入カストリングなどの別の電子エンティティが分類される可能性があるカテゴリである。

20

【0019】

アンチラベルは通常、ラベルにより記述されるクラスに属さない電子エンティティを含むカテゴリである。多項式クラスの文脈では、アンチラベルは、ラベルクラスに属さないすべての列挙されたクラスを含む。

【0020】

カスタムラベルは通常、所望ラベルカテゴリのインジケーションとしてユーザにより入力されるユーザ定義自然言語記述である。

【0021】

ラベル付けサービスは通常、ラベル又はラベル確率を自然言語ストリングなどの電子アイテムへ割り当てるアプリケーションである。

30

【0022】

ラベル採点サービスは通常、適用される可能性がある他の代替ラベルの文脈内のラベルからの候補の距離を測定するための候補自然言語入カストリングを採点するアプリケーションである。一般的に、ラベルスコアは、確率などの測度であり得、そして候補をラベルに関連する1又は複数のカテゴリ（ラベル、アンチラベル、ラベルのサブカテゴリ、又はアンチラベルのサブカテゴリなど）へ分類するために使用され得る。

【0023】

変換サービスは通常、用語又は一組の用語を取り入れそして同意語、反意語、語形などの操作に従ってそれらを変換するサービスである。

【0024】

優先キーワード抽出サービス（例えば図9）は通常、テキストストリングを取り込み、キーワードを抽出し、それらを順序付ける（例えば重要度の降順で順序づけられたキーワードのリストなどのラベル構造内で）サービスである。

40

【0025】

文脈意識キーワード抽出サービス（例えば図7）は通常、ラベルの文脈内の候補テキストを表わすキーワード抽出サービスである。

【0026】

用語類似度サービス（例えば図10）は通常、キーワードの構造（グラフなどの）に作用し、そして用語類似度を表わす（例えばグラフの用語間の重み付けグラフリンクージュにより）サービスである。

50

【0027】

検索／回収サービスとしても知られている検索サービス164は通常、文書のコーパス全体にわたるクエリーに作用し、そしてクエリーに特に関連する文書の一部分を提供するテキストスニペットと共にコーパスから文書の関連性ランク付けリストを戻す検索サービスである。

【0028】

自然言語処理（NLP：Natural Language Processing）アプリケーションは通常、自然言語入力ストリングに対するコンピュータ化操作を行うために音声入力又はテキスト入力などの自然言語入力に作用するコンピュータ化アプリケーションである。

【0029】

自然言語生成（NLG：Natural Language Generative）モデルは通常、生成入力に基づき自然言語テキストを生成するアプリケーションである。生成入力は例えば、トークン、一連のトークン、又は数の系列／ベクトルのような他のいくつかの入力機構であり得る。従って、これらのシステムは通常、教師無しラベル分類器の機能を行うことができないかもしれない。NLGモデルの例はGPT-2、GPT-3及びDeBERTaを含む。

【0030】

生成型事前学習済みトランスフォーマ（Pre-trained Transformer）モデルは通常、深層学習に基づきニューラルネットワークを使用する自己回帰言語モデルである。

【0031】

トランスフォーマモデルは通常、入力の広範な文脈を分類判断に関連し得る他の入力の文脈内に組み込むために注意喚起機構（attention mechanism）を利用する深層学習モデルである。

【0032】

移転学習（transfer-learning）モデルは、モデルが大きな教師無し及びラベル無しデータから少なくとも部分的に学習するニューラルネットワークモデルである。このようなモデルは、データにより（好適には同様な領域からモデルのアプリケーションへのデータにより）更に微調整され得る。

【0033】

zero-shot生成モードは通常、特定タイプのデータによる微調整無しにテキストを生成することができる生成NLPモデルのモードである。生成NLPモデルは通常、入力テキストストリングを受信し、そしてテキストである生成結果（入力テキストストリングのプロンプティングにおいて生成される）を生成する。

【0034】

教師無しラベル分類器は通常、ユーザにより提供されるラベル付けクラスの例を必ずしも必要としないが性能を強化するためにユーザ提供例を利用し得るラベル分類器を示す。

【0035】

意味論的検索モデルは通常、クエリー文書から一組の文書内の別の文書までの言語論的意味空間内の距離を測定し、そしてコサイン類似度などの測度（一組の文書内の当該文書までのクエリー文書の近さを表現する）を戻す深層学習モデルなどの学習モデルである。（意味論的検索モデルの例はDSMを含む）

【0036】

本明細書において説明される技術のいくつかの態様の概観について概説したので、本明細書において説明される技術のいくつかの態様が実装され得る例示的動作環境が、様々な態様の一般的文脈を提供するために以下に説明される。

【0037】

次に図1を参照すると、本開示のいくつかの態様が採用され得る例示的動作環境100を示すブロック図が提供される。本明細書において説明されるこの配置及び他の配置は単に一例として記載されるということが理解されるべきである。他の配置及び要素（例えば、機械、インターフェース、機能、順番及び機能のグルーピングなど）が、示されたもの

10

20

30

40

50

に加えて又はその代わりに使用され得、そしていくつかの要素は明確化のために纏めて省略され得る。更に、本明細書において説明される要素の多くは、個別部品若しくは分散部品として、又は他の部品に関連して、任意の好適な組み合わせ及び場所で、実装され得る機能的エンティティである。1又は複数のエンティティにより行われるとして本明細書において説明される様々な機能は、ハードウェア、ファームウェア、及び／又はソフトウェアにより実行され得る。例えば、いくつかの機能はメモリ内に格納された命令を実行するプロセッサにより実行され得る。

【0038】

示されない他の部品の中でも、例示的動作環境100は、ユーザデバイス105、サーバ125、クラウドサービス199、アプリケーションサービス175、ファブリックコントローラ179、サーバクラス176、サーバ177、ストレージサービス180、ネットワーク186及びネットワーク103などの多くのコンピュータデバイスを含む。図1に示される部品の各々は、任意のタイプのコンピューティングデバイス（例えば図8に関連して説明されるコンピューティングデバイス800など）を介し実装され得る。これらの部品は、ネットワーク103又はネットワーク186（限定しないが1又は複数のローカルエリアネットワーク（LAN）及び／又は広域ネットワーク（WAN）を含み得る）を介し互いに通信し得る。例示的実装形態では、ネットワーク103及びネットワーク186各々は、多種多様な可能な公的及び／又は私的ネットワークのうちの任意のものの中でも、インターネット及び／又はセルラーネットワークを含む。

【0039】

一態様では、本技術は、テキストをユーザ定義テキストラベルに属する又は属さないかのいずれかとして分類する方法を行うコンピュータ化システム（例えば、動作環境100で示される）に向けられる。動作環境100内のラベル付けアプリケーション110はディスプレイ120上にユーザへのプロンプトを提示し得る。ディスプレイ120は視覚的ディスプレイ又はスピーカであり得る。デバイス105内のマイクロホン、マウス又はキーボードなどのユーザ入力デバイス115がユーザからの入力を受信する。いくつかの実施形態では、この入力はユーザ定義テキストラベルとして働く自然言語ストリングであり得る。一実施形態では、オペレーティングシステム107は音声信号入力をテキストストリングへ変換し、そしてラベル付けアプリケーション110はこのテキストストリングを入力として受信する。一実施形態では、オペレーティングシステム107はキーボード115からキーストロークを受信し、そしてテキストストリングをラベル付けアプリケーション110へ提供する。ラベル付けアプリケーション110はまた、同様なやり方でユーザから分類される候補テキストを受信する。候補テキストは、ユーザ入力から又はシステム文書のコーパス154内の文書からラベル付けアプリケーション110により受信される可能性がある。このプロセスの終わりに、ラベル付けアプリケーション110は、ディスプレイ120上に提示される「候補テキストがユーザ定義ラベルに属する可能性が高い」というインジケーションなどの分類の結果を提供する。

【0040】

コンピュータデバイス105及びサーバ125は動作環境100のクライアント側のクライアントデバイスであり得る一方で、サーバ125、サーバ177、クラウドサービス199、アプリケーションサービス175、ファブリックコントローラ179、サーバクラス176及びストレージサービス180は動作環境100のサーバ側に在り得る。コンピュータデバイス105は通常、オペレーティングシステム107、タッチスクリーンセンサ又はマウスなどのユーザ入力デバイス115、及びディスプレイ120を含む。コンピュータデバイス105はまた、重要なことには、例えばブラウザ、プラグイン、ダウンロード可能アプリケーション、検索アプリケーション、情報管理システム、専用アプリケーション、ラベル付けアプリケーション、ラベル支援検索アプリケーション、ラベル支援型分類プログラム、文章作成支援、自動化コンプライアンスアプリケーション、顧客関係管理アプリケーションなどであり得るラベル付けアプリケーション110を含む。ラベル付けアプリケーション110はまた、サーバ177上に示されたアプリケーションに関

連するこれらのアプリケーション機能のうちの1又は複数を行うユーザインタフェース部品であり得る。一実施形態におけるリモートサーバ177上のアプリケーション及びデバイス105上のアプリケーションはサーバ125上に存在する。

【0041】

一実施形態では、ラベル付けアプリケーション110は、ユーザのために提供された機能をラベル付けアプリケーション110により協働で行うためにリモートサーバ177上の部品と通信する。例えば、ラベル付けアプリケーション110と協働する部品は、ラベル付けサービス142、ラベル採点サービス168、用語変換サービス144、検索サービス164、優先キーワード抽出サービス146、自然言語生成（NLG）モデルレポジトリ162、文脈埋め込み生成モデル158、文脈意識キーワード抽出サービス148、ベクトル化機能156、用語類似度サービス152、コーパス130、コーパス195、及びコーパス154を含み得る。これらの部品は、一組のコンパイル済みコンピュータ命令若しくは機能、プログラムモジュール、コンピュータソフトウェアサービス、又は例えば図8に関連して説明されるコンピューティングデバイス800などの1又は複数のコンピュータシステム上で行われるプロセスの配置として具現化され得る。サーバ177は、本開示において論述されるフィーチャ及び機能の任意の組み合わせを実装するためにユーザデバイス105上のクライアント側ソフトウェアと協働して働くように設計されたサーバ側ソフトウェアを含み得る。例えば、サーバ177は、デバイス105の情報管理システム（知識グラフ内の情報へのアクセス及びその使用を管理する）を実行し得る。サーバ177は、格納のための文書、スプレッドシート、電子メール、ソーシャルメディアポスト、ユーザプロフィール等々のファイルなどのデジタル資産を多くのユーザに属する多数のユーザデバイスから受信し得る。動作環境100のこの分割は好適な環境の一例を示すために提供されており、従って各実装形態に関し「サーバ177とユーザデバイス105との任意の組み合わせが別個のエンティティとして残る」という要件は無い。

【0042】

ユーザデバイス105及びサーバ125などのコンピューティングデバイスは、ユーザによる使用が可能な任意のタイプのコンピューティングデバイスを含み得る。例えば、一態様では、ユーザデバイス105及びサーバ125は、図8に関連して本明細書において説明されるタイプのコンピューティングデバイスであり得る。一例としてそして制限すること無く、コンピューティングデバイスは、パーソナルコンピュータ（PC）、ラップトップコンピュータ、モバイルデバイス、スマートフォン、タブレットコンピュータ、スマートウォッチ、ウェアラブルコンピュータ、適合性追跡器、仮想現実ヘッドセット、拡張現実眼鏡、パーソナルデジタルアシスタント（PDA）、MP3プレーヤ、全地球測位システム（GPS）又はデバイス、ビデオプレーヤ、ハンドヘルドコミュニケーションデバイス、ゲーミングデバイス又はシステム、娯楽システム、車両コンピュータシステム、埋め込み型システムコントローラ、リモートコントロール、アプライアンス、消費者電子デバイス、ワークステーション、ファイルサーバ、ウェブサーバ、アプリケーションサーバ、ホストコンピュータ、企業サーバ、サーバのクラスタ、データセンタ、検索アプライアンス、仮想サーバ、デーモン、メインフレーム又はこれら列挙されたデバイスの任意の組み合わせ、又は任意の他の好適なデバイスとして具現化され得る。

【0043】

本開示は、自然言語処理（NLP）をトレーニングすることに繋がる代表的データを生成するためにそうでなければ直接的又は間接的に使用される可能性がある表現を生成するために、又は興味ある1又は複数のクラス（クラスラベル）全体にわたって候補入力テキストをマッピング／分類する可能性があるテキスト分類モデルを生成するために、代表的ラベル付きデータ又は人間等級付け器の援助の必要性無しに、テキスト分類モデルをトレーニングするシステム及び方法を説明する。

【0044】

一般的に、不偏（unbiased）非代表的トレーニングデータに関しトレーニングされた（代表的ラベル付きデータが無い場合に）不偏テキスト分類モデルは、2進法分類に関し

10

20

30

40

50

てせいぜい50%精度を主張する可能性がある。これは、すべてのデータを2進法ラベル分類モードの「ポジティブラベルクラス」又は「ネガティブラベルクラス」としてラベル付けする人間認識に匹敵する。これは任意の候補モデルを比較するための科学的根拠（参照：ROC曲線のベースライン）として使用される。

【0045】

本システムの技術は、上記人間発見的又は不偏モデル分類（非代表的データに関しトレーニングされた）により可能であるものより正確であるだけでない。いくつかの実験の結果は、より良い精度と、現実シナリオにおけるモデルの有用性に関するより多くの客観的情報を意志決定者に与える確立された「再現率及び／定義済み偽ポジティブ率」（FPR：False Positive Rate）とを実証した。

【0046】

図2を参照すると、例示的グラフィックディスプレイ200は、顧客関係管理（Customer Relationship Management）システムの機能を行う例示的ラベル付けサービス142のためのラベル付けアプリケーション110として行うブラウザアプリケーションのユーザディスプレイを示す。CRMシステムのコーパス154は、販売、マーケティング、及びテキスト、ウェブ及び電子メール上のサービスコミュニケーションを収納する。グラフィックエリア202は制御要素を提供する。当初、ユーザは候補テキストを定義するためにテキストを提供する。ラベル付けアプリケーション110は候補テキストを受信する。ユーザがラベルを定義するためのテキストをグラフィックコントロール206へ入力すると、ラベル付けアプリケーション110はラベルテキストストリングを受信する。ラベルを定義するテキストストリングは、任意の概念又はアイデアのワード、用語又は記述であり得る。

【0047】

ラベル付けアプリケーション110はラベル付けサービス142へ2つのストリング（候補テキスト及びラベル）を送信する。ラベル付けサービス142は、ラベル処理を行い、そして1又は複数の結果をラベル付けアプリケーション110へ提供する。次に、ラベル付けアプリケーション110は、グラフィック表示エリア209、231、235、261、262、257、251、253、212、214、216、292及び204内に表示された結果などの結果を含むようにグラフィックディスプレイ200を更新する。

【0048】

表示エリア204は、ラベル付けサービス142により判断されるような候補テキストを表わす一組の文脈意識キーワードを示す。表示エリア212は、候補ラベルテキストから導出されたアンチラベルを表わすキーワードの順序付きリストを示す。表示エリア292は、候補ラベルのアンチラベルを表わしそして候補ラベルテキストから導出されるキーワードの順序付きリストを示す。表示エリア214は、候補ラベルテキストから導出された一組のアンチラベルキーワードを示す。表示エリア216は、候補アンチラベルテキストから導出された一組のアンチラベルキーワードを示す。表示エリア209は「候補テキストが規定ラベルクラスに属する」確率の推定を示す。加えて、表示エリア209は、確率の推定に適用されるラベルクラスメンバーシップの閾値判断に基づく結果（真又は偽など）の表示を提供する可能性がある。

【0049】

ラベル付けサービス142は、2進法分類出力を提供するために候補ラベルクラス予測（真に対し1、偽に対し、等々）を戻す。一般的に、すべての描写された結果データは、ラベル付けサービス142からラベル付けアプリケーション110により受信され、ラベル付けアプリケーション110はディスプレイ120上に提示されるべき描写された結果データを提供する。一般的に、描写されるラベル付け結果は、その表示又は使用がラベル付けシステムの部品へ提供される任意のラベル関係情報アイテム（例えばラベル付けサービス142が「候補テキストが判断基準を許容可能に満たす」ということを判断した場合に動作環境100内に示される）である。例えば、許容可能性判断基準は、ラベルクラスの確率の推定が許容可能閾値の上にあるということかもしれない。本システムは、規定ラ

10

20

30

40

50

ベルへ正しく分類されるテキストのいかなる例もユーザが必然的に提供することを必要としないので、描写された結果は教師無し方法において提供されることができる。

【0050】

一実施形態では、グラフィックディスプレイ200は、追加候補テキストを入力するようにユーザに促すためにグラフィックコントロール202からの候補テキストをクリアする一方で許容可能性判断基準を満足する候補テキストをポジティブ例表示エリア231内へ置くために更新される。この候補テキスト入力を繰り返すことにより、ユーザはまた、グラフィック表示エリア231及び235内に示されるものなどのようなポジティブ例並びに表示エリア261、262、257、251及び253内に示されるものなどのネガティブ例を生成するために意味論的言語処理を行うコンピュータ支援によりポジティブ例及びネガティブ例のラベルのライブラリを構築することができる。本方法は、候補の自動分類を提供し、そしてポジティブ及びネガティブ例及びキーワード構造を含むように一組の入力データを拡張する。候補テキスト入力を繰り返すことで、追加アンチラベル定義をグラフィック表示エリア224、222及び218内だけでなく追加ネガティブ例表示エリア267、277及び287内にも常駐させ得る。グラフィックディスプレイ200内の大文字A、B、C、D、E及びFは「アンチラベル表示エリア212、214、216、224、222及び218は、ラベル付けサービス142が表示エリア260、250、256、265、270及び280内に示されるアンチラベル例に対応するために判断したサブカテゴリである」ということを示す。このようにして、豊富な一組のアンチラベルサブカテゴリは、動作環境100内に示されるラベル付けシステムにより判断され、そして直観的及び有用なユーザグラフィックディスプレイ200においてユーザへ表示される。このディスプレイは、一組のアンチラベルキーワードと対応例とをペアにし、そして、アンチラベルの有用性又はアンチラベルに関係する対応例の有用性に関するフィードバックをユーザが提供することを可能にする。一般的に、結果を提供するいかなる表示エリアも232などの関連制御を有し得る。表示エリア231などの例示的表示エリアは、ユーザが「隣の例がシステムにより割り当てられたラベルに適合する」という確認をオーバーライド又は提供することを可能にする対応グラフィックコントロール232と共に示される。グラフィックコントロール232は「はい」のチェックボックス及び「いいえ」のチェックボックスを有するプロンプト（「これは良い例ですか？」など）を含み得る。代替的に、ディスプレイは、「グリーン」又は「良」とマーキングされ、そして選択されると悪い例と平凡な例とを示すために「レッド」又は「イエロー」とをトグルするラジオボタンであり得る。グラフィックコントロール232と同様なのはグラフィックコントロール236、259、252及び254である。グラフィックコントロール239は、すべての示されたポジティブ例が表示エリア231及び235に関し確認又は拒絶されることを許容する。同様に、グラフィックコントロール255は、表示エリア251及び253が1つのコントロールにより確認又は拒絶されることを許容する。

【0051】

グラフィックディスプレイ200は本明細書において開示される技術の一例を提供するが、システムは、候補テキスト（例えばグラフィックコントロール202内に示されるテキスト）が異なるユーザから又は文書コーパス154（セールスマンの電子メールからの文章など）から来るとクラスラベル付けを判断するように動作可能である。グラフィックディスプレイ200のユーザは、ラベル入力の最低限の定義（「楽しく及びビジネスライク：Pleasant, and Business-like」など）だけをグラフィックコントロール206内へ提供するCRMマネージャである可能性がある。次に、ラベル付けサービス142はラベルクラスを定義するためにコーパス154内の文書全体にわたり検索し、書かれた文章を試験し、そしてラベル例ライブラリを構築することによりライブラリを構築することを開始し得る。従って、表示エリア200は、はるかに明瞭な表示を当初提供し、ユーザに可視であるグラフィックコントロール206だけを提供する可能性がある。ユーザがラベルをグラフィックコントロール206内へ入力した後、数回の繰り返しが行われる可能性があり、そしてグラフィックディスプレイ200は、コーパスの上に提示されたラベル

10

20

30

40

50

の実行可能性の推定を表示する可能性がある、そうでなければ一組のリンクをユーザにより提示された記述に最も近い文書又は文書の一部へ提供する可能性がある。加えて、ラベルベース文書検索能力が、ユーザにより開発されたラベル定義を論理的に組み合わせることにより提供される可能性がある。各ラベル分類器が十分な性能に達した後、ラベル分類器は、ユーザのライブラリ内に置かれる可能性がありそしてユーザが定義したラベルの組み合わせ文脈において高スコアを提供する文書を発見するために組み合わせられる可能性がある。

【0052】

グラフィックディスプレイ200は、分類器を改善するためにユーザが低レベルフィードバックをシステムへ提供することを可能にする多くの表示エリアを提示する。ディスプレイは通常、アンチラベル表示エリア210、クラス定義表示エリア201、ポジティブ例エリア203及びネガティブ例エリア205を含む。一般的に、グラフィックディスプレイ200は、任意のユーザ入力結果を、ラベル付けアプリケーション110からラベル付けサービス142へ信号伝達されるデータにおいて提示する。

10

【0053】

一実施形態では、入力テキストは任意の長さのものである。ラベルは単なる短い文章又は文書である可能性がある。必要とされるラベルはポジティブクラスのラベルの観点で与えられる；ネガティブクラスはポジティブクラスの欠如として扱われる。これは、ワード、用語、又は任意の概念又はアイデアの記述のいずれかである可能性がある。

20

【0054】

本明細書において説明される技術は広範に適用可能である。本技術は多くのシステムに権限を与える可能性がある。例えば、この技術の1つの使用は自動化されたコンプライアンスのためのものであり、ここでは、テナントアドミが、全企業データコーパス（電子メール、チャット、文書レポジトリ、契約書などを含む）が当該時に必要と感じられ得る任意の概念に関しラベル付けされることを要求する可能性があり、そして法的理由のためだけでなくビジネス理由のためにもタイムリーな応答が重要である。このような洞察が必要とされる（以下の1又は複数の理由のために単独人間又は人間チームの集合のいずれによっても整合されることができなかった規模、速度、目的又は公平レベルで）いくつかの他のアプリケーションが存在する。

30

【0055】

本明細書において説明される技術は効率的にスケーリングされる。本明細書において説明される技術は、いかなる数の人間も所望目的のために手動で及び客観的に処理することが実現可能でない企業規模データ（電子メール、チャット、文書レポジトリ、契約書などを含む）向けに意図されている。説明される技術は低遅延を有し、従って大量のテキスト入力を効率的に処理し得る。本技術は、いかなる数の人間も所望目的のために手動で及び公正に処理するのが実現可能でない大きなデータを処理する必要があり並びにビジネス目的及び法的目的のための効果的及び有用であると考えられる合理的時間内に出力を供給する必要があるアプリケーション用に意図されている。

40

【0056】

本明細書において説明される技術はユーザプライバシー及び機密性を維持する。クリティカルデータの人間処理は、いくつかのリスクが、人間解析者をバルクラベル分類努力に関与させることに関与するので脆弱である。法的及びコンプライアンス要件に加えて、非企業データに関してすら、このようなデータを単独ユーザ又はユーザのチームに晒すことは賢明でない又は更には実現可能ではないかもしれない。

【0057】

本明細書において説明される技術は優れた公平性を有する。人間認識は、独特なものではなく、所与の概念の知識から、所与の文脈の理解から、及び特定言語の知識からしばしば偏らされる又は制限される。従って、このようなタスクは、様々な人間により行われると、所望どおりに広範なアプリケーションにおいて確実なものとなれないし制御もされない可能性がある偏りのクリティカルリスクを孕む。本明細書において説明される技術のキ

50

ワード文脈をユーザへ晒すことにより、ユーザは、貧弱に定義されたラベルを、又は「ワードが意味したとユーザが考えたものを正確に意味しなかったワード」を使用したラベルを修正することができる。

【0058】

本技術は優れた客観性を有する。本システムの目的は、任意の候補テキストの任意の候補ラベルレベルの予測だけでなく、多くの下流アプリケーションにおいてそして本開示が権限を与える関連ソフトウェアフィチャにおいて必要とされる又はそうでなければ有用である関連信頼度である。人間の認識は、個人の制限された理解のおかげで概して偏らされるので、認識の精度又は特定候補ラベルレベルのいかなる客観的に定義され及び監査可能な信頼度数も生成し得ない。

10

【0059】

本明細書において開示される技術は優れた多言語能力を有する。いかなる人間も、様々な言語の自身の知識及び運用能力により、そして既知の言語においてすら様々な概念の運用能力により制限される。従って、単独人間の認識は十分ではないかもしれない、そして人間の認識の集合は言語、概念及び専門知識の様々な組み合わせ全体にわたって一貫していないかもしれない。

【0060】

本明細書において開示される技術は優れた監査可能性及び再現性を有する。コンプライアンスを必要とする複数のドメインにおいて及びアプリケーションにおいて、客観性をプロセスに組み込むことだけでなく再現性及び一貫性を実証することはクリティカルであり得る。人間認識ベースシステムはこれらのドメイン及びアプリケーション内へ採用され得ない。

20

【0061】

本明細書において説明される技術は優れた経済性及び信頼性を有する。現在、多くのラベル付け要件がいくつかの所定ラベルに適用されており、従って多くのラベル付け要件は莫大な費用で、劣悪信頼性で、及び非常に限られた規模で達成される。通常、競合方法は、有料割り当て（解析者、売り手又は人間契約者）を介し、又はクラウドソーシングを介し行われる。有料割り当てに関与するものは高価である。クラウドソーシングに関与するものは信頼できない。

【0062】

30

図3は、表示エリア209内に示されるような結果（ラベル付けサービスが妥当な推定を提供するためにエラーメッセージ無しに完了するとラベル付けアプリケーション110へ送信される）を描画するコンピュータ化方法を行うラベル付けサービス142の処理フローを示す。一般的に、サーバ177上のラベル付けサービス142を初期化することが有利であり得る。ラベル付けサービス初期設定の一部として、NLGモデルがサーバ177のメモリ内へロードされる。一実施形態では、NLGモデルは、大規模サービスを提供するために複数の現実又は仮想サーバを使用することによりクラウドサービス199内でホストされる。レポジトリ162内で利用可能な生成NLPモデルがロードされる（又は事前ロードされたままである）。より良い結果のために、より大規模及びより表現豊かなモデルが使用され得る。モデルは、好適には事前トレーニングされ（転移学習（モデルが大きな教師無し及び無ラベルデータから部分的に学習する）の概念）そしてデータ（好適にはアプリケーション要件と同様なドメインからの）により更に微調整され得る。同様なモデルのいくつかの例は、GPT-3及びMicrosoft DeBERTaなど（好適には、良いzero-shot生成能力モード（「モデルが特定タイプのデータによる微調整無しにテキストを生成する可能性がある」モード）を有するモデルで）である可能性がある（限定しないが）。NLP生成モデルにおける技術（SOTA）の現在状態は、大規模（100億超のトレーニング可能パラメータ）トランスフォーマベース（Transformer based）モデルである。本開示は、本開示をこれらのモデルの使用に制限しないので、本明細書において開示された1又は複数の採点機構と互換性があるようにさせる可能性がある任意の利用可能モデルが使用される可能性がある。

40

50

【0063】

ラベル採点サービス168における工程を行うためにレポジトリ162から取られそしてラベル付けサービス142により採用されたNLGモデルは通常、ラベル無し自然言語コーパス全体にわたってトレーニングされる。同様に、NLPモデルは、モデル158のグループ内に格納されそして文脈埋め込みを生成するために使用されるか、変換サービス144を行うために又はベクトル化156を行うために採用されるかにかかわらず、同様に、一般的に、ラベル無し自然言語コーパス全体にわたってトレーニングされる。このようなモデルは通常、トークンマスキング技術を適用することによりトレーニングされる。一実施形態では、本サービスにおいて採用されたNLPモデル又はNLGモデルは、ウェブコーパス、企業データコーパス、又は別のコーパス全体にわたってトレーニングされる。本明細書において開示される技術は、他のモデルの中でも、ニューラルネットワークモデル、非ニューラルネットワークモデル、部分的事前トレーニング済みモデル、十分にトレーニングされたモデル、及び調整済みモデルにより操作可能である。

10

【0064】

ラベル付け結果を描画する方法（例えば方法300）は、ラベル付けサービス142が表示ページをラベル付けアプリケーション110へサービスすると工程303で始まる。工程305では、ラベル付けサービス142は、コーパス154内の文書から又はラベル付けアプリケーション110から候補テキストを定義するテキストストリングを受信する。工程310では、ラベル付けサービス142はラベルを定義するテキストストリングを受信する（例えばラベル付けアプリケーション110から）。工程307では、受信されたラベルが複数のワードを有すれば、ラベルのキーワード構造がラベル付けサービス142により判断される（例えば図9の例に関し説明されるように）。ラベルが複数のワードを有する場合、方法300はラベルを残りのプロセスと互換性がある関連レベルにするために追加サブ工程を使用する。一実施形態では、キーワードアルゴリズムは利用可能抽出テキスト要約及びキーワード抽出アルゴリズム（available extractive text summarization and keyword extraction algorithm）である。一実施形態では、例示のために、入力が「楽しく及びビジネスライク」である場合、出力は、タグディスプレイ1160も示す図11に示すラベルグラフ1130に示すように順序付き候補ラベル（「サービス」スコア＝0.6）、（「ハーモニー」スコア＝0.4）である。

20

【0065】

図9を手短に参照すると、工程710を行う方法において優先キーワード抽出のためのコンピュータ化方法（工程903で始まる）が提示される。本方法は工程905へ移り、ここで要約されるべきテキストがキーワード抽出サービス146により受信される。本例では、ラベルテキスト「楽しく及びビジネスライク」が受信される。加えて、工程710を行う方法は、生成される構造のサイズを制限する制約を受信する。例えば、サイズ制約は、保持すべきサービスのトップキーワードの最大数である可能性があり、そして工程710を行う方法によりストレージサービス180から受信され得る。別の実施形態では、サイズ制約は、工程710を行う方法によりストレージサービス180から受信されるキーワード強度閾値であり得る。次に、サイズ制約は、非重要用語をフィルタリングするために工程940において後で適用される。工程910では、テキストは、余分な文字が削除されそしてさらなる処理のためのテキストが用意されるように浄化されそして事前処理される。一実施形態では、テキストは追加処理を単純化するためにすべて大文字へ変更される。本方法は工程915へ移り、ここで浄化されたテキストが用語へトークン化される。一実施形態では、テキストの元の表現は同意語を介しよりコンパクトなボキャブラリへ変換される。工程920では、文章の用語はベクトル化される。そして変換が適用される。ベクトル化機能は通常、一組の用語を意味のある数値表現へ変換する機能である。ベクトル化機能の例はTerm Frequency Inverse Document Frequent (TF-IDF)、Global IDF、Entropy Weightingを含む。工程925では、ベクトル変換メトリックに関する閾値が、非重要用語をフィルタリングするために使用される。残りの用語は、グラフの頂点を形成するために

30

40

50

工程 930 において使用される。工程 935 では、グラフ内の各頂点（用語）は、用語間の類似度を表すエッジ重み付けによりエッジをグラフ内の互いの頂点に対し描写することによりグラフ内の他の用語との類似度に関して定量化される。

【0066】

一実施形態では、工程 935 は、共起（co-occurrence）ベース用語類似度を計算する方法 1000 を利用する。類似度計算は工程 1003 で始まり、そして工程 1005 へ移り、ここで用語のグラフが受信される。本文脈におけるグラフはグラフ内で接続される優先キーワードのグラフである。工程 1010 では、連結検索用語カウント又は「用語距離：Term Distance」が取得される、又はデフォルト値が使用される。例えば、入力無しでは、「用語距離」のデフォルトが文章内の用語の数の平方根として取られる。一実施形態における連結検索語計算は 2～10 の間の整数であり、これは、連語（collocated term）の検索においてどれだけの用語を考慮すべきであるかを伝える。連結検索は 2～10 の間のカウントが割り当てられれば隣用語と 9 番目の隣接用語との間の用語に関して行われることになる。本方法は工程 1015 へ進み、ここで各用語が用語距離内のペア毎に連結される回数が発見される。グラフ内の各頂点（用語）はグラフ内の別の用語との関係性に関して考慮される。2つの用語が「用語距離」内に共存する回数がカウントされる。工程 1020 では、共起頻度は 1 を加えるように正規化されスケーリングされる。工程 1025 では、各正規化されスケーリングされた頻度が 2つの頂点間のグラフリンク重み付けへ割り当てられる。工程 1030 では、用語重要度が必要に応じて計算される。頂点（用語）毎に、用語重要度が頂点からのすべて出力エッジの正規化スコアに応じて判断される。工程 1035 では、グラフエッジ重み付けが戻される。本方法は 1097 で完了する。

【0067】

工程 940 では、サイズ制約（例えば閾値）が非重要用語をフィルタリングするために適用される。このフィルタは弱いキーワードを削除する。工程 945 では、キーワード構造が出力される。一実施形態では、出力はその結果のグラフ構造である。グラフは、優先頂点、それぞれのエッジ重み付け及び頂点スコアを有するサブグラフであり得る。一実施形態では、出力は順序付けされた一組のキーワードである。本方法は工程 997 において完了する。

【0068】

図 3 に戻ると、工程 307 において判断されたキーワード構造は、とりわけ方法 300 による工程 330 における候補ラベルの一例の生成を支援するためにラベル付けサービス 142 により格納される。本方法は工程 372 へ進み、ここでアンチラベル構造がラベル付けサービス 142 により生成され格納される。アンチラベル構造は、とりわけ工程 345 における候補アンチラベルの一例を生成する手段として使用される。アンチラベルの生成の多くの様々な方法が採用され得る。例えば、表示エリア 212 内に示されるアンチラベル “bossy disharmony” は表示エリア 292 内に示される個々のキーワードの反転により生成された。加えて、全一組のラベルキーワードは、用語変換サービス 144 により採用される可能性がある文脈意識反転サービスを介し、又は文脈内で使用されるワードの反意語を供給する NLP ベクトル化埋め込みアルゴリズムなどの最先端ベクトル化技術により反転され得る。加えて、ラベル付けサービス 142 により格納された一組のラベル及びアンチラベルは、別個のラベル付け文脈を形成するために、関連する明示的又は暗黙的ユーザ承認と共にライブラリ内に格納され得る。この手法は、同様な言語文脈又は共有言語文脈を有するユーザ又は一群のユーザのラベル付け努力をより注意深く追跡し使用するために採掘され得るコミュニケーションの個別エリアとしてラベル用語の抽象的使用をインデックス付けする可能性を有する。追加反転技術として、使用される言語反転方法は用語「ビジネスライク」に焦点を合わせ、そして表示エリア 214 内に示される「インフォーマル」を見つけた。更に、例示的埋め込みベクトル反意語定位機能は、表示エリア 216 内に示されるあり得る反意語「自己中心的：self-focused」を戻した。アンチラベルカテゴリはしばしば複合的であるので、方法 300 は、グラフィックディスプレイ 200

に示すようにいくつかの例を生成するために発見された1つの又はすべてのアンチラベルを採用することができる。更に、同様な意味論的方法は、ラベルクラスを使用することにより同様な方法のラベル候補同意語を増やして、説明されたアンチラベルクラスと同じぐらい意味論的に豊かなラベルクラスを取得するためにラベルクラスへ適用され得る。

【0069】

方法300は、一例を含むように入力データを拡張する（又は換言すれば候補ラベルからいくつかの例を取得する）方法を行う工程315へ進む。図3に示す例は、ユーザがラベル又はアンチラベルのいずれかの2つ以上の例を提供すると1つのポジティブ例と1つのネガティブ例とのバランスのとれた初期セットを提供する。工程320において、利用可能候補ラベルのポジティブ例があれば、本方法は工程335へ進み、ラベル付けサービス142は、1又は複数のラベル採点方法により受信されるべき入力としてポジティブ例を受信する。同様に、工程325では、ユーザにより供給されたネガティブ例又はアンチラベル例があれば、方法300は工程340へ進み、そしてラベル付けサービス142は1又は複数の採点方法により受信される入力としてネガティブ例を受信する。ユーザがポジティブ例を提供していなかった場合、本方法は工程320から工程330へ進み、ここでラベル候補の一例が生成される。同様に、ユーザがネガティブ例を提供していなかった場合、本方法は工程325から工程345へ進み、ここで候補アンチラベルの一例が生成される。ラベルに関する情報から工程330においてラベル例を生成するために又はアンチラベルに関する情報から工程345においてアンチラベルの例を発見するために行われる方法は、同様なプロセスであるが異なる入力を有するプロセスに従い得る。

【0070】

工程330においてポジティブ例を取得する例示的方法は、ラベルから導出された順序付きキーワードを使用することによりそして図6に示す拡張方法600の少なくとも一部を使用することによりコーパス154全体にわたる検索を行うことを伴う。具体的には、コーパス154全体にわたる検索は、クエリーとしてラベルの優先キーワードを使用することにより工程620において行われる。工程625では、テキストスニペットが取得され、そして本方法は「テキストスニペットがラベルクラスに属する」という信頼度を定量化する工程630へ進む。クラス信頼度を定量化する例示的方法はテキストスニペットのキーワード構造を構築することである（例えば図9の工程710を行う方法を使用することにより）。テキストスニペットのキーワード構造とラベルキーワード構造との間の全体的意味論的類似度を評価する例示的方法は、グラフ用語のベクトル化変換に基づくコサイン類似度の使用又はベクトル化機能156により提供される他の方法である可能性がある。本明細書において開示される他の方法は、類似度スコア又はラベルがテキストスニペットへ正しく適用される確率の推定を提供する。確率が判断635において低過ぎれば、本方法は別のテキストスニペットを取得する工程625へ戻り、この別のテキストスニペットは工程630において定量化されそして工程635において試験される。クラス信頼度が判断635において十分であると、本方法は工程645へ進み、ここで入力是一例として十分なスニペットを含むように拡張される。同様な方法が、ネガティブ例を生成するために工程372において生成されたアンチラベルに合致する例を生成するために工程345において適用される。

【0071】

候補アンチラベルの例を生成する工程340の別の例示的方法は、検索サービス164を使用することにより、ラベルから優先キーワードの順序付きリストを使用することにより、そして低ランクのエントリのテキストスニペットを使用することにより、文書のコーパス全体にわたる検索を行うことに関わる。キーワードベースインデックスに関して、この手順は、コーパス内で有力である結果であるが「当該エントリが、ラベル内の他のワードの文脈と全く関係がないコーパス内のワードの共用に合致した」という理由のためだけで含まれる結果を戻す可能性が高い。テキストスニペットは最低ランク付けされた戻りから取得される（k番目戻り：ここでk番目戻りは、異なる文脈において使用される有力なコーパスワードを発見する可能性が高い）、ここで例えば $K=100$ 。同様に、コーパス

全体にわたるアンチラベル優先キーワードのクエリーは、アンチラベルクラスを定量化しない有力なキーワードを戻すだろう。このような遠い戻りの距離はまた、アンチラベルクラスからのラベルの分離可能性に関する重要な情報を提供し得る。

【0072】

一例を生成する別の例示的方法は一例がユーザにより提供された場合の追加（又はバランシング）例の計算に関わる。例えば、「図2の表示エリア231内に示される例が、ユーザにより、表示エリア231内に表わされるグラフィックコントロール内へタイプ入力された」ということを仮定する。この場合、工程315における方法は工程335へ進みポジティブ例を受信することになるだろう。工程325では、本方法は利用可能な候補アンチラベル例が存在しないので工程345へ進むだろう。この場合、ラベル付けサービス142は工程345において、方法600の少なくとも一部（拡張方法600が始まる工程603で始まる）を行う。本方法は工程605へ進み、ここで候補テキストが方法600により受信される。本方法は工程610へ進み、ここで本方法は候補アンチラベルの表現としてアンチラベルキーワード構造を受信する。工程615では、候補テキストのための一組の優先キーワードが用意される。本例では、これは、優先グラフを有するポジティブサンプルテキストを要約するために図9に示す工程710を行う方法を行うことによりポジティブ例の優先キーワードを最初に取得することにより発生するだろう。次に、グラフは反転される（例えばラベルのグラフは工程372において反転された）。次に、本方法は、図7の工程615を行う方法を行うことに進み、アンチラベルの文脈内の反転済みグラフの一組の文脈意識キーワードを生成するために工程720において始まる。工程720において、ネガティブテキストキーワードの優先用語の埋め込みベクトルが取得される（例えば文脈埋め込み生成モデル158から）、そして高優先用語だけが保持される。工程725において、アンチラベルキーワードの優先用語の埋め込みベクトルが取得される。例えば、反転済みテキストの各用語には埋め込みベクトルが提供され、そしてこのリストは優先用語だけを保持するためにフィルタリングされる。工程730において、優先アンチラベル用語と優先反転済みキーワードとの間の類似度が取得される。これは類似度（例えばアンチラベル内の各優先用語の埋め込みベクトルと反転済みテキスト内の各優先用語とのコサイン類似度）を計算することにより取得される可能性がある。工程735において、優先テキスト用語の文脈重要度が計算される。一実施形態では、各要約キーワード用語の文脈重要度は、アンチラベル内の各用語間の類似度の正規化加重平均として計算され、ここで重み付けはアンチラベル用語の重要度スコアである。工程740において、本方法は、文脈重要度及びキーワード優先度から文脈意識優先度を判断する。例えば、各要約キーワードの文脈意識優先度は、文脈重要度とキーワード優先度との正規化積として計算され得る。工程615を行う基本的方法は、工程703において通常始まるということ以外は、ポジティブテキスト及びポジティブラベルのような様々な入力に関して同じである。工程615を行う方法はまた、判断705において入力ラベルが複数の用語を有するかどうかの試験を含み得、そして真であれば、本方法は工程710へ進み、そして工程715へ戻る前に他のどこかに示される操作を行い、ここで優先テキストキーワードを提供する候補テキスト構造が判断される。工程615を行う方法は工程797で終わり、そして本方法は本事例では図6の方法600の工程620に戻る。次に、本方法はどこかに示すように工程620において一組のランク付けされた検索回収結果を取得することに進み、次に工程625においてエントリからテキストスニペットを取得し、そして次に工程630においてテキストスニペットのラベルクラス信頼度を定量化する。しかし、この場合、本方法は、アンチラベル及びポジティブ例のテキストに基づき工程380を行うラベル採点方法のうちの1つを利用し得る。方法600は工程635へ続き、そしてクラス信頼度が十分な場合、本方法は工程645へ進み、ここでこの場合、方法600が完了し、そして図3に示すように方法300の工程380に戻る。

【0073】

工程380を行う方法は図4の方法400、図5の方法500又は図12の方法1200において提示される1又は複数の採点方法を採用する。採点方法は通常、いくつかのポ

10

20

30

40

50

ジティブ又例はネガティブ例、ラベル及びアンチラベルを受信し、そして、ラベルが存在する確率の候補テキストをNLGモデルを使用することにより採点する。本方法はまた、生成された例に基づき、提供された入力を有する分類のGPT-3を利用し、そしてまた戻り、GPT-3の精度の以前の経験に対するラベルの類似度に基づきGPT-3確率の精度を推定する可能性がある。

【0074】

本開示は、ラベル採点の工程380を行う少なくとも4つの異なる方法を示唆する。図4の方法400として示された第1の方法は数値クラス(NC: Numeric Class)方法として知られている。図13の第2の方法1300はストリングラベル(SL: String Label)方法として知られている。図5の第3の方法500は検索スコア(SS: Search-Score)方法として知られている。図12の第4の方法1200はログ確率(LP: Log Probability)方法として知られている。加えて、ラベル採点の方法は、テキストの生成がどれだけ危険であるかをNLGモデルにより制御するリスクパラメータに基づきパラメータ化され得る。工程380を行う単独ラベル採点方法は例えば、リスクパラメータの制御によるハイリスク生成、中リスク生成、又は低リスク生成でもって操作され得る。従って、規定された方法は増されそして並列に操作され得る。パラメータ化され得る4つの方法は12まで拡張され得る。この理由のため、図3の工程380はラベル採点方法のアプリケーションを規定する。複数のラベル採点方法が同じ入力に関して操作され得、そして、結果のベクトル(ラベル採点方法の2つ以上の結果を提供する)が取得され得る。従って、ラベル採点サービス168は通常、本明細書において説明されるような複数のラベル採点方法の結果のベクトルである。NC、SL、SS及びLPラベル採点方法の各々は、ラベルの出力確率、判断されたクラスのインジケーション、結果が未決であるというインジケーション、結果が未決である理由の説明(例えば生成方法失敗、拡張失敗、弱過ぎる拡張、ラベル採点方法失敗、クラス分離の欠如、又は妥当でない閾値)を提供する。

【0075】

セットアップパラメータは、ラベル採点サービス168内に格納された利用可能モデルのうちのどれだけのモデルがラベル採点のために工程380において採用されることになるかをラベル採点サービス168から望まれるそれらの方法を選択することにより判断する。一実施形態では、セットアップパラメータはラベル及び／又はアンチラベルの特性に基づき判断される。モデル選択工程では、様々なラベル予測モードが選択される。一実施形態では、単独モードは、NC方法又はモードなどのデータ(ラベル付きデータのトレーニング／検証セット)を扱う分類システムにより使用されるデフォルト又は標準モードである。しかし、本方法は、それぞれが様々な条件に関して利点を有する他の採点システムも同様に行う。

【0076】

図3の工程380では、選択されたモードの結果が生成される。例えば、NC、SL、SS及びLP方法のすべてのロードされたモードが行われると、複合出力が次のベクトル出力を含む可能性がある: [NC: (サービスハーモニー: 0, 信頼度: 0.55), SL: (サービスハーモニー: 1, 信頼度: 0.8), SS: (サービスハーモニー: 1, 信頼度: 0.9), LP: (サービスハーモニー: 0, 信頼度: 0.6)]

【0077】

工程385において、ラベル採点サービス142は、性能、記録推定、類似度重み付け及びクラスラベルを既知性能のライブラリ内へ蓄積する。ラベル採点サービス168は、ベクトル及び類似度アルゴリズムのレポジトリを利用し、そして現在のスコアのラベルがライブラリ内で利用可能なラベル付け方法に似ているかどうかを判断する。ラベル採点サービスはまた、NLPベクトル化及びベクトル化機能内で利用可能な埋め込みアルゴリズムのレポジトリを利用する。

【0078】

工程390において、重み付けが利用可能ならば適用される。本方法は所与のモデル／アルゴリズムのために利用可能な2つ以上の採点方法を有し得る。また、このようなケー

スでは、様々な機構からの予測は変動し得る、又は少なくとも関連確率が変動する可能性がある。このような場合、本方法は予測と関連確率とを調停する必要がある。デフォルトとして、サブシステムが存在しなければ、工程 390 において、利用可能な重み付けは存在しない。本方法は、結果及びラベル確率推定を判断するためにデフォルト重み付け又は投票判断基準を使用する。追加情報が利用可能である場合、本方法は重み付けを出力評価に取り込む。重み付け結果を適用する例は事前ラベルが分類器のライブラリ内に発見されるケースを含み、このケースは「SS 検索及び SL 検索が、正しい結果を生じる可能性が他の利用可能分類器の 2 倍高い」ということを示し、したがって重み付け応答は $(2 * S + 2 * SL + LP + NC) / 6 =$ サービスハーモニーの重み付け尤度 (70%) である。

10

【0079】

工程 395 において、使用可能結果がラベル採点サービスにより取得されたということを性能条件が示す場合、推定に基づき 1 又は複数の結果が描画される。次に、本方法は工程 395 へ進み、ここで結果は確率推定に基づき描画される。この時点で、ラベル付けサービス 142 は、使用可能結果が利用可能な場合図 2 のグラフィックディスプレイ 200 内に表示されるすべての結果を戻し、そして、本方法は工程 397 へ進み、ここで新しい入力が見込まれる。新しい入力が見込まれる場合、例えばユーザからの追加入力が表示される。

【0080】

ラベル採点の NC 方法が図 4 の 400 内に示される。本方法は工程 315 から工程 410 へ進み、ここで生成モデル入力のいくつかの例がフォーマット化される。このラベル予測及び採点システムは、その出力の観点では任意の標準的 2 進法／多項式分類システムのように振る舞う。本システムの出力は、ブール／多項式クラス指示子（ブールクラス指示子に関して、ポジティブクラスは 1 として示されネガティブクラスは 0 として示される）及び関連確率／尤度である。このシステムに関して、本システムは、ブール／多項式インデックス付きクラスを有するいくつかの任意のポジティブ及びネガティブ文章と、次にモデルがその関連「トークン確率」と共に同様なブール／多項式クラスラベルを生成すると期待される入力文章とにより zero-shot モードにおける生成モデルを探索する。関連トークン確率は、予測確率／尤度として使用される歴史的モデル特定範囲パラメータにより正規化／スケールされる。追加照査が、トークン確率とラベルとを整合しそして出力を生成する前に、「生成されたテキストが必要クラスラベルを含む」ということを保証するために行われる。これらの照査が失敗した場合、この採点機構が方法 300 の工程 390 における最終予測重み付け機構から抜け出したということを示す「NONE」出力が送信される。工程 380 におけるラベル採点の NC 方法は、ラベルクラスを表わすための数とアンチラベルクラスを表わす別の数とを使用する。従って、2 進法ケースに関して、ラベル = 1、アンチラベル = 0。例示的实施形態では、NLG モデルが zero-shot モードにおいて使用される。例えば、モデルプロンプトが、文章クラスセパレータを使用することによりいくつかの例とそれぞれのラベルとを組み合わせることにより用意される可能性がある。別個の例は文章マスクブレイク (sentence mask break) を使用する。次に、プロンプトは別の文章マスクにより継続され、「文章クラスセパレータ」が続く、「予測開始」プロンプトが続く。1 つのポジティブ生成例及び 1 つのネガティブ生成例を有するケースに関して、プロンプトは次のとおりであり得る：「「ポジティブ例」「文章クラスセパレータ」1「文章区切り」「ネガティブ例」「文章クラスセパレータ」0「文章区切り」「候補テキスト」」。

20

30

40

【0081】

工程 420 において、プロンプトが GPT-3 などの生成モデルへ適用される。生成されたテキスト及びトークンの「ログ確率」が、生成されたテキスト内のトークン毎に受信される。工程 430 では、生成出力は数字「1」及び「0」に関して検索される。これらの数字ラベルのいずれも存在しなければ、本方法は失敗し、そしてエラー応答がラベル付けサービス 142 へ戻される。数字が存在すれば、トークン確率が工程 430 において生

50

成出力から判断される。例えば、2進法ケースに関して、生成出力は数値ラベル1又は0に関して検索される。発見されたシンボル1又は0のトークン確率はラベル確率の推定を判断するために使用される。トークン確率は、必要であれば組み合わせられ、そして予測確率として使用されるために正規化される。加えて、採点方法はそれ自身の閾値を適用し得、そして候補テキストがラベルに属するかどうかを判断する。NCラベル採点サービスの結果は、例えばラベル付けサービス142へ戻されると格納される。

【0082】

ラベル採点のSL方法が図16の1600に示される。本方法は工程315から工程1610へ進み、ここで生成モデル入力のいくつかの例がフォーマット化される。SL方法は通常、NC方法の操作を行う。しかし、プロンプトが発生するやり方には差がある。基本的差は、数値ラベルよりむしろテキストラベルが使用されるということである。複数のワードにより表わされる任意の概念はNLPシステムが理解するのが困難である。この理由のため、本方法は、ラベル生成のための重要なキーワードベース概念の順番を生成する。これらのラベルがモデルを促すために使用されれば、いかなる任意の概念も直接使用される可能性がある。出力は同様に、ユーザへ提示するための元の「任意の概念」又は用語へ後でマッピングされる促された概念を含む。予測確率は、生成出力内のキーワード又はキーワードの同意語を発見することによりそして出力内のキーワード又は同意語のトークン確率から判断することにより工程1630において生成／計算される。SLモードに関して、数値ラベルを使用することの代わりに、プロンプトが、それぞれのラベルを組み合わせることにより使用される。例えば、ラベルクラスが優先キーワードリスト「サービスハーモニー」を有し、アンチラベルクラスが優先キーワードリスト「ディスハーモニー」を有し、そして本方法が1つのポジティブ例及び1つのネガティブ例を処理すれば、プロンプトは次のとおりである可能性がある：「「あなたのために何かできることがあれば私に知らせてください。喜んでお助けします」「文章クラスセパレータ」「サービスハーモニー」「文章区切り」「これはあなたの問題であり私の問題ではない」「文章クラスセパレータ」「ディスサービス ディスハーモニー」「文章区切り」「候補テキスト」」。

【0083】

工程1620では、プロンプトはGPT-3などの生成モデルへ適用される。トークンの生成されたテキスト及び「ログ確率」が、生成されたテキスト内のトークン毎に受信される。工程1630では、生成出力はラベルのキーワード及びアンチラベルのキーワード（例えば「サービス」「ハーモニー」「ディスサービス」「ディスハーモニー」、又はこれらの同意語）に関して検索される。これらのキーワード又はそれらの同意語のいずれも存在しなければ、本方法は失敗し、そしてエラー応答がラベル付けサービス142へ戻される。キーワード又は同意語のうちの1つが存在すれば、トークン確率は工程1630において生成出力から判断される。例えば、このケースに関して、生成出力はラベル「サービス」及び「ハーモニー」に関して検索される。次に、発見された「サービス」及び「ハーモニー」のトークン確率はラベル確率の推定を判断するために使用される。トークン確率は、必要であれば組み合わせられ、そして予測確率として使用するために正規化される。加えて、採点方法はそれ自身の閾値を適用し得、そして候補テキストがラベルに属するかどうかを判断する。ストリングNCラベル採点サービスの結果は、例えばラベル付けサービス142へ戻されると格納される。工程1630では、ラベル採点サービス168のSL方法は、候補ラベル又はアンチラベルからの用語の出力され生成されたテキスト、又は生成された文脈内の非常に似た意味／埋め込みを有する用語の出力された生成テキストを検索する。そうでなければ、SL方法はNC方法が行うような操作を行う。

【0084】

ラベル採点のSS方法は図5の方法500において示される。SSラベル採点方法は、工程315から図5に示す工程380におけるSSラベル採点方法の実施形態へ進む。この方法は、入力ラベル内の概念と似ているとして選択された似たテキスト及び似ていないテキスト／好適にはアンチテキストの1又は複数のサンプルを使用し得る。本システムは入力テキストと共にサンプルを専用検索ランク付けサブシステム／モデル（様々な文章／

10

20

30

40

50

テキストの検索ランクを提供する)へ送信する。回収された検索サンプル及び検索ランク付けに基づき、本方法は入力テキストのラベル及びラベルの確率を判断する。検索スコアは、予測確率へ変換されるために追加処理を必要とする。ほとんどの検索サブシステムに関して、スケーリング又は正規化された検索ランク／スコア範囲が尤度の代用として使用される可能性がある。1つの特別な考慮は以下のことである：似た／似ていないテキスト生成のための追加サブシステムは、概念及び検索クエリーに対処する機構であり、従って伝統的分類システムと同類ではなく、従って、その生来の処理されていない／フィルタリングされていない形式でこれらのシステムから生成／回収されたデータは分類器をトレーニングするために直接使用され得ない。

【0085】

工程510では、ラベルの例の集合内の各例及びアンチラベルの例の集合内の各例は、NLGモデルを使用することによりそして出力結果に関連ラベルをタグ付けすることによりテキストを生成するために使用される。従って、ラベルの2つの例(EX-L1及びEX-L2と表された)が存在するケースを考察する。更に、アンチラベルの2つの例(EX-A1及びEX-A2と表された)がある。次に、例示的入力からの結果は、例示的名前に指示子「GR-」を前に付けることにより表される。従って、工程510において、生成モデルをEX-L1へ適用することでGR-EX-L1を生成する。生成モデルをEX-L2へ適用することでGR-EX-L2を生成する。生成モデルをEX-A1へ適用することでGR-EX-A1を生成する。生成モデルをEX-A2へ適用することでGR-EX-A2を生成する。

【0086】

工程520では、生成モデルは対応生成出力(GR-CTと表された)を取得するために候補テキスト(CTと表された)へ適用される。本方法は、工程510において生成された生成例により生成された文書集合から候補生成されたテキスト(GR-CT)の検索スコアを計算するために工程530へ進む。一般的に、SS方法のアイデアは、検索エンジンにおけるクエリーとして候補の生成出力(GR-CT)を使用することと、「ラベル例(GR-EX-L1及びGR-EX-L2)から生成された結果がアンチラベル例(GR-EX-A1及びGR-EX-A2)から生成された結果よりクエリー(GR-CT)により近いかどうか」を判断するためのメトリックとしての検索結果ランクを測定することである。一般的に、クエリーと結果との間の意味論的距離を測定するトレーニング型構造的／意味論的類似度モデル検索エンジン(例えばMicrosoft(登録商標)DSSM)が望ましい。代替的に、GPT-3検索ランクが使用される可能性がある。

【0087】

一実施形態は、工程540においてラベル確率推定を判断するために検索ランクに関する調停規則又は文書集合全体にわたる異なるやり方でラベル付けされた文書のスコアを使用する。第1の調停規則は、最良検索ランク(最も高い検索スコア)を有する文書のラベル及び検索ランク／スコアを使用することである。第2の調停規則は、ラベル例から生成されたすべて文書のグループの発見的な検索スコアの発見的なグループ(平均などの)を判断し、そしてこれとアンチラベル例から生成されたすべて文書の他のグループの発見的な検索スコアとを比較することである。第3の調停規則は、検索スコア又はランクに基づき候補文書をショートリストに載せ、そして次に第2の規則をこのショートリストに対し行うことである。例えば、最高位ランク～最低位ランク(GR-EX-L1, GR-EX-A1, GR-EX-A2, GR-EX-L2)を仮定する。ラベルクラスは規則1下で選択されるだろう。次に以下のことを仮定する：検索エンジンの検索スコアは、関連スコア(GR-EX-L1=0.5, GR-EX-A1=0.3, GR-EX-A2=0.21, GR-EX-L2=0.5)を生じる意味論的空間内の文書間のコサイン類似度である。第2の規則下で、ラベルがまた選択されるだろう。しかし、同じ検索スコアにより、調停規則3は0.08の検索スコア閾値が使用されればアンチラベルを選択するだろう。次に、ラベルの確率推定は検索スコアを正規化することにより形成される。次に、方法500は図31の工程380の出力に戻る。

10

20

30

40

50

【0088】

このラベル予測及び採点方法は、NLP検索及びランク付け能力を有するモデルを必要とする。これらは純粋なNLP生成モデル又は他のSOTA検索ランク付けモデルであり得る。加えて、このシステムは、事前（ユーザにより提供された又は使用ケースに固有な）トレーニングデータ無しに特別要件に基づきテキストを生成／回収する可能性があるテキスト生成又は回収サブシステムを必要とする。一実施形態では、これは規則ベースウェブ検索回収システムである可能性がある。固有検索判断基準／概念（しばしば繰り返される）に関して、これらの要件は高められ、そして人間作成候補検索ランクテキストと置換される可能性がある。

【0089】

ラベル採点のLP方法が図12の方法1200に示される。LPラベル採点方法は、工程315から図12に示す工程380のLPラベル採点方法の実施形態へ進む。LP方法はまた、工程380を行うデュアルパス生成ログ確率ベースラベル採点方法（dual-pass generative Log Probability based label scoring method）として知られている。このシステムでは、一実施形態におけるNCクラスインデックス又はSLラベル採点のいずれかの機構がサブ工程を支援するものとして使用される。LP方法では、本システムにラベルを生成するように頼む代わりに、本方法は、各あり得るクラスインデックス又は（ストリング）クラスラベルの入力テキストを複製し、そしてシステムに次のテキストを生成するように頼む。生成されたテキストは直接使用されないかもしれないが、様々なインデックス／ラベルのための提出されたクラスインデックス／ラベルのトークンログ確率が使用され、そして、本方法は、これらのログ確率を1へ再スケーリングするsoftmax関数を適用した後に最も高いログ確率を有するものを選択する。

【0090】

方法1200に示す工程380を行うラベル採点方法に入った後、3つの経路が並列に動作する。工程1210では、方法1200は、候補テキストを有する入力としてポジティブ例を取る（例えばテキスト例とラベルタイプとを組み合わせる文接続技術を使用することにより）。工程1215では、ラベルのログ確率がこの入力から判断される。工程1220では、方法1200は、候補テキストを有する入力としてネガティブ例を取る（例えばテキスト例とアンチラベルタイプとを組み合わせる文接続技術を使用することにより）。本方法は工程1225へ進み、ここでアンチラベルのログ確率がこの入力から判断される。同様に、工程1230では、次のテキストは例と候補テキストとのすべての組み合わせにより予測される（例えばテキスト例とラベルタイプとを組み合わせる文接続技術を使用することにより）。本方法は工程1235へ進み、ここでキー用語／トークンのログ確率が導出され、そして閾値インジケーションとして使用される。本方法は工程1240へ進み、ここで「取得された閾値が、ラベルに関連する候補テキストのログ確率間の分離がアンチラベルに関連する候補テキストのログ確率から十分に分離されているということを保証するかどうか」を見るための試験が行われる。閾値が妥当でなければ、本方法は工程1245へ進み、ここでエラー信号が生成される。そうでなければ、本方法は工程1250へ進み、ここでポジティブ確率及びネガティブ確率が予測確率を生成するようにスケーリングされ、そしてより高いスコアを有するクラスに有利である予測が生成される。

【0091】

本開示は、最先端NLPモデルのデータ拡張のためのシステム及び方法を説明する。ここで「最先端」NLPモデルは通常、文章の文脈により焦点を当てることを学習したNLPモデルのクラスであってそして多くのデータから多くの豊かな表現を学習するために十分に複雑であるNLPモデルのクラスを指す。このようなモデルのいくつかの例はトランスフォーマーアーキテクチャ上で作られた転移学習ベースモデル（例えばBERT、TURING、GPT3など）である。

【0092】

伝統的データ拡張技術（1又は複数の手法を使用することにより既存トレーニングデータから擾乱（perturbation）を生成することに基づく）は不十分だった。失敗した企て

10

20

30

40

50

の例は、テキストを1つの言語から別の（恐らく幾つかの）言語へ翻訳しそして次に元の言語に戻る逆翻訳（Back Translation）を含む。このような変換は、恐らくワードの若干異なる選択であるが同じ意味を伝える選択により同じテキストの別の表現を生成することになる。そのままでは不十分な別の企てはEasy Data Augmentation（EDA）である。これらは、同意語置換、ランダム挿入／削除／スワップのような方法を使用することによりテキスト内の特定ワード／用語を変更するために組み合わせで適用される一組の簡単な技術である。これらはまた、文章の同じアイデアを保持し、そしていくつかのワードだけを修正する。再び、NLPシーケンス／Tone Alterationを行うだけでは不十分である。この方法では、文章内のワードの順番／シーケンスが変更される。これは、ランダムであり得る、又は或る単純論理（第1の人から第3の人へ）を伴い得るが当該アイデアを変更しない。単なる埋め込みベースワード／用語変更を使用することもまた不十分だった：これらの技術は、GloVe、Word2VecなどのようなNLPベクトル化モデルからのワード埋め込みを使用し、そして次に、文章内のいくつかのワードを変更するためにいくつかのワードのベクトルの近い／似た表現（又は反意語のための逆ベクトル）を選ぶ。

10

20

30

40

50

【0093】

伝統的データ拡張技術は現代の最先端NLPモデルの助けにはならない。これらの伝統的技術は、以下の理由のためにこのような最先端モデルのトレーニングデータを拡張するためには好適ではない。第1に、文脈内のモデル（例えばBERT、Turingなど）は、トランスフォーマ学習に基づいており、そしてそれらの事前学習段階から同じワードの様々な定式化を既に知っているのので、いくつかのワードがそれらの同意語又は似た埋め込み用語により変更される場合、新しい何かを学習するためにこれらのモデルの最小限の新しいアイデアを生成する。第2に、これらの大規模／最先端NLPモデルはたいていは、それらのほとんどが、マスクされた用語を予測することを事前トレーニング中に学習するのでランダム挿入及び削除ベース擾乱に対して免疫性がある。第3に、これらのモデルのほとんどは、多言語対応であり、従って同質ベクトル空間内の複数言語のベクトル表現に対し作用し；従って翻訳ベースアイデアに対し免疫性がある。第4に、これらの大規模モデルは、文脈意識的であり、従って非文脈意識的変換（例えば「lay-egg」内の文脈内の「lay」を「lye」の同意語により変更すること）はこれらのモデルの性能を悪化さえし得る。第5に、これらのモデルは、何十億ものトレーニング可能パラメータを有し、従ってトレーニングデータの「豊かな」コーパスを必要とする。ここで「豊かな」は、同じラベルクラス（「文脈」）内のアイデアにおける量及び著しい多様性の両方により修飾される。上記不十分な技術だけでは、大量のトレーニングデータを生成することができなく、そしてまた、様々なアイデア（同じラベルクラス内の）を有するデータを生成する際に惨めに失敗する可能性がある。第6に、同様なアイデアの偏りの問題が存在する。これらの大規模モデルは文章内のアイデアの表現から学習し、同じアイデアが複数回繰り返されれば（伝統的拡張技術を使用することにより）、モデルは、当該アイデアに過剰適合し、そして同じ文脈内の異なるアイデアを有するテキストに対してもよく働く可能性が高い。第7に、大量のトレーニングデータが効果的学習のために必要とされる。伝統的NLPモデルがそれらの学習欲求を飽和させるための数千のトレーニングサンプルを必要とした場合、最先端NLPモデルは、文脈の基礎をなす様々なアイデアの様々なスタイルの表現を学習するために所与の文脈全体にわたり何百万ものラベル付きデータを必要とする。従って、データの手動検索及び等級付けはそれらに関して非常に高価であり得る。

【0094】

最先端NLPモデルのデータ拡張要件を緩和するための他の代替Data Scarce Approachも同様に不十分である。豊か及び最先端大規模NLPモデルを効果的にトレーニングするために文章内のより豊かな「文脈」及び「アイデア」を必要とするモデルをトレーニングするための大規模「拡張」データを得ることは挑戦的である。従って、これらのモデルのデータ拡張挑戦を緩和するために使用される現在の手法は以下のものである。

【0095】

第1に、Non-Scalable and Costly手法は不十分だった。これは第1に：Manual Data Source Scavenging and Gradingを含む。これは、最先端モデルをトレーニングすること取得する（正確には拡張しない）ための最も有力な手法である。文脈要件（ラベルクラス仕様）に基づくこの第1の不十分な方法では、データのいくつかの多様な源が取得され、次に、これらの源の各サンプルが手動で又はクラウドソーシングを介し等級付けされる。第2に、スケーリング可能であるが余り効果的でない手法もまた例えばFew-Shot Classificationとして不十分である。この手法では、NLPモデル（たいていはトランスフォーマベースモデル）は、ラベル無しの「ウェブ」又は「企業」データの大きなコーパスに対し事前トレーニングされる。これは「より豊かな」文脈及び「アイデア」（上述のような合成的伝統的拡張技術）を有する実際の人間生成データに関する学習の適合性を提供する。しかし、このようなデータはラベル付けされない。しかし、ラベル付きデータのいくつかのサンプルによってすら、このようなモデルは、同じトレーニングデータから生成された拡張と組み合わせられた同じトレーニングデータに関しトレーニングされた伝統的モデルよりはるかにうまく機能するということが分かった。別のスケーリング可能であるが不十分な技術は単なるzero-shot学習である。この手法では、非常に大きな（何十億ものパラメータ、例えばGPT-3）NLPモデルは更に大きなラベル無しトレーニングデータ上にテキストを生成する（テキストを分類することとは反対に）ようにトレーニングされる。以下のことが仮定される：利用可能ないくつかのトレーニングサンプルがテキストを生成するためのプロンプトとして使用される場合、モデルはpseudo-NLP-classificationモデルとして役立ち、従って、大きなラベル付きトレーニングデータによりトレーニングする必要性を軽減する可能性が高い。

10

20

【0096】

図2に戻ると、グラフィックディスプレイ200はまたグラフィックコントロール293、294、295を含む。これらのコントロールは例えばラベル付けサービス142により使用又は生成されるデータ項目全体にわたって一組の操作を行う際にユーザを支援するために使用され得る。このようなコントロールは、ラベル付け基準、文書コーパス、基準変更ログ、ラベル付け性能ログ、ラベル付けインデックス及びラベル付けインデクサーなどの電子アイテムのために使用され得る。電子アイテムは通常、格納され、回収され、修正され、そしてストレージ180又はサーバ177のメモリを使用することによりラベル付けサービス142により表示される。本明細書において使用される「ラベル付け基準」は通常、データ項目の集合を指す。データ項目の集合は併せて、ラベル付けサービス142がモデルに基づき判断（ラベルが新しい候補に正しく属するか否かを判断する判断）を提供することを可能にする。「文書コーパス」は通常、新しい候補がラベル付け基準に影響を与える判断を下すために引き出される一組の文書である。「基準変更ログ」は通常、ラベル付け基準に対するデータ項目追加及び削除の記録である。「ラベル付け性能ログ」は通常、ラベル付け基準に関係する事象の記録であり、これは、拒絶の頻度、手動で追加される例の平均信頼度、最近追加された候補の平均信頼度、拒絶された候補の平均信頼度、これらの統計値のうちの1つの統計値の標準偏差、又は当該ラベルが管理され確認された一組の管理文書に対するラベル付け基準の成功率などの不満足を示す可能性がある。いくつかの例が手動で追加されると、ラベル付けサービス142は、ラベル付け基準の精度の推定値を得るために、エントリを追加する前にエントリに対しラベル付け基準を実行し得、そしてこれらの推定を最近追加された候補の平均信頼度へ取り込み得る。「ラベル付けインデックス」は通常、ラベルが正しく適用される文書コーパスの一部分を示す記録である。「ラベル付けインデクサー」は通常、文書コーパスのラベル付けインデックスを構築しそしてコーパス内のどの文書がラベル付けのために走査されたかを追跡するアプリケーション機能を指す。

30

40

【0097】

グラフィックコントロール293は、選択されると、ユーザがコンテンツ管理に関係す

50

る操作を行うこと（例えば、ラベル付け基準を保存する、ラベル付け基準をロードする、ラベル付け基準を保存する、ラベル付け基準に関連するコーパスを定義する、ラベル付け基準の論理的組み合わせを定義する、ラベル付け基準を閉じる、新しいラベル付け基準を開く、最近使用されたラベル付け基準をロードする等々）を可能にするドロップダウンメニューを提供する。「ラベル付け基準の論理的組み合わせを定義する」機能は、2つ以上の規定ラベル付け基準が第3のラベル付け基準を形成するために論理的に組み合わせられることを可能にする。例えば、劣悪顧客サービスを定義する3つのラベル付け基準は、これらのラベルのうちの少なくとも1つを有した通信の一部を識別するためにOR機能を介し論理的に組み合わせられる可能性がある。別の例として、映画データベース内の4つの特定プロット要素を捜す人は、各プロット要素のラベル規則を生成し、そして次にプロット要素の各ペアの論理的組み合わせ機能（ペアの6つの論理的組み合わせの結合に関係するラベル付け基準を定義する組み合わせられた規則を生成する機能）を介しプロット要素のうちの少なくとも2つを含むプロットを発見する論理的規則を生成する可能性がある。

10

【0098】

グラフィックコントロール294は選択されると通常、ユーザが開発、運用、解析に係る操作を行いそしてロードされたラベル付け基準の以下の履歴を使用することを可能にするドロップダウンメニューを提供する：変更ログを見る、性能ログを見る、コーパスをラベル付け基準によりインデックス付けする、ラベル付け基準を手動で拡張する、新しい例をインポートする、インデックス粒度を設定する、ラベル閾値を設定する、ラベル付け基準の例を拡張する、ラベル付け基準のアンチラベルを拡張する、ラベル付け基準のラベルを拡張する、すべての部品を拡張する等々。「ラベル拡張の手動モード」は、表示エリア203内の235などの空のグラフィックコントロールを提示するためにコンテンツをクリアすることによりグラフィックディスプレイ200により提供され得る。ユーザがテキスト入力を完了した後、新しいテキストが、確認された状態により設定されたポジティブ例へ追加される。代替的に、ラベル拡張の手動モードの選択が、文書コーパスの上で動作する伝統的キーワードインデックス検索エンジンであるがランク付けされた戻り結果内の各テキストスニペットに隣接した制御を提供する伝統的キーワードインデックス検索エンジンを提供し得る。ユーザがポジティブ例又はネガティブ例を示すための制御を選択すると、テキストスニペットが適切な指定と共にラベル付け基準へ追加される。「サンプルをインポートする」機能は、ポジティブ及びネガティブとしてマーキングされた例を含みそしてデータセットをラベル付け基準へ取り込む既に定義されたデータセットを取り得る。例えば、手動検索又はエントリを行ったユーザは、電子メールをアタッチメント（恐らくいかなるラベルの定義も有しないがラベル付け基準構造内に格納されたそれらの例を含むアタッチメント）と共に送信し得る。ラベル付け基準ファイルがローカルに保存されると、ラベル付け基準ファイルは、いくつかの例を別のラベル付け基準へインポートするために任意のファイルブラウザにより選択され得る。「インデックス粒度を設定する」機能は、文章、パラグラフ、いくつかのワード、又は文書などの候補テキストを形成する分量を定義する。「インデックス粒度を設定する」機能はまた、ユーザが「どれくらい正確にポジティブラベルインジケーションの場所が記録されることになるか」を定義することを可能にする。例えば、文書レベル精度は、文書がラベルに関する試験で陽性であるということを記録するだろうが、文書当たり1つのインジケーションだけが記録されることになる。「ラベル付け基準の例を拡張する」機能は通常、ラベルの文脈内の現在の例の豊さを反映する利用可能例のコンピュータ実施型拡張を提供する機能を行う。「ラベル付け基準内のアンチラベルを拡張する」機能は例示的拡張機能のように動作するが、例を単に追加する代わりに、代替アンチラベルキーワード構造が、追加例に加えて又は追加例を追加する代わりにアンチラベルエリア210へ追加される。「ラベル付け基準のラベルを拡張する」機能は例示的拡張機能のように動作するが、例を単に追加する代わりに、代替ラベルキーワード構造が、追加例に加えて又は追加例を追加する代わりにラベル定義表示エリア201へ追加される。一実施形態では、一組のラベルキーワード構造が、発見されたキーワードの代替ラベルセットを提供するためにアンチラベル表示エリア210のような表示

20

30

40

50

エリアにおいてユーザへ提示される。

【0099】

グラフィックコントロール295は通常、ラベル付けサービス操作のうちの1つがユーザのために行われることを可能にする機能活性化コントロールである。グラフィックコントロール295を選択することにより、この機能は即座に行われる。本明細書において更に説明される一実施形態では、グラフィックコントロール295は「ラベル付け基準の例を拡張する」機能に割り当てられる。ユーザは、同僚により手動で入力された10個のポジティブ例及び10個のネガティブ例の新しい組を受信し、そして新しい例をグラフィックディスプレイ200内に示されるラベル付け基準へインポートしたならばこのようなコントロールを選択する可能性がある。別の理由は「ユーザが規則を適用するためのコーパス定義を変更した」ということである可能性があるので、蓄積された例が新しいコーパスの文脈内の分類された例を拡張するために使用され得る。例えば、ユーザは最初に、文書コーパスが、高水準の顧客サービスを有する可能性が高い「販売電子メール」として定義した。文書コーパスを「技術サポート」コーパスへ変更すると、ユーザは、異なり及びより豊かな例を発見する可能性が高く、そしてよりバランスのとれた一組のネガティブ例を利用することができる。入力ラベル付け基準内のサンプルの数は極めて小さいので（最先端大規模トランスフォーマベースNLPモデルの要件と比較して）、他のモデルは、少数のデータサンプルにより効果的にトレーニングされることができない可能性が高い。これらの大規模最先端モデルは文脈要件を全体論的に表わすのに必要とされるアイデアの偏りの観点で非常に豊かな多様な（全体論的に文脈要件を表わすのに必要なアイデアにおける多様性の観点で）トレーニングデータを必要とする。小さなデータセットだけによりこのような豊さを表わすことは通常実現可能ではない。しばしば、より小さなラベル付け基準データセットでは、このデータセットから全体論的に必要文脈表現を効果的に学習するためのモデルのデータの十分な多様性及び豊さは無い。

【0100】

拡張機能の結果として、開示された方法600は、両方のクラスにわたる十分なデータ（豊かなデータ）によりそしてうまく選択されたコーパス（人間生成された）のおかげでこのデータセットを拡張する。従って、結果として拡張されたデータは現実のシナリオを表し、雑音耐性がある。従って、拡張の結果は、ラベル付け基準から実行される改善された安定性及び関連性モデルである。本明細書において開示される方法は、最先端大規模NLPモデルをトレーニングするために好適なラベル付け基準データセットを生成することができる。

【0101】

本システムは、極めて小さいデータセットを非常に豊かな多様性により拡張する。出力データセットは、シソーラスが提供するであろうような個々のワードのより豊かな表現だけではなくまた文脈要件に関する新しいアイデアでもって豊かでもある。出力データセットは文脈要件意識的やり方で人／企業生成データを発見する。本明細書において開示される拡張の方法は、ワード／用語／翻訳／生成などをランダムに単に置換するだけでなく、ラベル記述により提供される特定文脈要件に関する新しいアイデアを全体論的に発見する。提示される拡張の方法は雑音耐性的やり方で働く。生成された拡張データセットは、大規模及び最先端NLPモデルをトレーニングするために直接使用される可能性がある。加えて、分類する既存（無ラベル）データセットを必要とするzero-shot/few-shot分類技術とは対照的に、開示された方法は拡張要件及び事前分類要件の両方を満たす。開示される拡張方法は、任意の分類モデルの準備ができて正しいデータサブセット内のデータサンプルを自動的及びインテリジェントに取得し汲み出す。

【0102】

ユーザが拡張機能呼び出すためにグラフィックコントロール295を選択した後、ラベル付けサービス142は、アプリケーション110から制御信号を受信し、そしてそれに応じて、拡張方法600に参与する拡張操作を行う。一般的に、特定文脈要件を表すいくつかのポジティブ例及びいくつかのネガティブ例が方法600により受信される。ラベ

ル付け基準における例及びラベルは受信された例及びラベルに基づき一組の例を拡張する拡張操作を行うために方法600により受信される。グラフィックコントロール295の選択により呼び出された拡張機能の出力は通常、多様性に富むより多くのポジティブ及びネガティブクラス固有データサンプルを有するより大きな文脈要件意識データセットを有する改善されたラベル付け基準である。即ち、一組の例は、これらのアイデアが入力サンプルの極めて小さい組の中に存在しない場合ですら必要な文脈に関する多様なアイデアを有する。加えて、生成されるサンプルは非合成的である：すなわち、生成されるサンプルは生成モデルを使用することによりストリングの単なるスポット擾乱により生成されない。このデータセットは、現在の手動取得及び等級付けが必要とされる大量の豊かなデータを必要とする最先端大規模NLPモデルをトレーニングするために理想的に適している。

10

【0103】

本拡張方法は通常、ラベル付け基準内の一組の現在定義されている例などの一組の例を受信する（例えばストレージサービス180からラベル付け基準を受信することにより）。次に、本拡張方法は一組の例をループし、一度に1つの例と関連ラベルとを取る。一実施形態では、選択されたラベルはネガティブ例に関連付けられたアンチラベル又はポジティブ例に関連するラベルである。複数の利用可能ラベルが存在する場合（例えば利用可能なくつかのアンチラベルが存在する場合）、ラベルと例との複数の組み合わせが使用され得る。別の実施形態では、ラベルは、同じクラスの一組の利用可能ラベルからランダムに選択される。

【0104】

20

例とラベルとが選択されると、方法600は工程603において拡張方法を開始する。工程605では、方法600は現在の例から候補テキストを受信する。例示的ケースでは、グラフィックコントロール202内に示される以前に分類されたサンプル「あなたのsprocketの注文をお手伝いできて嬉しいです：I would be happy to help you with your sprocket order」がポジティブ例として分類され、従って方法600により受信される。工程610において、方法600は、グラフィックコントロール206内に示される入力ラベルに対応するグラフなどの入力ラベル又はリスト「サービスハーモニー」から成るグラフィックコントロール292内に示される順序付きリストを受信する。

【0105】

工程615において、一組の優先キーワードが用意される。この工程では、要約キーワードが抽出され、そしてそれぞれの強さが文脈意識的やり方で計算される。即ち、各優先キーワードの強さが計算される。この計算はラベル記述内の文脈要件を意識する。この文脈意識的な一組のキーワードはネガティブ例及びポジティブ例の両方に関して取得される。一般的に、記述的ラベルテキスト入力は複数の用語を含む生テキストストリングであり得、候補テキストは複数の用語を含む生テキストストリングである。ラベル付け基準は候補ラベルペアの優先キーワードを格納する可能性がある。このケースでは、優先候補ラベルキーワードは一組の優先キーワードを用意するためにストレージサービス180から方法600により受信される。代替的に、候補テキスト及び／又はラベルのキーワード要約構造がラベル付け基準内で利用可能であり得る。これらの構造は利用可能な場合はストレージサービス180から受信される。工程615を行う方法は工程703で始まり、そして工程705へ進む。ラベル構造がストレージ180から利用可能でなければ、ラベルが複数のワードを含むかどうかを判断するための試験が行われる。多くの文脈要件は単一用語で説明することが可能ではない。より複雑なラベルアイデアはアイデアの集合を必要とする。大規模最先端トランスフォーマベースモデルを使用する現代のNLPは、このようなデータをスマートに分類する可能性がある豊かなモデルを生成することに優れている。しかし、これらのモデルはまた、概念要件を作る様々なアイデアの多様な表現下で基本的概念を全体論的に学習するために豊かなトレーニングデータを必要とする。すべての基本的アイデアは、非常に小さな入力データサンプル内においてもラベル要件の単一用語においても全体論的に表現されることができないので、文脈要件は単一用語ラベル要件の代わりにラベル記述として表現される。

30

40

50

【0106】

ラベルが複数のワードを含めば、本方法は工程710へ進む。要約キーワード構造は、図9の工程710を行う方法により説明されるように入力ラベル記述から判断される。本方法は工程715へ戻り、ここで優先テキストキーワードを提供する候補テキスト構造が取得される。工程715の方法は、要約するための異なる入力テキスト（すなわち候補テキスト）により工程710の方法のように進む。例えば、候補テキスト「あなたのsprocket orderの注文をお手伝いできて嬉しいです：I would be happy to help you with your sprocket order」は、[helping, community-focus, happy, customer, sprocket]などの有意なキーワードのリストを判断し得る。優先度を有する優先キーワードの順序付きリストは[(helping, 0.35), (community-focus, 0.35), (happy, 0.2), (customer, 0.1)]である。helping頂点1112、community-focus頂点1114、happy頂点1116及びcustomer頂点1118を有する候補グラフ1110のイラストを示す結果グラフが図11の構造表示1100内に示される。タグディスプレイ1160は、社会的価値タグがhelping、community-focus及びserviceへ割り当てられたことを示す。peopleタグは顧客へ割り当てられた。feeling/sentimentタグはhappy及びハーモニーへ割り当てられた。示されたグラフ構造は、より豊かな用語を提供し、そしてまた順番だけでなく強さ及び類似度も含むより豊かな順番記述も提供する。以下のプロセスのためのより豊かなクエリー構築のためのタグ、リンケージ及び方向が利用可能である。工程710を行うキーワード要約方法が終わると、優先キーワードは[helping, community-focus, happy, customer]である。一実施形態では、異なるサイズ判断基準が候補テキストを要約するケースのために使用される。本方法は工程720へ進み、ここでテキストキーワードの優先用語のための埋め込みが取得される。工程725において、ラベルキーワードの優先用語の埋め込みベクトルが取得される。例えば、候補テキストの各用語には埋め込みベクトルが提供され、そしてリストは優先用語だけを保持するためにフィルタリングされる。工程730において、優先度ラベル用語と優先度候補キーワードとの間の類似度が取得される。これは類似度（例えばアンチラベル内の各優先用語の埋め込みベクトルと反転済みテキスト内の各優先用語とのコサイン類似度）を計算することにより取得される可能性がある。工程735において、優先テキスト用語の文脈重要度が計算される。一実施形態では、各要約キーワード用語の文脈重要度はラベル内の各用語間の類似度の正規化加重平均として計算され、ここで重み付けはラベル用語の重要度スコアである。工程740において、本方法は、文脈重要度及びキーワード優先度から文脈意識優先度を判断する。例えば、各要約キーワードの文脈意識優先度は文脈重要度とキーワード優先度との正規化積として計算され得る。本例では、文脈意識優先キーワードは「helping, happy, customer」である。文脈意識優先キーワードの計算は797において終了し、そして本方法は図6の工程620へ戻る。

【0107】

工程620において、一組のランク付けされた検索回収結果が取得される。コーパス154などのラベル付け文書コーパス全体にわたる検索サービス164は文脈意識キーワードをクエリーとして使用することにより行われる。トップランク検索戻りの数が検索サービス164から取得される。例えば、「文脈意識」用語が所与の検索エンジンに関する関連文書を検索するために使用される場合はトップn（例えばn=10）検索結果が回収される。ほとんどの検索エンジンはまた「回収された検索結果がクエリーに関連する」ということをなぜ信じるかということを示すテキストのスニペットを生成する。工程625では、本方法は、データベースを拡張するためにこれらのスニペットを収集する。一実施形態は検索エンジンのAPI版を使用する。一実施形態は、トップN検索結果の検索回収のクライアント版とそれぞれのスニペット抽出とを使用する（工程625において）。この工程への入力の例示的順序付けられた文脈意識キーワード用語は「helping, happy, customer」である。この入力クラス要件プロンプト（すなわち、ポジティブ文章が生成されるということとネガティブ文章が生成されるということとを保証するためのプロンプト）に基づき更に豊かにされ得る。例えば、グラフィックコントロール236は、発見

10

20

30

40

50

されたポジティブ例を確認するようにユーザに促し得る。グラフィックコントロール 2 5 2 におけるプロンプトは、ネガティブ例が発見されたことを確認するようにユーザに促し得る。

【0108】

工程 6 2 5 では、テキストスニペットが取得され、そして本方法は「テキストスニペットがラベルクラスに属する」という信頼度を定量化する工程 6 3 0 へ進む。クラス信頼度を定量化する例示的方法はテキストスニペットのキーワード構造を構築することである（例えば図 9 の工程 7 1 0 を行う方法を使用することにより）。テキストスニペットのキーワード構造とラベルキーワード構造との意味論的類似度全体を評価する例示的方法は、グラフ用語のベクトル化変換に基づくコサイン類似度の使用、又はベクトル化機能 1 5 6 により提供される他の方法である可能性がある。本明細書において開示される他の方法は、類似度スコア又はラベルがテキストスニペットへ正しく適用される確率の推定を提供する。確率が判断 6 3 5 において低過ぎれば、本方法は、ストレージサービス 1 8 0 内の失敗したスニペットを記録することにより 6 4 0 において失敗を示し、そして別のテキストスニペットを取得する工程 6 2 5 に戻り、この別のテキストスニペットは工程 6 3 0 において定量化されそして工程 6 3 5 において試験される。クラス信頼度が判断 6 3 5 において十分であると、本方法は工程 6 4 5 へ進み、ここで入力是一例として十分なスニペットを含むように拡張される。一実施形態では、工程 6 3 0 は、テキストスニペットを工程 3 0 5 における候補入力としてそしてラベルを工程 3 1 0 における候補ラベルとして使用することにより、ラベルが候補テキストへ正しく適用されるという信頼度を判断する方法 3 0 0 を使用する。次に、方法 3 0 0 の推定されたラベル確率の出力はクラス信頼度として使用される。工程 3 0 7 において、ラベルは既に知られているので、本方法は工程 3 7 2 に進む。工程 3 7 2 において、一実施形態では、アンチラベルはアンチラベルをメモリ内に格納したラベル付け基準からラベル付けサービス 1 4 2 により生成され、そして本方法は工程 3 1 5 に進む。

10

20

【0109】

工程 3 2 0 において、本方法は候補ラベルの一例が利用可能であると判断したので、本方法は工程 3 3 5 に進み、ここで候補ラベルの一例が受信される。一実施形態では、ポジティブラベルの K 個の例が利用可能ならば受信され、ここで K は非負整数である。工程 3 3 5 の一実施形態では、一例は、一組のポジティブ例からランダムに選択される。工程 3 3 5 の一実施形態では、ラベルの一組の最高信頼度例が、ポジティブ組内のトップ L 個の例のうちの K 個をランダムに選択するために使用される。工程 3 3 5 の一実施形態では、ポジティブ例を取得するために使用される一組の例は、同様な例の同じクラスに属する一組のポジティブ例となるように制限される。

30

【0110】

工程 3 2 5 では、本方法は候補アンチラベルの一例が利用可能であるということを判断するので、本方法は工程 3 4 0 に進み、ここで候補アンチラベルの一例が受信される。一実施形態では、ネガティブラベルの K 個の例が利用可能ならば受信され、ここで K は非負整数である。工程 3 4 0 の一実施形態では、アンチラベルの一組の最高信頼度例が、ネガティブ組内のトップ L 個の例のうちの K 個をランダムに選択するために使用される。工程 3 4 0 の一実施形態では、ネガティブ例を取得するために使用される一組の例は、同様な例の同じクラスに属する一組のネガティブ例となるように制限される。

40

【0111】

一実施形態では、K 及び／又は L は拡張方法 6 0 0 を制御するためにユーザにより設定されるパラメータである。一実施形態では、K 個のネガティブ例と K 個のポジティブ例とのバランスのとれた一組が利用可能ならば取得される。

【0112】

本方法は工程 3 1 5 から工程 3 8 0 へ進み、ここで 1 又は複数のラベル採点方法が適用される。工程 3 8 5 において、性能記録が蓄積され、そして現在のラベルと同様なラベルの利用可能重み付けが捜される。工程 3 9 0 において、重み付けが発見されればそれらが

50

適用され、そして重み付けラベルスコアが判断される、そうでなければラベルスコアは工程 380 において判断された一組のラベルスコアから判断され、そして工程 395 において、結果が推定に基づき描画される。方法 300 の事例では、描画された結果は、判断されたラベルスコアをラベルクラス信頼度として方法 600 へ提供され、工程 635 において試験される。次に、方法 300 は工程 397 に進み、ここで新しい入力が入力ユーザ又は拡張機能から待たれる。特に所定場所で行われない擾乱システムの拡張時の雑音は挑戦である。

【0113】

拡張時の雑音はまた、複雑なモデルをトレーニングするためのデータを拡張又は生成するように意図された AI ベース代替システムの挑戦である。わずかなデータが利用可能な場合にサンプルがあるクラスに属する一定の確率があったとしても、クラスに含まれるがクラスの良い代表ではないいくつかのサンプルもあることになる。これらのサンプルからの雑音は低減される必要がある。本システムは、小さなサンプルサイズに関して働く雑音低減方法を提供する。

【0114】

判断 635 において、クラス信頼度が十分な場合、本方法は工程 645 へ進み、ここで一組のポジティブ例が、ラベル付け基準内のポジティブ例としてテキストスニペットを格納することにより増加される。次に、本方法は、判断 650 においてラベル付けサービス 142 からの追加ユーザ入力又は追加入力を照査することに進み、そして追加ユーザ入力がない場合、本方法は判断 655 に進み、ここで試験は、発見された各新しい例がネガティブ例（発見されると新しいポジティブ例を補足する）により「バランスされる」又は補足されるべきかどうかを判断するために行われる。拡張されたデータの豊さ及び規模を増加するために、拡張機能は各例からのネガティブクラス拡張及びポジティブクラス拡張の両方をその元のクラスに関係なく生成する。従って、例えば、ポジティブクラスサンプルはまた、生成されたバランシングサブグラフが存在するということを保証するためにネガティブクラスサンプルに合成的に変換される。データが特定アイデアの表現のバランスを有するかに関するいくつかのラベル判断における利点がある。バランスのとれた例を生成する実施形態は、例えばシソーラスベース方法、反意語置換方法、ネガティブベクトルベース埋め込み方法などを含む可能性がある。

【0115】

判断 655 を行うための判断基準は、ラベル付けサービス 142 のユーザ設定、ラベル付け基準設定又は拡張機能設定であり得る。新しく発見されたポジティブ例がバランスされるべきであるということを実行 655 が行えば、本方法は工程 660 へ進み、ここで最近発見されたポジティブ例に関係するアンチラベル例を取得するために必要とされるデータが判断される。一実施形態では、工程 660 において判断されたアンチラベルデータは、テキストスニペットの一組の優先キーワード、テキストスニペットの一組の優先キーワードの反転、アンチラベルの一組の優先キーワード、及びアンチラベルの優先キーワードの文脈内のテキストスニペットの優先キーワードの反転の一組の文脈意識キーワードを含む。アンチラベルの文脈の観点のテキストスニペットの反転の一組の文脈意識キーワードを取得すると、本方法は工程 620 へ進み、ここで一組の文脈意識キーワードの一組のランク付けされた検索結果が取得される。次に、本方法は、本明細書においても説明されるように、ネガティブ例を発見するための適切及び補足的なポジティブ例の本明細書において説明された方法を使用するが異なる入力を使用することにより工程 625、630、635 及び 640 を介しネガティブ例を発見することに進む。方法 600 により受信される入力はネガティブテキスト文脈意識キーワード（候補テキストを表現するための）及びアンチラベル（候補ラベルを表現するための）を含む。ネガティブ例が拡張方法 600 により捜される場合、補足データが、十分なクラス信頼度の工程 645 において拡張ネガティブ例を取得するために使用される。

【0116】

判断 650 において、ユーザ入力があるかどうか又はまだ拡張されていない残りの例が

あるかどうかを見るために試験が行われる。追加入力が受信されれば、方法 600 は工程 665 に進み、ここで追加入力が処理される。拡張されるべき追加例があれば、本方法は工程 605 に進み、ここで候補テキストが受信され、そして本方法は新しい入力データに関して繰り返す。工程 650 において、ユーザが追加入力を提供すれば、本方法は、改善された拡張を提供するために工程 665 において追加入力を使用する。例えば、新しく取得された例が表示エリア 235 においてユーザへ表示され、そしてユーザがコントロール 236 を使用することにより「確認」又は「グリーン」を選択すれば、工程 665 はこの例を強い例として記録し、そして追加例を生成するためにこの例を一組のサンプルへ追加することにより工程 620 に進む。代替的に、新しく発見された例が劣悪であるとユーザが判断したならば、ユーザは「拒絶」又は「レッド」の入力をコントロール 236 へ入力し、そして、本方法は、拡張するために、優先キーワードを規定した一組の例から新しい例を使用することにより工程 620 へ進むだろう。代替的に、キーワードがまだ定義されていなければ、本方法は工程 605 へ進むだろう。加えて、ユーザがラベルを修正し、修正されたラベル定義入力をグラフィックコントロール 206 内へ提供すれば、本方法はリセットし、そして新しいラベルにより工程 606 において拡張方法を開始し、新しいラベルに照らして複製されるべきすべての例を見る。

10

【0117】

新しいユーザ入力がないと判断 650 が行い、そして拡張のために考慮されるべき追加例がない場合、本方法は、拡張完了通知を表示し、そして、追加入力があるまで判断 650 において定期的に入力状態をサンプリングすることにより効果的に待つ。グラフィックコントロール 295 を選択することにより呼び出されるデータ拡張の工程では、拡張されたテキストは「選択されたスニペットの予測クラスが意図されたクラスに整合する」ということを判断するために雑音フィルタリングを使用することにより検証される。一実施形態では、閾値は、サンプルを許容するための信頼度の許容レベルを設定する。一実施形態では、文脈意識キーワードからのクエリー戻りの数は好適な候補を発見することなく消費される。この場合、この例は効果的にスキップされ、そしてエラーメッセージが格納される。拡張方法が完了すると、拡張の統計が、ユーザのために要約され、そしてユーザが拡張機能の成功の程度のインジケーションを受信するようにグラフィックディスプレイ 200 などの表示エリアにおいてユーザへ提示される。一実施形態では、追加される多くの成功ポジティブクラス例がエリア 203 内に表示され、追加される多くのネガティブクラス例がエリア 205 内に表示され、そして多くのスキップされたサンプルが表示エリア 201 内に表示される。

20

30

【0118】

ラベル付け基準を文書化する際、「サービスハーモニー」などのポジティブクラスの記述が搜される。動作環境 100 内に示されるシステムは、「楽しく、ビジネスライク：Pleasant and business-like」などの記述的入力からこのような表現を判断する。次に、ポジティブクラスは通常、顧客に役立つポジティブフィーチャと顧客幸福度及び忠実度を促進するポジティブフィーチャとを示す文章である。搜されるクラスの豊かな定義を得るために、「ディスサービス」又は「ディスハーモニー」のいずれかを反映する文章の例を有することも有用である。動作環境 100 内に示されるようなシステムは、意味論的に豊かで、多様なアイデアを有し、バランスがとれ、そして表現の強さのためにフィルタリングされる一組の例の生成及び拡張を提供する。文章がサービスハーモニーラベルにおけるポジティブ傾向又はネガティブ傾向のいずれも示さない場合、文章は通常、「非活性的：inert」又は「イエロー」とラベル付けされる。いくつかの文脈は、反対ケースからの距離ではなく「非活性的」ケースからの距離を与えられる文章の 2 つの閾値がセットアップされることを可能にする。「非活性的」ケースを反映するこのようなサンプルは、ポジティブ例又はネガティブ例のいずれかから引き出され、そして親例に特に近くはないということが分かった。

40

【0119】

テキストのアイデアは通常、文章内で使用されるワード／シーケンスに対する特定属性

50

を有しない文章の意味である。例えば、以下のこれらの2つの文章は同じアイデアを有する：

【0120】

犬は余りに疲れていたので通りを横断することができなかった。

【0121】

獵犬は疲労していたので路を横断することができなかった。

【0122】

翻訳の豊さは、同じアイデアを維持する一方で概して多様な表現である。以下は、1つの形式から別の形式へのテキストの拡張において多くの豊さがあるが同じアイデアを依然として表現する2つの例である。

【0123】

犬は余りに疲れていたので通りを横断することができなかった。

【0124】

私のペットはサンピエトロ寺院の反対側までの一層の努力をしたくないように見えた。

【0125】

アイデアにおける豊さは通常、様々な観点からの同じアイデアの表現である。下記は、同じ文脈（例えば「サービスハーモニーを描写する文章」）下のものであるが非常に異なるアイデアを有する2つの文章の例である。

【0126】

問題を抱える顧客をサービス及び親切を介しファンにさせることが私を良い気分にするので私は顧客サービスで働くことが好きです。

【0127】

私はあなたの言いたいことが分かる；私があるあなたの質問に対する答えを知らないということで苛立っているとあなたは言っているので、いくつかの文書と一緒に見よう、そしてあなたが必要とする情報を私が得ることができるかどうかを確かめよう。

【0128】

一般的に、最先端NLPモデルは、ラベル分類器を効果的にトレーニングするために同じラベルに属するトレーニングサンプル全体にわたるアイデアの豊さを必要とする。

【0129】

開示される解決策は、他の手法より優れている。特定トレーニング文脈要件の実際の、豊かで、及び人間生成されたデータは、いかなるNLPモデルもトレーニングするための「ゴールド」基準である。しかし、他の方法は、最先端及び大規模NLPモデルのための「豊かな」、「人間生成された、文脈要件意識、トレーニングデータを拡張するための効果的方法を提供しない。データ取得及びラベル付けの「手動」モードは、これらの最先端モデルに必要とされるデータの規模（いかなる伝統的NLPモデルの100×～100000×である可能性がある）に関しスケーリング可能でもないしコスト効率が高くもないので、本開示は、他のスケーリング可能手法との比較により評価され得る。

【0130】

データ拡張の他の方法は豊さ又は文脈意識を提供しない。最終モデル結果の比較は、本明細書において開示される方法を使用しない他の手法により行われることになる、又は本開示における1又は複数の方法及びサブ方法を使用することにより拡張を利用する場合はそれらの同じ方法により行われることになる。最良ベースラインは、最先端及び大規模事前トレーニング済みトランスフォーマベースNLPモデルを有するFew-Shot手法（例えばMicrosoft Turing）を使用しながら従われる。

【0131】

提案された方法は、主要な知的で、スケーリング可能で、及び文脈要件意識データ拡張方法を有することに加えて、拡張されたデータを雑音耐性的にするための追加方法も有する。データは、現代の最先端及び大規模トランスフォーマベースNLPモデル（例えばMicrosoft Turing）により供給されるベースラインの性能と比較して、孤立した主データ拡張モジュールと雑音低減アドオンを有する主データ拡張モジュールとの

10

20

30

40

50

両方の以下の性能により示される（これらのモジュール無しに同じデータサンプルに関し）。本方法は、特定文脈要件／ラベル記述に関し20個（各ポジティブ傾向及び各ネガティブ傾向毎に10個）～100個の記録の範囲内「だけ」のサンプルを有する標準化されたデータセットの様々なサイズのサブセットを取る。このような大規模モデルに関して、トレーニングサンプルのこの数は他の人により極めて小さいと考えられ、そして任意の現実アプリケーションのまずまずの性能モデルをトレーニングすることは不可能であると考えられる。この事実は、この技術無しにこのような小さな代表的データセットに関するトレーニング済みモデルの準最適性能により検証される。NLPシステムの「再現率」は20～100個のサンプルの範囲のサンプルサイズに関して4%～8%の範囲である。最終トレーニング済みモデルのすべての検証／試験が、同じ文脈要件（ラベル記述）のために作られた検証データセットに対し行われたが、モデルをトレーニングするために利用可能な一握りのトレーニングサンプル内に実際に提示され得るより文脈要件に関する多くの様々なより豊かで多様な「アイデア」を有する。このシナリオは、制限されたデータ源から（例えば、所与のトピックのポータルから、又は特定クラス／サブセットの標的視聴者が頻繁に訪れるポータルから）取得されるはるかに大きな（1000×）トレーニングデータセットのケースすら変則的に表現する。

10

【0132】

次に、同じデータサンプル（単に似たサイズのデータではない）を使用することにより、動作環境100のシステムは、開示された方法を、雑音低減アドオンモジュール無しスクーリング可能、インテリジェント、及び文脈要件意識方法フローだけにより一度実施する。本方法は、20～100個のサンプルサイズに関し8%再現率～17%再現率を提供した。雑音低減が行われると、性能は似ているが、サンプルサイズ40において14%の再現率の早期利点がある。両方の条件では、開示された方法はベースラインよりはるかに良い結果を提供した。より小さなデータサンプルサイズによっても、開示された雑音低減アドオンモジュールは更に良い結果を提供した。

20

【0133】

開示された方法は、基本的モデルがより良い再現率／FPR／精度を提供する可能性がある大規模最先端NLPデータのトレーニングデータを拡張し、そして拡張する可能性があるデータのアイデアの豊さ及び多様性に起因して、モデルは、文脈をより良く及びより全体論的に学習する可能性があり、これは、モデルが新データ／ドメインに関して合理的により良く行う可能性があるということの意味する。拡張されたサンプルは、検索ベースであり、従って人／企業生成された実際のサンプルであり、これは、現実アプリケーション下でこれらのシステムに関しトレーニングされたモデルがより信頼可能及び安定であるということを保証する。

30

【0134】

開示された方法は、アイデアの非常に多様な表現に豊かな文脈を学習することを要求する最先端トランスフォーマベースNLPモデルに関してすら、膨大な量の現実的及び人／企業生成されたトレーニングデータを拡張する可能性がある。

【0135】

開示された方法は文脈要件意識的である（その同意語／反意語により又はランダムワードを追加／置換することによりいずれかのワードを単に変更することとは対照的に）。これは、いかなる下流モデルの雑音も大いに低減するだけでなく下流のモデルのより関連するトレーニングデータも保証し、従ってモデルの性能、精度、関連性、信頼性及び安定性を改善するので巨大な便益である。

40

【0136】

大規模モデルのデータを手動で取得しそして等級付けすることは以下の理由で不十分である：

【0137】

第1に、最先端及び大規模トランスフォーマベースNLPモデルは、非常に豊かな表現データの最小でも何千ものサンプルを必要とするので。このようなデータは単一源から取

50

得するのが困難であり、時間がかかり、非常に高価である。このような手法は伝統的モデルに関しては過去においてうまく働いてきた（他の非 AI ベース拡張技術と併せて）が、現代 NLP エコシステムへはスケーリングされない。

【0138】

第2に、複数のデータ源からこのようなデータを取得することに関してすら、これらのデータは時間、費用、及び最も重要なことにはこのようなデータの等級付けに関係する偏りのすべてのニュアンスで等級付けされる必要がある。

【0139】

次に図13-15を参照すると、本明細書において説明される方法1300、1400及び1500の各ブロックは、ハードウェア、ファームウェア、及び／又はソフトウェアの任意の組み合わせを使用することにより行われ得るコンピューティングプロセスを含む。例えば、様々な機能はメモリ内に格納された命令をプロセッサが実行することにより実行され得る。本方法はまた、コンピュータストレージ媒体上に格納されたコンピュータ使用可能命令として具現化され得る。本方法は、いくつかの例を挙げるとスタンドアロンアプリケーション、サービス又はホストされたサービス（スタンドアロン、又は別のホストされたサービスとの組み合わせた）、又は別の製品へのプラグインにより提供され得る。加えて、方法1300、1400及び1500は、図1-12のシステム及び方法に関する一例として説明される。しかし、これらの方法は、制限しないが本明細書において説明されたものを含む任意の1つのシステム、又はシステムの任意の組み合わせにより追加的に又は代替的に実行され得る。

【0140】

図13は、本開示のいくつかの実施形態によるクラスラベルとテキストとの対応を判断する方法1300を示すフローチャートである。方法1300は、ブロック1302において、候補テキストを受信することを含む。図2を参照して前に説明したように、候補テキストはユーザインタフェースを介し受信され得る。代替的に、候補テキストは一群の文書、電子メール又はテキストの他の源であり得る。いくつかの態様では、候補テキストはより大きな文書の一部（文章、文書の句又はパラグラフなど）であり得る。方法1300は、ブロック1304において、ラベル記述を受信することを含む。図2を参照して前に説明したように、ラベル記述はユーザインタフェースを介し受信され得る。ユーザは、1又は複数の文書、電子メール、テキスト、ソーシャルメディアポスト又は他のテキストコンテンツがラベル記述に対応するかどうかを判断する目的でラベル記述を提出し得る。例えば、ユーザは、顧客サービスを具現化する文書を識別したいかもしれない。方法1300は、ラベル記述が候補テキストに対応するかどうかを判断し得る。ラベルは、テキスト及びラベル記述における概念が同様な意味を有すると候補テキストに対応する。

【0141】

方法1300は、ブロック1306において、クエリーを生成するためにラベル記述を使用することを含む。例えば、ラベルから導出された優先キーワードは、図6の工程615及び工程620により説明されたようなクエリーとして使用される。代替的に、ラベルから導出された優先キーワードは、図7に示すように工程615により一組の文脈意識的キーワードを形成するための例から導出された優先キーワードと連結して使用される。

【0142】

方法1300は、ブロック1308において、クエリーを検索エンジンへ伝達することを含む。ラベル付けサービス142がクエリーを検索サービス164へ送信する。一実施形態では、検索サービス164は検索エンジンのAPI版である。一実施形態では、検索のクライアント版が使用される。検索サービス164はクエリーを受信し、そして文書コーパス154全体にわたる検索を行う。検索エンジンは、ランク付けされた回収結果（各結果のランク及び各結果の検索スコアを含む）のブロックと、クエリーに関連する場所において文書をサンプリングするテキストスニペットとを判断する。検索サービス164は図6の工程620との関連で論述されたように一組のランク付けされた検索結果を取得する。

10

20

30

40

50

【0143】

方法1300は、ブロック1310において、クエリーに応じたテキストストリングを検索エンジンから受信することを含む。検索サービス164は、各ランク付けされた検索結果のテキストスニペットを含む一組のランク付けされた検索結果を含む結果ページをラベル付けサービス142へ送信する。いくつかのケースでは、高ランク又は高検索スコア関連性のエントリは、テキストストリングとしてテキストスニペットを選択することによりラベル付けサービス142により選択される。いくつかのケースでは、図6に関連して説明したように、「ランク付けされた検索結果のリストが、工程625において説明されたようにテキストスニペットを取得することにより、テキストスニペットのラベルクラス信頼度を工程630において定量化することにより、そしてテキストスニペットが十分な信頼度を有する適切なクラスを有すれば判断635において判断することにより評価される」ループが形成される。そうでなければ、本方法は、工程640において当該スニペットの失敗を記録し、そして工程625に戻る。十分な信頼度のテキストスニペットが工程635において発見されると、十分な信頼度のものであると分かったテキストスニペットが、クエリーに応じたテキストストリングとして選択される。

10

【0144】

方法1300は、ブロック1312において、テキストストリング及び候補テキストを生成モデルへ入力することを含む。テキストストリングは、基本的にポジティブ例又はネガティブ例であるので、本明細書において開示される例示的处理と関連して使用される。いくつかのケースでは、生成モデルのリスクの量を示すパラメータがストレージ180から回収される。例示的テキストストリングを候補モデルへ入力するための本明細書において説明される以下の4つの基本的方法がある：NC方法（図4、工程410及び工程420）、SL方法（図4、工程410及び工程420）、SS方法（図5、工程510及び工程520）、及びLP方法（図12、工程1210、工程1220及び工程1230）。いくつかの実施形態における生成モデルのモードはzero-shotモードである。

20

【0145】

方法1300は、ブロック1314において、複数のトークン及び関連確率を含む生成されたテキストを生成モデルから受信することを含む。生成されたテキストは、モデルにより生成されるテキストトークンの実際のストリームだけでなく各トークンに関して報告された関連トークン確率及びログ確率のベクトルも広範に含み、ここで、各ログ確率は、モデルが選択した可能性があるトークンに対応する一定数の尤度を記述する。上述のように、生成モデルから生成されたテキストを受信するための本明細書において開示される（NC、LP、SS及びLP方法において開示される）4つの基本的方法がある。図4に示すNC方法では、テキストが、受信され、そして工程420のNC実施形態に関連して説明されるようにクラスラベルに関して走査される。図4にも示されるSL方法では、テキストが、受信され、そして工程420のSL実施形態に関連して説明されるようにラベル及びアンチラベルのキーワードに関して走査される。図5に示すSS方法では、生成されたテキストは、工程530において説明されるように検索クエリーにおいて使用される。図12に示すLP方法では、ログ確率が工程1215、1225及び1235との関連で使用される。

30

40

【0146】

方法1300は、ブロック1316において、生成されたテキストに基づきラベル確率推定を判断することを含む。再び、我々は、NC、SL、SS及びLP方法において開示されるようにラベル確率を判断する4つの基本的方法を有する。NC方法では、工程430において、ラベル番号又はアンチラベル番号のトークン確率は、いくつかの実施形態では実験的に推定されたスケーリング係数を使用する近似に対する入力として使用される。SL方法では、工程430において、ラベル又はアンチラベルの又はそれらの同意語のキーワードのトークン確率が、アンチラベルインジケーションとは対照的にラベルインジケーションの強さの近似を形成するために使用される。SS方法では、工程540において、調停規則が、ネガティブ例文書とは対照的にポジティブ例文書のランクのバランスを保

50

つために使用される。LP方法では、工程1250において、予測可能性の閾値を越える結果は、ラベル確率を近似するためにポジティブ確率に対する及びネガティブ確率に対するスケーリングを提供する。

【0147】

方法1300は、ブロック1318において、候補テキストがラベル確率推定に基づきラベル記述に対応するかどうかのインジケーションを出力することを含む。図2を参照して前に説明したように、インジケーションはユーザインタフェースを介し出力され得る。一態様では、このインジケーションは2進法「はい／いいえ」又は同様なインジケーションであり得る。他の態様では、このインジケーションは相関の度合い又は強さを表現し得る。

10

【0148】

図14は、本開示のいくつかの実施形態によるクラスラベルとテキストとの対応を判断する方法の方法1400を示すフローチャートである。方法1400は、ブロック1402において、候補テキストを受信することを含む。図2を参照して前に説明したように、候補テキストはユーザインタフェースを介し受信され得る。代替的に、候補テキストは一群の文書、電子メール又はテキストの他の源であり得る。いくつかの態様では、候補テキストはより大きな文書の一部（文章、文書の句又はパラグラフなど）であり得る。方法1400は、ブロック1404において、ラベル記述を受信することを含む。図2を参照して前に説明したように、ラベル記述はユーザインタフェースを介し受信され得る。ユーザは、1又は複数の文書、電子メール、テキスト、ソーシャルメディアポスト又は他のテキストコンテンツがラベル記述に対応するかどうかを判断する目的でラベル記述を提出し得る。例えば、ユーザは、顧客サービスを具現化する文書を識別したいかもしれない。

20

【0149】

方法1400は、ブロック1406において、生成モデルから、生成モデルへの入力としての候補テキストを有する候補結果を生成することを含む。方法1400は、ラベル記述が候補テキストに対応するかどうかを判断し得る。ラベルは、テキスト及びラベル記述における概念が同様な意味を有すると候補テキストに対応する。一実施形態では、ラベルは、それぞれがラベルを又はラベルに適合する具体例を具現化するいくつかの例を正しく説明する抽象的概念又はカテゴリである。生成モデルから、生成モデルへの入力としての候補テキストを有する候補結果を生成する工程が図5の工程520において説明される。グラフィックディスプレイ200からの候補テキスト入力の一例はグラフィックコントロール202に示すような「あなたのスプロケットの注文をお手伝いできて嬉しいです」である。

30

【0150】

方法1400は、ブロック1408において、生成モデルから、生成モデルへの入力として、ラベル記述を具現化するポジティブ例テキストを有するポジティブ例結果を生成することを含む。工程1408及び1410は図5の工程530において概して説明される。グラフィックディスプレイ200において示される例では、ポジティブ例テキストは、表示エリア231に示すような「あなたのために何か他にできることがあれば教えてください。喜んでお手伝いします」である可能性がある。

40

【0151】

方法1400は、ブロック1410において、生成モデルから、生成モデルへの入力としてラベル記述の反対の概念を具現化するネガティブ例テキストを有するネガティブ例結果を生成することを含む。グラフィックディスプレイ200において示されるようなネガティブ例テキストの一例はグラフィック表示エリア261に示すような「これはあなたの問題であり私の問題ではない」である可能性がある。

【0152】

方法1400は、ブロック1412において、ポジティブ例結果及びネガティブ例結果を含むコーパス全体にわたる第2のクエリーとして候補結果を検索エンジンへ提出することからの応答に基づきポジティブ例結果の第1のランクスコアを判断することを含む。ラ

50

ンクスコアは数値ランク 1、2、3 であり得、ここでより低い数値はより高いランク（第 1 リスト）を実際に反映する。ランクスコアは、候補結果とポジティブ例結果との間のコサイン類似度であり得る。

【0153】

方法 1400 は、ブロック 1414 において、ネガティブ例結果及びネガティブ例結果を含むコーパス全体にわたる第 2 のクエリーとして候補結果を検索エンジンへ提出することからの応答に基づきネガティブ例結果の第 2 のランクスコアを判断することを含む。ランクスコアは例えば候補結果とネガティブ例結果との間のコサイン類似度であり得る。類似性測度は意味論的検索エンジンを使用することにより深層ベクトル空間内で測定され得る。

10

【0154】

方法 1400 は、ブロック 1416 において、ポジティブ例結果の第 1 のランクスコアとネガティブ例結果の第 2 のランクスコアとを比較することによりラベル確率推定を判断することを含む。本明細書において開示される調停規則は確率を推定するために使用され得る。一実施形態では、ラベル確率は、平均ポジティブ例コサイン類似度と平均ネガティブ例コサイン類似度との間のスケーリングされた比較である。一実施形態では、スケーリング係数は、減衰係数としてランダムに選択されたテキストのコサイン類似度を発見することにより判断される。一実施形態では、スケーリング係数は係数としてユーザ確認の比率を測定することにより判断される。

【0155】

方法 1400 は、ブロック 1418 において、候補テキストがラベル確率推定に基づきラベル記述に対応するかどうかのインジケーションを出力することを含む。一態様では、このインジケーションは 2 進法「はい／いいえ」又は同様なインジケーションであり得る。他の態様では、このインジケーションは、相関の度合い又は強さを表現し得る。

20

【0156】

図 15 は、本開示のいくつかの実施形態による分類器のトレーニングデータを拡張する方法 1500 を示すフローチャートである。

【0157】

方法 1500 は、ブロック 1502 において、分類器のための、クラスラベルに関連する例示的テキストを含むトレーニングデータインスタンスを受信することを含む。トレーニングデータインスタンスはインターフェースを介しユーザにより提供され得る。別の態様では、トレーニングデータはトレーニングデータの集合から引き出される。

30

【0158】

方法 1500 は、ブロック 1504 において、例示的テキストの一組の優先キーワードを判断することを含む。優先キーワードは、例えば図 9 に関連して説明されたように判断される。

【0159】

方法 1500 は、ブロック 1506 において、クラスラベルの一組の優先キーワードを判断することを含む。クラスラベルのために判断される一組の優先キーワードは、例えば図 3 の工程 307 及び図 9 において説明されるように判断される。

40

【0160】

方法 1500 は、ブロック 1508 において、一組の優先キーワード及び一組の優先キーワードから一組の文脈意識キーワードを判断することを含む。一組の文脈意識キーワードを判断する方法は図 7 において説明された。文脈意識キーワードの一例は、グラフィックディスプレイ 200 の表示エリア 204 に示すような「helping, happy, customer」であり得る。

【0161】

方法 1500 は、ブロック 1510 において、一組の文脈意識キーワードを含むクエリーを検索エンジンへ伝達することを含む。ラベル付けサービス 142 は、文脈意識キーワードを含むクエリーを検索サービス 164 へ送信する。一実施形態では、検索サービス 1

50

64は検索エンジンのAPI版である。一実施形態では、検索のクライアント版が使用される。検索サービス164はクエリーを受信し、そして文書コーパス154全体にわたる検索を行う。検索エンジンは、ランク付けされた検索結果のブロック（各結果のランク及び各結果の検索スコアを含む）とクエリーに関連する場所において文書をサンプリングするテキストスニペットとを判断する。検索サービス164は図6の工程620との関連で論述されたように一組のランク付けされた検索結果を取得する。

【0162】

方法1500は、ブロック1512において、テキストスニペットをクエリーに応じて検索エンジンから受信することを含む。検索サービス164は、各ランク付けされた検索結果のテキストスニペットを含む一組のランク付けされた検索結果を含む結果ページをラベル付けサービス142へ送信する。いくつかのケースでは、高ランク又は高検索スコア関連性のエントリはラベル付けサービス142により選択される（これによりテキストスニペットを選択する）。いくつかのケースでは、図6に関連して説明したように、「ランク付けされた検索結果のリストが、工程625において説明されたように潜在的テキストスニペットを取得することにより、潜在的テキストスニペットのラベルクラス信頼度を工程630において定量化することにより、そして潜在的テキストスニペットが十分な信頼度を有する適切なクラスを有すれば判断635において判断することにより評価される」ループが形成される。そうでなければ、本方法は、工程640において当該スニペットの失敗を記録し、そして工程625に戻る。十分な信頼度の潜在的テキストスニペットが工程635において発見されると、十分な信頼度のものであると分かった潜在的テキストスニペットが、クエリーに応じて、戻されるべきテキストスニペットとして選択される。

【0163】

方法1500は、ブロック1514において、テキストスニペット及びクラスラベルを含む拡張されたトレーニングデータインスタンスを生成することを含む。一実施形態では、ラベル付け基準は、テキストスニペットを含む追加例であってクラスラベルに関連付けられた追加例を含むことにより増加される。本明細書において開示されるラベル付け基準の追加例を含むためにラベル付け基準を格納、修正、強化する方法は、テキストスニペット又はクラスラベルの新しい例を含む拡張されたインスタンス（又はラベル標準）を生成する例である。

【0164】

方法1500は、ブロック1516において、拡張されたトレーニングデータインスタンスを有するトレーニングされた分類器を使用することにより候補テキストをクラスへ分類することを含む。

【0165】

方法1500は、ブロック1518において、候補テキストがクラスに対応するラベルに対応するというインジケーションを出力することを含む。一態様では、このインジケーションは2進法「はい／いいえ」又は同様なインジケーションであり得る。他の態様では、このインジケーションは、相関の度合い又は強さを表現し得る。

【0166】

例示的動作環境

添付図面を概して参照すると、そして当初特に図8を参照すると、本明細書において説明される技術の態様を実装するための例示的動作環境がコンピューティングデバイス800として概して示され指定される。コンピューティングデバイス800は、好適なコンピュータ環境の単に一例であり、したがって本明細書において説明される技術の使用の範囲に関するいかなる制限も示唆するようには意図されていない。コンピューティングデバイス800は、示された部品のうちの任意の部品又はその任意の組み合わせにも関係するいかなる依存性も要件も有すると解釈されるべきでない。

【0167】

本明細書において説明される技術は、コンピュータ又は他の機械（パーソナルデータアシスタント又は他のハンドヘルドデバイスなどの）により実行されるコンピュータコード

10

20

30

40

50

又は機械使用可能命令（プログラム部品などのコンピュータ実行可能命令を含む）の一般的文脈において説明され得る。一般的に、ルーチン、プログラム、及びオブジェクト、部品、データ構造等々を含むプログラム部品は、特定タスクを行う又は特定抽象的データ型を実装するコードを指す。本明細書において説明される技術は、ハンドヘルドデバイス、民生用電子機器、汎用コンピュータ、特定コンピューティングデバイスなどを含む多種多様なシステム構成で実行され得る。本明細書において説明される技術の態様はまた、タスクが通信ネットワークを介しリンクされる遠隔処理デバイスにより行われる分散コンピューティング環境内で実行され得る。

【0168】

図8を引き続き参照すると、コンピューティングデバイス800は、以下のデバイスに直接的又は間接的に結合するバス810を含む：メモリ812、1又は複数のプロセッサ814、1又は複数のプレゼンテーション部品816、入出力（I/O）ポート818、I/O部品820及び例示的電源822。バス810は、1又は複数のバスであり得るもの（アドレスバス、データバス又はその組み合わせなど）を表現する。図8の様々なブロックは明確化のために線で示されるが、現実的には、様々な部品の描写はそれほど明白ではなく、そして隠喩的には、線はより正確には灰色及び曖昧だろう。例えば、ディスプレイデバイスなどのプレゼンテーション部品はI/O部品であると考えられ得る。また、プロセッサはメモリを有する。本発明者らは、このようなものが本技術の本質であるということ認識し、そして図8の線図が、本明細書において説明される技術の1又は複数の態様に関連して使用され得る例示的コンピューティングデバイスの例示に過ぎないということ繰り返して述べておく。「ワークステーション」、「サーバ」、「ラップトップ」、「ハンドヘルドデバイス」などのようなカテゴリは、すべてが図8の範囲内で考慮されそして「コンピュータ」又は「コンピューティングデバイス」を指すので区別されない。

【0169】

コンピューティングデバイス800は通常、多種多様なコンピュータ可読媒体を含む。コンピュータ可読媒体は、コンピューティングデバイス800によりアクセスされ得そして揮発性及び不揮発性媒体、取り外し可能及び取り外し不能媒体の両方を含む任意の利用可能媒体であり得る。一例として、そして制限することなく、コンピュータ可読媒体はコンピュータストレージ媒体及び通信媒体を含み得る。コンピュータストレージ媒体は、コンピュータ可読命令、データ構造、プログラムモジュール又は他のデータなど情報の格納のための任意の方法又は技術で実装された揮発性及び不揮発性、取り外し可能及び取り外し不能媒体を含む。

【0170】

コンピュータストレージ媒体は、RAM、ROM、EEPROM、フラッシュメモリ又は他のメモリ技術、CD-ROM、デジタルバーサタイルディスク（DVD：digital versatile disks）又は他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ又は他の磁気ストレージデバイスを含む。コンピュータストレージ媒体は伝搬型データ信号を含まない。

【0171】

通信媒体は通常、コンピュータ可読命令、データ構造、プログラムモジュール、又は変調データ信号（搬送波又は他の移送機構など）内の他のデータを具現化し、そして任意の情報配送媒体を含む。用語「変調データ信号」は、その特性のうちの1又は複数が信号内の情報を符号化するようなやり方で設定又は変更された信号を意味する。一例として、そして制限することなく、通信媒体は、有線ネットワーク又は直接有線接続などの有線媒体及び音響、RF、赤外線及び他の無線媒体などの無線媒体を含む。上記のもののうちの任意のものの組み合わせもコンピュータ可読媒体の範囲内で含まれるべきである。

【0172】

メモリ812は、揮発性及び／又は不揮発性メモリの形式のコンピュータストレージ媒体を含む。メモリ812は取り外し不能、取り外し可能、又はその組み合わせであり得る。例示的メモリは固体メモリ、ハードドライブ、光ディスクドライブなどを含む。コンピ

ューティングデバイス 800 は、バス 810、メモリ 812 又は I/O 部品 820 などの様々なエンティティからデータを読む出す 1 又は複数のプロセッサ 814 を含む。プレゼンテーション部品 816 はデータインジェクションをユーザ又は他のデバイスへ提示する。例示的プレゼンテーション部品 816 はディスプレイデバイス、スピーカ、印刷部品、振動部品などを含む。I/O ポート 818 は、コンピューティングデバイス 800 が他のデバイス (I/O 部品 820 を含む) へ論理的に結合されることを可能にし、そのうちのいくつかは組み込まれ得る。

【0173】

I/O 部品はマイクロホン、ジョイスティック、ゲームパッド、衛星放送アンテナ、スキャナ、プリンタ、ディスプレイデバイス、無線デバイス、コントローラ (スタイラス、キーボード及びマウスなど)、ナチュラルユーザインタフェース (NUI: natural user interface) 等々を含む。いくつかの態様では、ペンディジタイザ (示されない) 及び付随入力器具 (もまた示されないが一例としてペン又はスタイラスだけを含み得る) が手書きユーザ入力をディジタル的に捕捉するために提供される。ペンディジタイザとプロセッサ 814 との間の接続は直接的であってもよいし、シリアルポート、パラレルポート、及び/又は当該技術領域において知られている他のインターフェース及び/又はシステムバスを利用する結合を介してもよい。更に、ディジタイザ入力部品はディスプレイデバイスなどの出力部品から分離された部品であり得る、又はいくつかの態様では、ディジタイザの使用可能入力エリアは、ディスプレイデバイスの表示エリアと共存し、そしてディスプレイデバイスと一体化されてもよいし、又は、それを覆う別個のデバイスとして存在してもよい、又は、そうでなければディスプレイデバイスへ付加されてもよい。あらゆるこのような変形形態及びその任意の組み合わせが、本明細書において説明される技術の態様の範囲内で企図される。

【0174】

NUI は、ユーザにより生成される空中ジェスチャ、音声又は他の生理学的入力を処理する。適切な NUI 入力は、コンピューティングデバイス 800 と関連したプレゼンテーションのためのインクストロークとして解釈され得る。これらの要求はさらなる処理のために適切なネットワーク要素へ送信され得る。NUI は、音声認識、接触及びスタイラス認識、顔認識、バイオメトリック認識、スクリーン上のジェスチャ認識及びスクリーンに隣接するジェスチャ認識、空中ジェスチャ、頭及び目追跡、及びコンピューティングデバイス 800 上の表示に関連する接触認識の任意の組み合わせを実施する。コンピューティングデバイス 800 は、ジェスチャ検出及び認識のためのステレオカメラシステム、赤外線カメラシステム、RGB カメラシステム及びこれらの組み合わせなどの深度カメラを装備し得る。加えて、コンピューティングデバイス 800 は、運動の検出を可能にする加速度計又はジャイロスコープを装備し得る。加速度計又はジャイロスコープの出力は、没入型拡張現実又は仮想現実を描画するためにコンピューティングデバイス 800 のディスプレイへ提供され得る。

【0175】

実施形態

本明細書において説明される技術は、制限的であるというよりむしろ例示的であるようにすべての態様において意図された特定態様に関連して説明された。本明細書において説明される技術は様々な修正形態及び代替構造の影響を受け易いが、そのいくつかの示された態様が添付図面において示されそして上に詳細に説明された。しかし、次のことが理解されるべきである: 本明細書において説明された技術を開示された特定形態へ制限する意図はないが、逆に、本意図は、すべての修正形態、代替構造、及び本明細書において説明された技術の精神及び範囲に入る等価物をカバーすることである。

【0176】

例えば、コーパス 154 全体にわたり文書をラベル付けするラベル付けサービス 122 は時には CRM データの企業コーパスを論述したが、ラベル付けサービスは、文書の任意のコーパス全体にわたり文書の一部をラベル付けし得る。コーパス 154 は、パーソナル

ハードドライブ、クラウドストレージ、一組のウェブページ、映画データベースなどの一部分である可能性がある。

【0177】

加えて、ラベル付けアプリケーション110は、ラベル付け結果を提供するアプリケーションとして一般的に説明された。ラベル付けアプリケーション110は有利な組み合わせを介し検索サービス164と組み合わせられ得る。例えば、検索サービス164からのより大きな一組の結果は、ラベルに適合しない戻りを削除するためにラベル付けサービスを介しフィルタリングされ得る。別の例として、検索サービス164は100個の最も関連する戻りを戻すように構成され得、そしてラベルに関連するそれらの戻りは、ランクリストのトップへ移動される可能性がある。一実施形態では、ユーザはラベル記述をグラフィックコントロール206内へタイプ入力し、そして、検索サービス164は可能なポジティブ例を提示する一組のエントリ及び可能なネガティブ例を提示する一組のエントリをユーザへ戻す。ユーザはポジティブ例及びネガティブ例を選択し、そして本方法は、ユーザ選択ポジティブエントリのテキストスニペットから取られたポジティブ例とユーザ選択ネガティブエントリのテキストスニペットから取られたネガティブ例とにより方法300を行うことに進む。次に、検索サービス164は処理をラベル付けサービス142へ渡す。ラベル付けサービス142は、エントリが、生キーワード類似度ではなくラベル確率に基づきランク付けされ、そしてウェブ結果の意味論的に関連するリストとしてユーザへ提示されるように、ユーザにより入力されたラベルに照らして、検索サービス164により戻された各エントリからの各テキストスニペットが候補テキストとして評価される方法300を呼び出すことにより検索サービス164の入力フィルタリングを行うことに進む。

10

20

【0178】

更に、ラベル付けアプリケーション110は、ラベル強さインデックスを提供する文書のコーパスの検索インデックスを生成しそしてキーワード関連性ではなくむしろラベル強度の組み合わせに基づき文書を戻すために使用される可能性がある。更に、検索ランクを判断するために重み付け組み合わせとしてキーワードインデックス及びラベル強度インデックスを重み付けるハイブリッド検索が生成され得る。

【0179】

本明細書において開示される分類レベルは時にはラベル及びアンチラベルとしての2進レベルだった。本明細書において説明される技術は、多項式ラベル分類器を提供するために多項式レベルを処理することができる。

30

【0180】

加えて、検索又はウェブ検索が本明細書において説明される場合はどこでも、意味論的近さに基づく意味論的検索が伝統的キーワード検索の代わりに行われ得る。

【0181】

実施形態1. クラスラベルとテキストとの対応を判断する方法であって候補テキストを受信することとラベル記述を受信することを含む方法。本方法はまた、クエリーを生成するためにラベル記述を使用することを含む。本方法はまた、クエリーを検索エンジンへ伝達することを含む。本方法はまた、クエリーに応じてテキストストリングを検索エンジンから受信することを含む。本方法はまた、テキストストリング及び候補テキストを生成モデルへ入力することを含む。本方法はまた、複数のトークン及び関連確率を含む生成されたテキストを生成モデルから受信することを含む。本方法はまた、生成されたテキストに基づきラベル確率推定を判断することを含む。本方法はまた、候補テキストがラベル確率推定に基づくラベル記述に対応するかどうかのインジケーションを出力することを含む。

40

【0182】

実施形態2. ラベル確率推定はラベルに対応する生成されたテキストのトークン確率から判断される、実施形態1に記載の方法。

【0183】

実施形態3. ラベルはポジティブラベル又はアンチラベルである実施形態2に記載の方

50

法。

【0184】

実施形態4．ラベル確率推定はラベル記述のキーワード又はアンチラベルのキーワードに対応する生成されたテキストのトークン確率から判断される実施形態1乃至3のいずれか一項に記載の方法。

【0185】

実施形態5．検索エンジンの検索エンジン技術は規則ベース検索、意味論的近さに基づく意味論的検索、又はトランスフォーマモデルを使用する文脈検索から成るグループから選択される実施形態1乃至4のいずれか一項に記載の方法。

【0186】

実施形態6．生成されたテキストに基づきラベル確率推定を判断することは、生成されたテキストに基づく第1のラベルスコアへ適用される第1の重み付けと、候補テキストが第2の生成モデルへ入力された場合に第2の生成モデルから受信された第2の生成されたテキストに基づく第2のラベルスコアへ適用される第2の重み付けとを使用することを含む実施形態1乃至5のいずれか一項に記載の方法。

【0187】

実施形態7．第1の重み付け及び第2の重み付けは、ラベル記述に似ている異なるラベル記述の一組の格納された重み付けを発見することにより判断される実施形態6に記載の方法。

【0188】

実施形態8．コンピューティングデバイスにより実行されるとコンピューティングデバイスにクラスラベルとテキストとの対応を判断する方法であって候補テキストを受信することとラベル記述を受信することを含む方法を行わせる命令を含むコンピュータ可読媒体。本方法はまた、生成モデルから、生成モデルへの入力としての候補テキストを有する候補結果を生成することを含む。本方法はまた、生成モデルへの入力として、生成モデルから、ラベル記述を具現化するポジティブ例テキストを有するポジティブ例結果を生成することを含む。本方法はまた、生成モデルへの入力として、生成モデルから、ラベル記述の反対の概念を具現化するネガティブ例テキストを有するネガティブ例結果を生成することを含む。本方法はまた、ポジティブ例結果及びネガティブ例結果を含むコーパス全体にわたる第2のクエリーとして候補結果を検索エンジンへ提出することからの応答に基づきポジティブ例結果の第1のランクスコアを判断することを含む。本方法はまた、ネガティブ例結果及びネガティブ例結果を含むコーパス全体にわたるクエリーとして候補結果を検索エンジンへ提出することからの応答に基づきネガティブ例結果の第2のランクスコアを判断することを含む。本方法はまた、ポジティブ例結果の第1のランクスコアとネガティブ例結果の第2のランクスコアとを比較することによりラベル確率推定を判断することを含む。本方法はまた、候補テキストがラベル確率推定に基づくラベル記述に対応するかどうかのインジケーションを出力することを含む。

【0189】

実施形態9．検索エンジンは意味論的検索エンジンである実施形態8に記載の媒体。

【0190】

実施形態10．生成モデルはzero-shotモードにおいて実行されるGPT3である実施形態8乃至9のいずれか一項に記載の媒体。

【0191】

実施形態11．インジケーションは、ラベル確率推定と、異なる方法により計算された第2のラベル確率推定との重み付け組み合わせに基づく実施形態8乃至10のいずれか一項に記載の媒体。

【0192】

実施形態12．候補テキストは文書のコーパスである実施形態11に記載の媒体。

【0193】

実施形態13．1又は複数のプロセッサ；及び1又は複数のプロセッサにより使用され

10

20

30

40

50

ると1又は複数のプロセッサに方法を行わせるコンピュータ使用可能命令を格納する1又は複数のコンピュータストレージ媒体を含むシステム。本方法は、分類器のための、クラスラベルに関連する例示的テキストを含むトレーニングデータインスタンスを受信することを含む。本方法はまた、例示的テキストの一群の優先キーワードを判断することを含む。本方法はまた、クラスラベルの一群の優先キーワードを判断することを含む。本方法はまた、一群の優先キーワード及び一群の優先キーワードから一群の文脈意識キーワードを判断することを含む。本方法はまた、一群の文脈意識キーワードを含むクエリを検索エンジンへ伝達することを含む。本方法はまた、テキストスニペットをクエリに応じて検索エンジンから受信することを含む。本方法はまた、テキストスニペット及びクラスラベルを含む拡張されたトレーニングデータインスタンスを生成することを含む。本方法はまた、拡張されたトレーニングデータインスタンスによりトレーニングされた分類器を使用することにより候補テキストをクラスへ分類することを含む。本方法はまた、候補テキストがクラスに対応するラベルに対応するというインジケーションを出力することを含む。

10

【0194】

実施形態14．例示的テキストはクラスラベルのポジティブ例である実施形態13のシステム。

【0195】

実施形態15．例示的テキストはクラスラベルのネガティブ例である実施形態13のシステム。

【0196】

20

実施形態16．本方法は例示的テキストの一群の優先キーワード及びクラスラベルの一群の優先キーワードをグラフ構造内に格納することを更に含む実施形態14又は15のいずれかに記載のシステム。

【0197】

実施形態17．本方法は更に、例示的テキストの一群の優先キーワードの用語の第1の埋め込みを取得することを含む実施形態14、15又は16のいずれかに記載のシステム。本方法はまた、クラスラベルの一群の優先キーワードの用語の第2の埋め込みを取得することを含む。本方法はまた、文脈意識キーワードを判断するために第1の埋め込み及び第2の埋め込みに対する操作を使用することを含む。

【0198】

30

実施形態18．操作を使用することは、例示的テキストの一群の優先キーワードの用語とクラスラベルの一群の優先キーワードの用語との間のコサイン類似度を計算することを含む実施形態17のシステム。

【0199】

実施形態19．一群の文脈意識キーワードを判断することはクラスラベルのキーワードの文脈に対する例示的テキストの一群の優先キーワードの各用語の関連性に従って例示的テキストのキーワードをフィルタリングすることを含む実施形態14、15、16、17、又は18のいずれかのシステム。

【0200】

実施形態20．本方法は「テキストスニペット及びクラスラベルを受信しそしてテキストスニペットがクラスラベルを具現化する確率は閾値を越えるというインジケーションを戻すラベル採点方法」を使用することによりテキストスニペットがクラスラベルを表現する可能性が高いことを確認することを更に含む実施形態14、15、16、17、18、又は19のいずれかのシステム。

40

【0201】

実施形態21．クラスラベルとテキストとの対応を判断する方法であって候補テキストを受信することを含む方法。本方法は更に、ラベル記述を受信すること；ラベル記述を具現化するポジティブ例テキストを受信することを含む。本方法は更に、ラベル記述と反対の概念を具現化するネガティブ例テキストを受信することを含む。本方法は更に、ポジティブ例結果を取得するためにポジティブ例テキスト及び候補テキストへ生成モデルを適用

50

することを含む。本方法は更に、ネガティブ例結果を取得するために生成モデルをネガティブ例テキスト及び候補テキストへ適用すること；ベースライン結果を取得するために生成モデルをポジティブ例テキスト、ネガティブ例テキスト及び候補テキストへ適用すること；ポジティブ例結果の関連ログ確率とベースライン結果の文脈内のネガティブ例結果の関連ログ確率とを比較することによりラベル確率推定を判断することを含む。本方法は更に、候補テキストがラベル確率推定に基づくラベル記述に対応するかどうかのインジケーションを出力することを含む。

【0202】

実施形態22．生成されたテキストのトークン確率はラベルに対応する数のトークン確率を含む実施形態2の方法。

10

【0203】

実施形態23．生成されたテキストのトークン確率はアンチラベルに対応するトークン確率を含む実施形態2の方法。

【0204】

実施形態24．ラベル確率推定はアンチラベルに対応する生成されたテキストのトークン確率から判断される実施形態2の方法。

【0205】

実施形態25．ラベル確率推定はストリングラベルのキーワードの同意語である生成されたテキストからの用語のトークン確率から判断される実施形態2の方法。

20

【0206】

実施形態26．ラベル確率推定はストリングのラベルのキーワードである生成されたテキストからの用語のトークン確率から判断される、実施形態2の方法。

【0207】

実施形態27．2つのトークン確率は全体的確率推定を形成するために組み合わせられる実施形態24又は25の方法。

【0208】

実施形態28．トークンラベル確率推定は、ストリングラベルのキーワード又はストリングラベルのキーワードの同意語である生成テキストから2つの用語の確率を取り込む実施形態25又は26の方法。

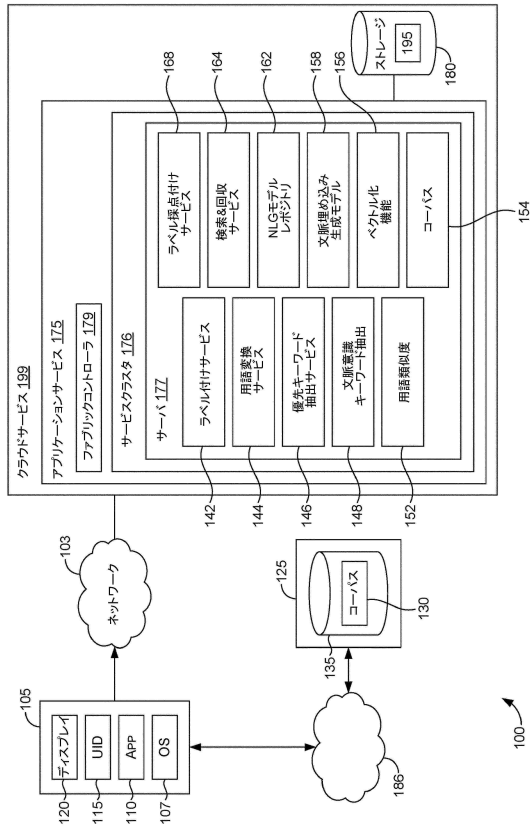
30

40

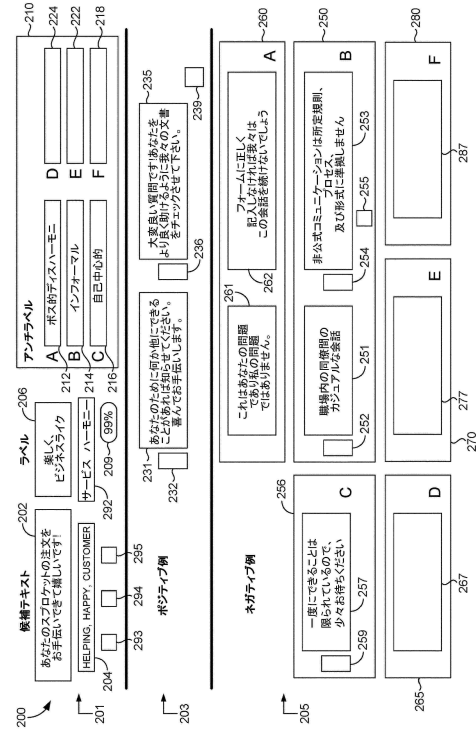
50

【凶面】

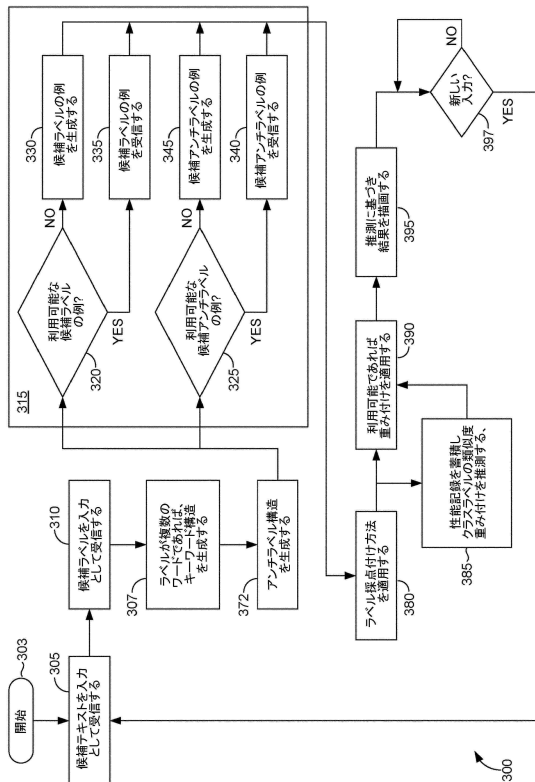
【図 1】



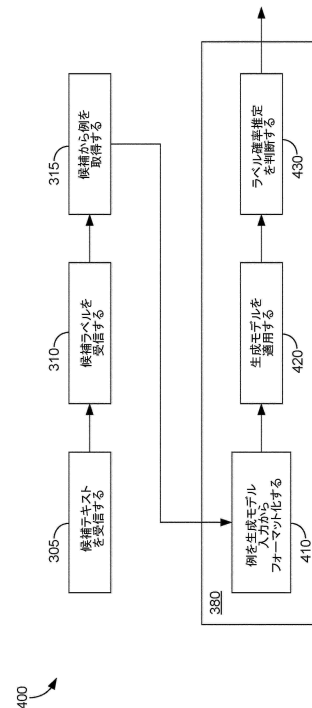
【圖 2】



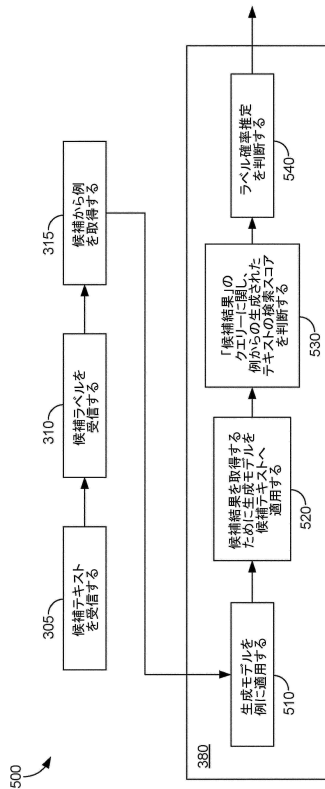
【図 3】



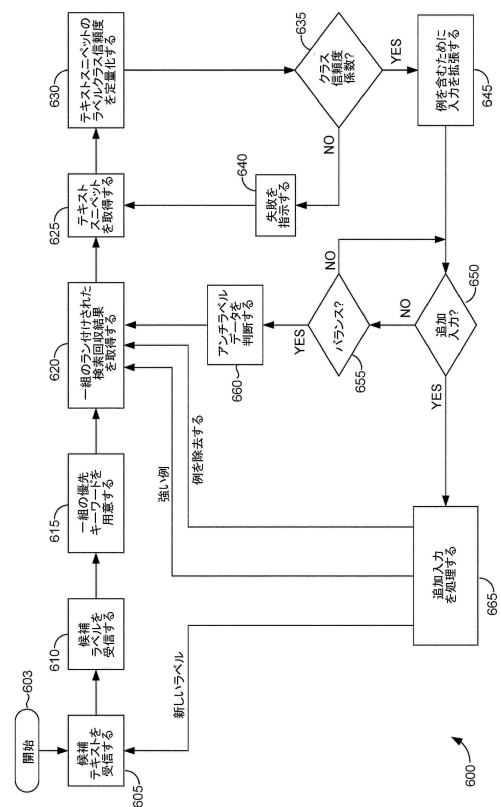
【図 4】



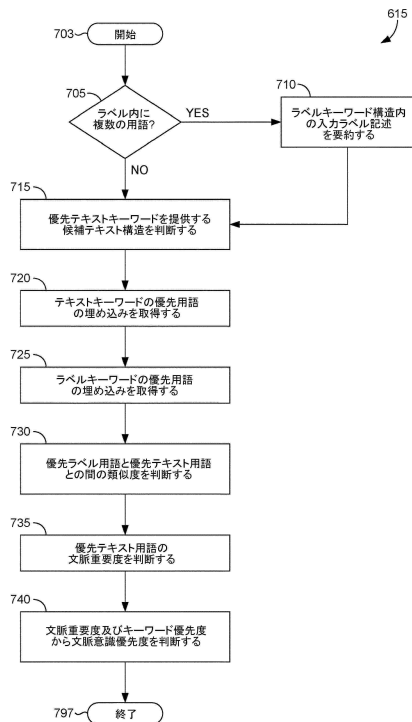
【圖 5】



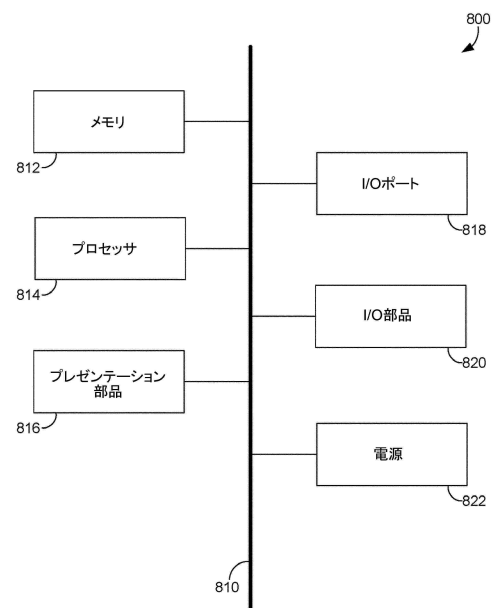
【圖 6】



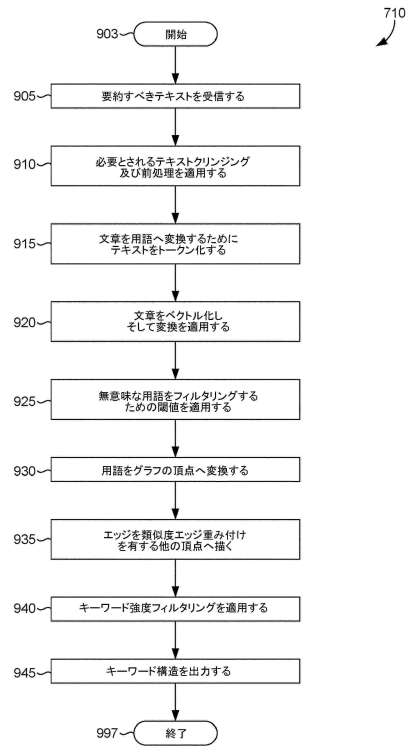
【図 7】



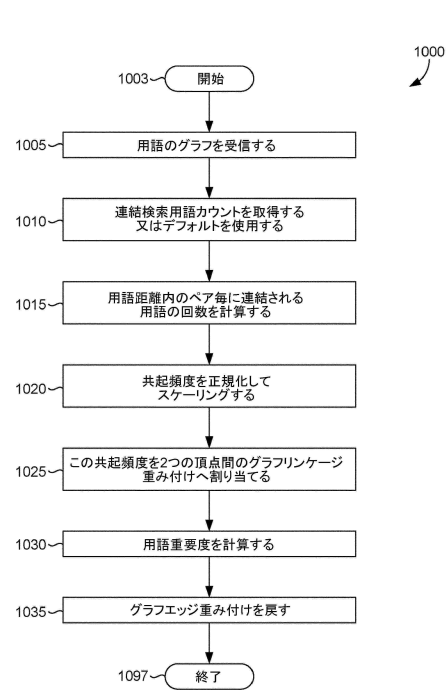
【図 8】



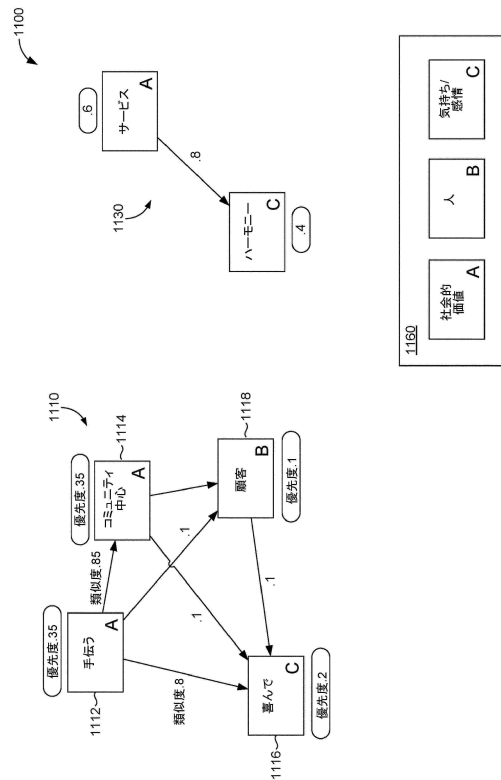
【図 9】



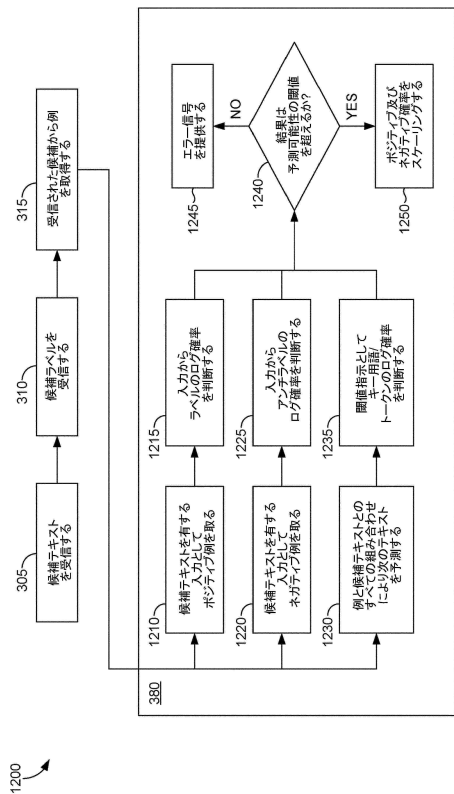
【図 10】



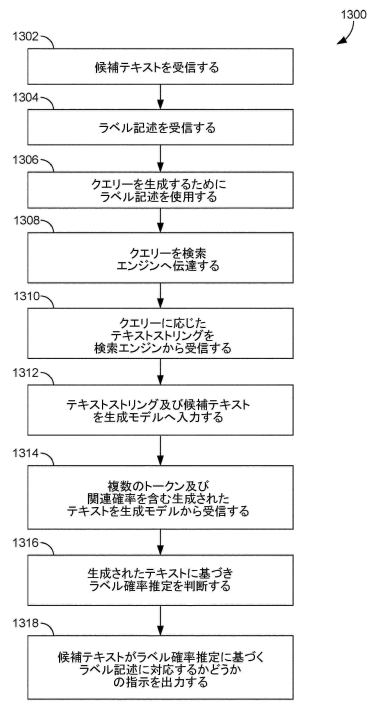
【図 11】



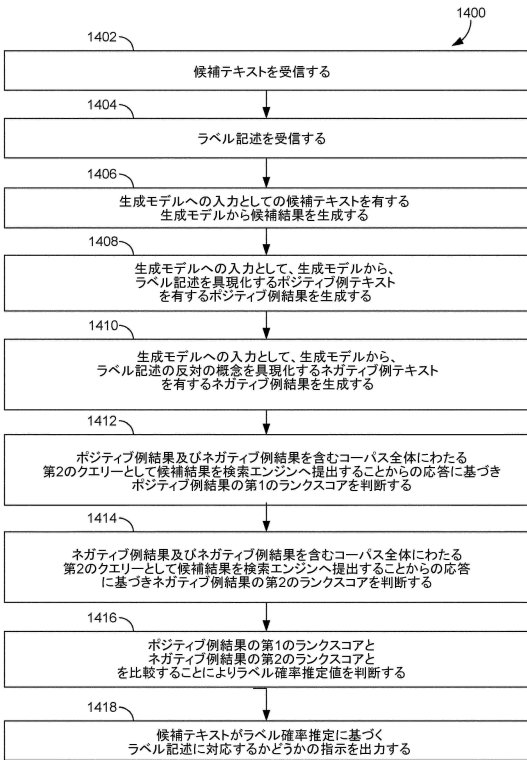
【図 12】



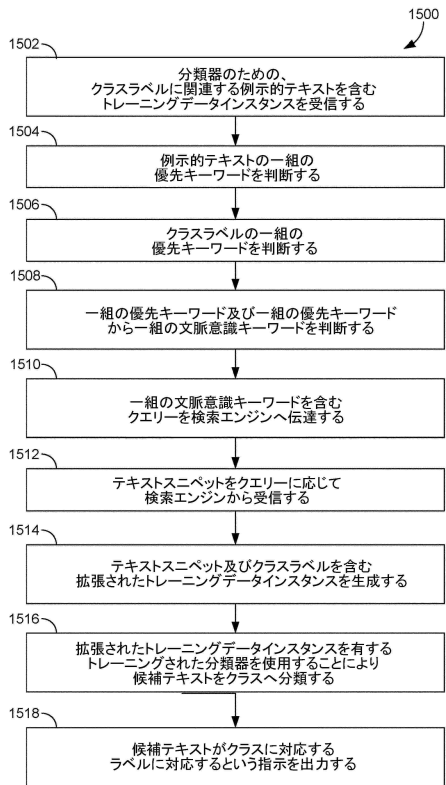
【図 1 3】



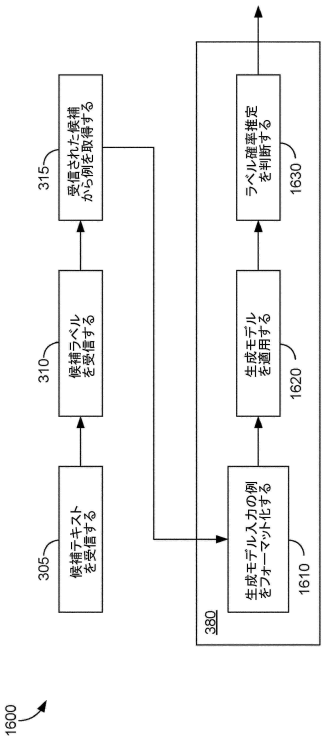
【図 1 4】



【図 1 5】



【図 1 6】



10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/030464

A. CLASSIFICATION OF SUBJECT MATTER INV. G06F16/33 G06F16/35 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) G06F		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018/052928 A1 (LIU MINGKUAN [US] ET AL) 22 February 2018 (2018-02-22) paragraph [0134] -----	1-12
A	NABARUN MONDAL ET AL: "Supervised Text Classification using Text Search", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 14 November 2020 (2020-11-14), XP081823872, page 2, paragraph 9 - page 3, last paragraph -----	1-12
☐ 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 25 August 2022		Date of mailing of the international search report 02/09/2022
Name and mailing address of the ISA/ European Patent Office, P.B. 5618 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Herri, Edmond

Form PCT/ISA/210 (second sheet) (April 2005)

International application No

PCT/US2022/030464

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2018052928 A1	22-02-2018	US 2018052928 A1	22-02-2018
		WO 2018035139 A1	22-02-2018

40

フロントページの続き

,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,D
K,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),O
A(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,B
B,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,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,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,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,Z
A,ZM,ZW

弁理士 阿部 豊隆

- (72)発明者 セワク, モヒト
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 ボルリ, ラヴィ キラン レディ
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 ブラム, ウィリアム
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 チャン, パク オン
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 リー, ウェイシェン
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 アーチャーリヤ, シャラダ シリシュ
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 ラドニック, クリスチャン
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 ベトサー, マイケル アブラハム
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 ドリニック, ミレンコ
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ
(72)発明者 リウ, シホン
アメリカ合衆国, ワシントン州 98052 レッドモンド ワン マイクロソフト ウェイ

F ターム (参考) 5B175 FB02