

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2025-515316
(P2025-515316A)

(43)公表日 令和7年5月14日(2025.5.14)

(51)国際特許分類

F I

G 0 6 F 40/30 (2020.01)

G 0 6 F 40/35 (2020.01)

G 0 6 F 40/30

G 0 6 F 40/35

審査請求 未請求 予備審査請求 未請求 (全33頁)

(21)出願番号	特願2024-563073(P2024-563073)	(71)出願人	524389858
(86)(22)出願日	令和5年4月24日(2023.4.24)		ギャン, インコーポレイテッド (ア デ
(85)翻訳文提出日	令和6年12月20日(2024.12.20)		ラウェア コーポレーション)
(86)国際出願番号	PCT/US2023/066123		アメリカ合衆国 マサチューセッツ州 0
(87)国際公開番号	WO2023/212524		2 0 6 2, ノーウッド, セントラル ス
(87)国際公開日	令和5年11月2日(2023.11.2)		トリート 1
(31)優先権主張番号	63/334,340	(74)代理人	110000659
(32)優先日	令和4年4月25日(2022.4.25)		弁理士法人広江アソシエイツ特許事務所
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	スリニヴァサン, ヴェンカテサン
(31)優先権主張番号	63/497,731		アメリカ合衆国 マサチューセッツ州 0
(32)優先日	令和5年4月23日(2023.4.23)		2 4 9 3, ウェストン, カッターズ ブ
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	ラフ 1 1
(81)指定国・地域	AP(BW,CV,GH,GM,KE,LR,LS,MW,MZ		ジャタフ, ヴィシャル
	最終頁に続く		インド マハラシュトラ州 4 1 2 2 0
			7, プネー, ワゴリ, カラディー, パン
			最終頁に続く

(54)【発明の名称】 説明可能な自然言語理解プラットフォーム

(57)【要約】

リアルタイムの人間らしい自然言語理解エンジンのためのシステムおよび方法（NLUプラットフォーム）。このプラットフォームは、自然言語コンテンツをその完全な構成コンテキストで理解し、完全に説明可能であり、トレーニングデータがまったくないか、またはほとんどなくても効果的である。本発明は、自然言語の機械表現を利用して、機械による人間らしい理解を可能にする。本発明は、知識ベースの言語学、談話モデル、構成性、および言語における修辭構造を組み合わせたものである。それは、コンセプトに関するグローバル知識を取り込むことで、その表現と理解を向上させることができる。このプラットフォームは、統計的に導出された分布意味論に依存せず、世界中のあらゆる自然言語理解タスクのために使用され得る。1つの例示的な用途では、NLUプラットフォームは、知識の自動取得に使用したり、インターネットなどの1つ以上のドキュメントリポジトリからの任意のトピックまたはトピックセットのための自動調査エンジンとして使用したりできる。

【選択図】 なし

【特許請求の範囲】

【請求項 1】

1つ以上のテキストドキュメントから自然言語コンテンツの人間らしい自然言語理解のタスクを自動的に実行するように構成された少なくとも1つのプロセッサを含む自然言語理解システムであって、前記システムが、

構成言語モデルを使用してドキュメントの意味の機械表現を作成する自然言語理解エンジンをホストするコンピュータシステムを含み、

前記エンジンが、前記ドキュメント全体をその構成構造に分解して、複数の意味の層を反映することにより、ドキュメント用のその機械意味表現を作成し、かつ可逆的な方法で前記機械意味表現を作成し、

前記エンジンが、前記ドキュメントに対して識別された談話モデルに従って前記ドキュメントを解析し、前記ドキュメントの主要なアイデアを識別し、前記ドキュメント内の新たなアイデアの始まりを識別し、アイデアによって前記ドキュメントをサブドキュメントに分解し、前記サブドキュメントをその構成要素に分解することによって前記ドキュメント全体の意味表現を作成することにより、前記ドキュメントをその構成構造に分解し、

前記コンピューティングシステムが、1つ以上の仮想サーバーまたは専用サーバー、あるいは同様のコンピューティングデバイスを含み、実行可能な命令でプログラムされている、システム。

【請求項 2】

前記構成言語モデルが、その意味表現グラフを作成するために、統計的機械学習、または単語埋め込みのような統計的に導出された分布意味論を利用しない、請求項 1 に記載のシステム。

【請求項 3】

前記自然言語理解エンジンが、計算言語学、言語の構成性および修辞構造を利用し、いずれのトレーニングまたはトレーニングデータもなしで1つ以上のドキュメントを処理できる、請求項 1 に記載のシステム。

【請求項 4】

前記自然言語理解エンジンが、完全に自然言語で動作し、前記ドキュメントのいずれの部分も実数値のベクトルエンコーディングに変換しない、請求項 1 に記載のシステム。

【請求項 5】

プログラミングなしで新たな談話モデルが作成され、統合される、請求項 1 に記載のシステム。

【請求項 6】

前記ドキュメントの前記主要なアイデアが、言語構造および修辞構造、属性、ならびに談話モデルの組み合わせを使用して、コンセプトの重要性に基づいて決定される、請求項 1 に記載のシステム。

【請求項 7】

前記新たなアイデアの始まりは、複数の言語規則、修辞関係、共参照関係を含む構成要素を結び付ける言語属性と修辞表現の性質に基づいて、段落、文、節を含む各連続する構成要素間の関係の性質を決定することによって決定され得る、請求項 1 に記載のシステム。

【請求項 8】

前記サブドキュメントを分解することは、第 1 のレベルで、前記サブドキュメントを構成する文レベルの構成要素とそれらの間の前記関係を、言語属性と、前記構成要素と前記サブドキュメントの主文および構成要素同士との言語関係の性質を使用して識別することを含む、請求項 1 に記載のシステム。

【請求項 9】

次のレベルで前記サブドキュメントを分解することは、言語属性と規則および修辞属性と規則を使用して、文レベルの構成要素内とそれらの間の前記修辞関係および言語関係の前記性質を識別することを含む、請求項 8 に記載のシステム。

10

20

30

40

50

【請求項 10】

前記分解が、単語レベルを含む各構成単位が前記意味表現グラフの一部になるまで継続される、請求項 9 に記載のシステム。

【請求項 11】

いずれのプログラミングもなしで構成要素間の関係を検出するための新たな規則が追加され得る、請求項 8 に記載のシステム。

【請求項 12】

前記意味表現が、コンセプト、前記修辭表現の表面形態または抽象的な修辭関係の種類によってクエリーできる前記分解された構成要素を含む多層グラフの形をとる、請求項 1 に記載のシステム。

10

【請求項 13】

前記意味表現グラフが、可逆的に完全なドキュメントに戻ることができる、請求項 12 に記載のシステム。

【請求項 14】

前記構成構造を分解することが、単語間の共参照関係を確立することを含み、そのような共参照関係の前記検出が、グローバル知識ストアに保存された外部知識によって改善される、請求項 1 に記載のシステム。

【請求項 15】

前記グローバル知識ストアが、各コンセプトについてセンスレベルで知識を保存し、いずれのプログラミングもなしで継続的に改善することができる、請求項 14 に記載のシステム。

20

【請求項 16】

前記グローバル知識ストアが、各コンセプトのためにセンスレベルで知識を保存することができる、請求項 14 に記載のシステム。

【請求項 17】

ドキュメントが、任意の談話モデルに基づく、テキスト、html、pdfを含む、任意のフォーマットにおける任意の長さのテキストドキュメント、およびテキストに転写されたビデオであり得る、請求項 1 に記載のシステム。

【請求項 18】

ドキュメントが、文法的に完全ではなく、不可解なソーシャルメディアメッセージを含む、請求項 1 に記載のシステム。

30

【請求項 19】

前記エンジンが、文間および各文内の、主語と目的語との間を含む構成単位間の修辭句を識別し、また、このような識別された修辭句を正規化された修辭関係のセットにマッピングする、請求項 1 に記載のシステム。

【請求項 20】

修辭表現と抽象的な修辭関係との間の前記マッピングが、構成可能であり、また、手がかり語、手がかり句、および品詞、文型、共参照関係などの前記談話の言語属性の機能であり得る、請求項 12 に記載のシステム。

【請求項 21】

前記エンジンが、前記マルチレベルの意味表現グラフに対応するドキュメントの拡張可能なマルチレベルの概要を作成する、請求項 1 に記載のシステム。

40

【請求項 22】

前記エンジンが、前記ドキュメント内で識別された前記修辭関係に基づいて動作する規則として表現されたユーザ選択に従って、ドキュメントの拡張可能なマルチレベルの概要を作成する、請求項 1 に記載のシステム。

【請求項 23】

前記エンジンが、いずれのトレーニングまたはトレーニングドキュメントのセットも必要とせずに、1つ以上のドキュメントを処理することができる、請求項 1 に記載のシステム。

50

【請求項 24】

請求項1のシステムの前記意味表現を使用して、関連ドキュメントの知識コレクションを作成できる、システム。

【請求項 25】

前記知識コレクションが、インターネット検索エンジン、プライベートドキュメントコレクション、およびソース、またはそれらの組み合わせを使用したクエリーに基づき得る、請求項24に記載のシステム。

【請求項 26】

前記エンジンが、各ドキュメントに最も関連性の高い一致を決定することにより、関連ドキュメントの前記コレクションを事前に決定されたサブトピックのセットに整理し、

事前に決定されたサブトピックのセットが使用不能である場合、前記エンジンが、請求項1の前記システムの前記意味表現を使用して、前記コレクション内の前記ドキュメントのそれぞれのために、関連ドキュメントの前記コレクションからトピックの固有のセットを表面化させる、請求項25に記載のシステム。

【請求項 27】

サブトピックの前記固有のセットが、前記ドキュメントのそれぞれにおける前記アイデアを集約することにより、ドキュメントの前記コレクション全体で最も代表的なアイデアとして決定される、請求項26に記載のシステム。

【請求項 28】

前記ドキュメントコレクション全体で最も代表的なアイデアを決定するプロセスが、前記グローバル知識ストアからの知識を統合して、前記エンジンが前記ドキュメントに明示的に存在しないアイデア間の関係を認識できるようにし、それが見つかったドキュメントにおけるその役割を含む言語特性を使用する、請求項27に記載のシステム。

【請求項 29】

前記ドキュメントコレクション内の各記事と事前に定義された各前記サブトピックとの関連度が、コンセプトの重要性、文の重要性およびコンテキストの強さの組み合わせを使用して決定される、請求項26に記載のシステム。

【請求項 30】

前記エンジンが、前記ドキュメントレベルで意味表現グラフ全体の冗長なコンセプトを削除した後、各前記ドキュメント用の意味表現グラフを集約して、ドキュメントの前記コレクション全体の組み合わせられた意味表現グラフを作成する、請求項24に記載のシステム。

【請求項 31】

前記エンジンが、前記ドキュメントの各々についての概要を集約し、冗長なコンテンツを削除することにより、ドキュメントの前記コレクション全体の集約概要を作成する、請求項24に記載のシステム。

【請求項 32】

前記エンジンが、そのサブトピックに分類された前記ドキュメントの概要を集約することにより、前記サブトピックの各々についての集約概要を作成する、請求項24に記載のシステム。

【請求項 33】

2つの文中のコンセプト間の前記共参照関係が、まったく同じコンセプトの前記使用、または外部のグローバル知識ストアに含まれる前記コンセプトへの代名詞的または実用的な参照であり得る、請求項1に記載のシステム。

【請求項 34】

1つ以上のテキストドキュメントから自然言語コンテンツの人間らしい自然言語理解機能を自動的に実行する方法であって、前記方法が、

構成言語モデルを使用して、ドキュメントの意味の機械表現を生成することと、

前記ドキュメント全体をその構成構造に分解して、複数の意味の層を反映することにより、ドキュメント用の機械意味表現を生成し、かつ可逆的な方法で機械意味表現を生成す

10

20

30

40

50

ることと、

前記ドキュメントに対して識別された談話モデルに従って前記ドキュメントを解析し、前記ドキュメントの主要なアイデアを識別し、前記ドキュメント内の新たなアイデアの始まりを識別し、アイデアによって前記ドキュメントをサブドキュメントに分解し、前記サブドキュメントをその構成要素に分解することによって前記ドキュメント全体の意味表現を作成することにより、前記ドキュメントをその構成構造に分解することと、を含む、方法。

【請求項 3 5】

前記構成言語モデルが、その意味表現グラフを作成するために、統計的機械学習、または単語埋め込みのような統計的に導出された分布意味論を利用しない、請求項 3 4 に記載の方法。

10

【請求項 3 6】

前記構成言語モデルが、計算言語学および言語の修辭構造を利用し、いずれのトレーニングまたはトレーニングデータもなしで 1 つ以上のドキュメントを処理できる、請求項 3 4 に記載の方法。

【請求項 3 7】

前記方法が、完全に自然言語で動作し、前記ドキュメントのいずれの部分も実数値のベクトルエンコーディングに変換しない、請求項 3 4 に記載の方法。

【請求項 3 8】

プログラミングなしで新たな談話モデルが作成され、統合される、請求項 3 4 に記載の方法。

20

【請求項 3 9】

前記ドキュメントの前記主要なアイデアが、言語構造および修辭構造、属性、および談話モデルの組み合わせを使用して、コンセプトの重要性に基づいて決定される、請求項 3 4 に記載の方法。

【請求項 4 0】

前記新たなアイデアの始まりが、複数の言語規則、修辭関係、および共参照関係を含む構成要素を結び付ける言語属性および修辭表現の性質に基づいて、段落、文、節を含む各連続する構成要素間の関係の性質を決定することによって決定され得る、請求項 3 4 に記載の方法。

30

【請求項 4 1】

前記サブドキュメントを分解することが、第 1 のレベルで、前記サブドキュメントを構成する文レベルの構成要素とそれらの間の前記関係を、言語属性と、前記構成要素と前記サブドキュメントの主文および構成要素同士との言語関係の性質を使用して識別することを含む、請求項 3 4 に記載の方法。

【請求項 4 2】

次のレベルで前記サブドキュメントを分解することが、言語属性と規則および修辭属性と規則を使用して、文レベルの構成要素内とそれらの間の前記修辭関係および言語関係の前記性質を識別することを含む、請求項 4 1 に記載の方法。

【請求項 4 3】

前記分解が、単語レベルを含む各構成単位が前記意味表現グラフの一部になるまで継続される、請求項 4 2 に記載の方法。

40

【請求項 4 4】

いずれのプログラミングもなしで構成要素間の関係を検出するための新たな規則が追加され得る、請求項 4 1 に記載の方法。

【請求項 4 5】

前記意味表現が、コンセプト、前記修辭表現の表面形態または抽象的な修辭関係の種類によってクエリーできる、前記分解された構成要素を含む多層グラフの形をとる、請求項 3 4 に記載の方法。

【請求項 4 6】

50

前記構成構造を分解することが、単語間の共参照関係を確立することを含み、そのような共参照関係の検出が、前記グローバル知識ストアに保存された外部知識によって改善される、請求項 3 4 に記載の方法。

【請求項 4 7】

前記グローバル知識ストアが、各コンセプトについてセンスレベルで知識を保存し、いずれのプログラミングもなしで継続的に改善することができる、請求項 4 6 に記載の方法。

【請求項 4 8】

ドキュメントが、文法的に完全ではなく、不可解なソーシャルメディアメッセージを含む、請求項 3 4 に記載の方法。

【請求項 5 0】

修辞表現と抽象的な修辞関係との間の前記マッピングが、構成可能であり、また、手がかり語、手がかり句、および品詞、文型、共参照関係などの前記談話の言語属性の機能であり得る、請求項 4 5 に記載の方法。

【請求項 5 1】

前記方法が、前記マルチレベルの意味表現グラフに対応するドキュメントの拡張可能なマルチレベルの概要を作成する、請求項 3 4 に記載の方法。

【請求項 5 2】

それが、前記ドキュメント内で識別された前記修辞関係に基づいて動作する規則として表現されたユーザ選択に従って、ドキュメントの拡張可能なマルチレベルの概要を作成する、請求項 1 に記載の方法。

【請求項 5 3】

前記エンジンが、いずれのトレーニングまたはトレーニングドキュメントのセットも必要とせずに、1 つ以上のドキュメントを処理することができる、請求項 1 に記載のシステム。

【請求項 5 4】

請求項 3 4 の方法の前記意味表現を使用して、関連ドキュメントの知識コレクションを作成できる、方法。

【請求項 5 5】

前記知識コレクションが、インターネット検索エンジン、プライベートドキュメントコレクション、およびソース、またはそれらの組み合わせを使用したクエリーに基づき得る、請求項 5 4 に記載の方法。

【請求項 5 6】

前記方法が、各ドキュメントに最も関連性の高い一致を決定することにより、関連ドキュメントの前記コレクションを事前に決定されたサブトピックのセットに整理し、

事前に決定されたサブトピックのセットが使用不能である場合、前記方法が、請求項 3 4 の前記方法の前記意味表現を使用して、前記コレクション内の前記ドキュメントのそれぞれについて、関連ドキュメントの前記コレクションからトピックの固有のセットを表面化させる、請求項 5 5 に記載の方法。

【請求項 5 7】

サブトピックの前記固有のセットが、前記ドキュメントのそれぞれにおける前記アイデアを集約することにより、ドキュメントの前記コレクション全体で最も代表的なアイデアとして決定される、請求項 5 6 に記載の方法。

【請求項 5 8】

前記ドキュメントコレクション全体で最も代表的なアイデアを決定する前記プロセスが、前記グローバル知識ストアからの知識も統合して、前記方法が前記ドキュメントに明示的に存在しないアイデア間の関係を認識できるようにする、請求項 5 7 に記載の方法。

【請求項 5 9】

前記ドキュメントコレクション内の各記事と、事前に定義された各前記サブトピックとの関連度が、コンセプトの重要性、文の重要性およびコンテキストの強さの組み合わせを

10

20

30

40

50

使用して決定される、請求項 5 6 に記載の方法。

【請求項 6 0】

前記方法が、前記ドキュメントレベルで意味表現グラフ全体の冗長なコンセプトを削除した後、前記ドキュメントのそれぞれについての意味表現グラフを集約して、ドキュメントの前記コレクション全体の組み合わせされた意味表現グラフを作成する、請求項 3 4 に記載の方法。

【請求項 6 1】

前記方法が、前記ドキュメントのそれぞれについての概要を集約し、冗長なコンテンツを削除することにより、ドキュメントの前記コレクション全体の集約概要を作成する、請求項 3 4 に記載の方法。

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

本発明は、自然言語理解、計算言語学、意味の機械表現に関する。具体的には、自動化自然言語理解タスクの全範囲における自然言語データへの人間らしい理解のための方法、コンピュータプログラムおよびシステムに関する。

【背景技術】

【0 0 0 2】

自然言語コンテンツは、パブリックインターネット上とプライベートコンテンツリポジトリ内の両方で、コンテンツの大部分を占めている。したがって、当然のことながら、自然言語コンテンツの自動機械処理に対するニーズと関心はかなり高まっている。自然言語の機械処理に現在利用可能なほぼすべての技術は、各単語の実数値のベクトル表現に基づくパターンを含む単語パターンの大規模なニューラル言語モデルに依存している。このようなアプローチは、テキストをその完全な構成コンテキストで理解するには設計されていない。人間は、テキストをその完全な構成コンテキストで理解する。

【0 0 0 3】

現在のすべての言語モデルは、単語のシーケンスの確率を予測することを学習する。生成された単語のシーケンスは、ユーザによるクエリーに関連付けられる。最も確率の高い次の単語がモデルによって出力として選択される。これらのモデルは、非常に大量の生のテキストでトレーニングされる。2つのカテゴリの統計手法が、言語モデルの構築のために使用されてきた。1つは隠れマルコフモデルを使用し、もう1つはニューラルネットワークを使用する。どちらも、予測のために自然言語テキスト（入力データ）の大規模なコーパスに依存する統計的アプローチである。入力データに関しては、単語の袋（bag of words）アプローチが、非常に人気がある。しかし、単語の袋モデルは、そのシンプルさゆえに好ましいが、任意のコンテキストを保持せず、異なる単語は、どのように使用されるかに関係なく、毎回異なる表現を持つ。これにより、類似した単語に対して類似した表現を作成しようとする、単語埋め込みと呼ばれる新たな技法が開発された。「単語は一緒にいる人によって知られる（A word is known by the company it keeps）」という有名な格言が、単語埋め込みの動機である。単語埋め込みの支持者は、これが単語の意味およびその使用法を捉えることに似ていると考えている。

【0 0 0 4】

単語埋め込みには、主にそのような埋め込みのコーパス依存の性質と、単語が使用される異なるセンスを処理できないことに関係するいくつかの重大な制限がある。類似語を決定するためにコーパスに基づいて学習された単語表現は、テキスト全体の意味の人間らしい機械表現と同等ではない。

【0 0 0 5】

現在の言語モデルには、構成コンテキストを保持できず、説明可能性が欠如し、ニューラル言語モデルと組み合わせるタスク固有の大規模なデータが必要であるなど、重大な制限がある。さらに、現在の言語モデルは、ある分野から次の分野に知識を「学習」または

10

20

30

40

50

「転送」する能力において、普遍的とは程遠いものである。

【0006】

より最近では、特にコンテキストの保持やコンテキスト固有のグローバル知識など、いくつかの制限に対処するために、外部知識をこれらのモデルに統合する傾向がある。これらの強化により、言語モデルのパフォーマンスは向上するが、依然として単語レベルのパターンに基づいている。それらは、構成の完全な文脈的理解、説明可能性および大規模なタスク固有のデータの必要性といった基本的な問題に対処していない。現実の世界では、ほとんどの自然言語処理または理解タスクでは、大量のデータが不足していることがよくある。

【0007】

機械による自然言語理解の必要性を示す例として、ほとんどの知識労働者が日常的に行う一般的なタスクである「調査」について考えるとする。今日において、世界にはおよそ10億人の知識労働者が存在し、その数は増加していると推定されている。知識労働者の例は、学生、教師、学術研究者、アナリスト、コンサルタント、科学者、弁護士、マーケティング担当者などを含む。今日彼らが行っている調査または知識獲得の作業は、インターネット上の検索エンジンの使用、および社内またはプライベートのドキュメントコレクションに対する検索技術の使用を伴う。検索エンジンは、これまでアクセスできなかった情報へのアクセスを提供することで大きな価値を提供してきたが、それは調査の出発点にすぎない。

【0008】

今日の検索エンジンは依然としてドキュメント内のテキストを完全に理解できないため、結果セットには偽陽性と偽陰性の両方が存在する。クエリーに基づいて、1ページに10項目を表示する設定では、関連性のない結果が1ページ目という早さから表示され得、それ以降のページでも確実に表示される。同様に、関連性の高い結果が、10～20ページ目のように、かなり後の方にリストされ得る。

【0009】

さらに重要なのは、今日の技術では、残りの工程、すなわち、関連性を決定し、各ドキュメントの知識を整理、読み取り、理解、および統合し、すべてのドキュメントにわたる集約的な理解を深めるという作業には何の支援も提供されていないことである。現在の検索技術または言語モデルは、これらのドキュメント内のテキスト全体を理解できないため、完全な調査または知識獲得プロセスを自動的に支援することはできない。

【0010】

非構造化コンテンツを完全に理解する必要がある他の例示的なユースケースとしては、非構造化コンテンツのタグ付け、盗用または重複を検出するために複数のドキュメントを相互に比較すること、インテリジェントな自然言語出力の生成を自動化することなどがある。誤解を避けるために、非構造化コンテンツとは、テキストに転写できるビデオまたはオーディオデータであり得る。

【0011】

したがって、説明可能で扱いやすい方法で、またデータがまばらな状況であっても、テキスト談話の完全な構成コンテキストにおける人間らしい理解を自動化する方法およびシステムが必要であることは明らかである。

【発明の概要】

【0012】

リアルタイムの人間らしい自然言語理解エンジンのためのシステムおよび方法（NLUプラットフォーム）が開示されている。このNLUプラットフォームは、自然言語コンテンツをその完全な構成コンテキストで理解し、完全に説明可能であり、トレーニングデータがまったくないか、またはほとんどなくても効果的である。本発明は、機械による人間らしい理解を可能にする自然言語の機械表現の新規な方法を開示する。本発明は、知識ベースの言語学、談話モデル、構成性、および言語における修辞構造を組み合わせたものである。それは、コンセプトに関するグローバル知識を取り込むことで、その表現と理解を

10

20

30

40

50

向上させる。本発明は、統計的に導出された分布意味論を利用しない。開示された発明は、普遍的に実行される自然言語理解タスクの全範囲に使用可能である。

【0013】

1つの例示的な用途では、NLUプラットフォームは、知識の自動取得に使用したり、インターネットなどの1つ以上のドキュメントリポジトリからの任意のトピックまたはトピックセットの自動調査エンジンとして使用したりする。これは、各項目の関連度を決定することと、関連項目を目次に整理することまたは各項目のその人間らしい理解から目次を自動的に導出することと、各関連項目のマルチレベルの概要を作成することと、各項目のマルチレベル意味グラフを作成することと、各項目の主要なトピックを識別することとを含む。複数のコンテンツ項目のコンテンツを統合して、統合されたクロスドキュメントのマルチレベルの意味表現、および複数のドキュメントにわたるマルチレベルの統合された概要を作成する。NLUプラットフォームは、主要なトピックなど、最初に発見された知識のあらゆる態様について、継続的に新たな知識を自動的に発見する。

10

【0014】

本発明の1つ以上の態様は、明細書の最後にある特許請求の範囲において、例として具体的に指摘され、明確に主張されている。本発明の上記およびその他の目的、特徴、および利点は、添付の図面と併せて以下の詳細な説明から明らかである。

【図面の簡単な説明】

【0015】

【図1】インターネット上の一般的な検索エンジンを使用して10個の検索クエリーの関連性を分析した結果を示す。

20

【図2】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム高レベルアーキテクチャを示す。

【図3】本明細書に記載の1つ以上の実施形態による、NLUプラットフォームパイプラインを示す。

【図4】本明細書に記載の1つ以上の実施形態による、例示的なNLUプラットフォーム意味表現グラフメイン層を示す。

【図5】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム意味表現グラフ文レベルを示す。

【図6】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム意味表現グラフコンセプトレベル層を示す。

30

【図7】本明細書に記載の1つ以上の実施形態による、いくつかの談話モデルを示す。

【図8A】ニュース記事の例を示す。

【図8B】ニュース記事の例を示す。

【図8C】ニュース記事の例を示す。

【図8D】ニュース記事の例を示す。

【図9A】NLUプラットフォーム200による、ドキュメントと関心のあるトピックとの関連度の決定を示す。

【図9B】NLUプラットフォーム200による、ドキュメントと関心のあるトピックとの関連度の決定を示す。

40

【図10】図2で報告された同じ10個のクエリーに対するNLUプラットフォーム200からの結果の分析を示す。注：コンセプト検出（ドキュメントのSKGでのコンセプトの出現）の後、複雑なコンセプト（複数のトピック）については、文の重要性（主要なアイデア、主要なクラスターなど）とコンセプトの重要性（主題など）の他に、NLUプラットフォーム200はコンテキストの強さも決定する。コンテキストの強さは、ドキュメントが複雑なコンセプト内の複数のコンセプトを互いのコンテキストでどれだけ適切に参照しているかを決定する。

【図11】本明細書に記載の1つ以上の実施形態による、図3のニュース記事の例の部分的なNLUプラットフォーム意味表現グラフコンセプト層を示す。

【図12】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム200

50

によって作成された図9のドキュメント用の例示的な概要を示す。

【図13】NLUプラットフォーム200を使用した、電気自動車のトピックに関する例示的な知識コレクションを示す。

【図14】本明細書に記載の1つ以上の実施形態による、NLUプラットフォームが任意のトピックに関する新たな知識を積極的に発見し、既存の意味表現グラフと統合する例示的なリアルタイム知識発見プロセスを示す。

【図15】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム200が、複数のドキュメントにわたるトピックを組み合わせ、ドキュメントのコレクションにわたるトピックの作成を示す。

【図16】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム200が、冗長な意味を排除することで個々のドキュメントの意味表現構造を統合した、例示的な集約意味表現の例を示す。

【図17】本明細書に記載の1つ以上の実施形態による、関連性または概要を決定する際にNLUプラットフォーム200が考慮するトピックに関するグローバル知識をユーザが入力できる方法を示す。

【図18】本明細書に記載の1つ以上の実施形態による、NLUプラットフォーム200による検討のために、ユーザがどのように異なる単語の特定の語義を識別するための記号規則を入力できるかを示す。

【図19】任意の機関が新たなトピックまたは特定のスキルに関連するトピックに関する学習コンテンツを迅速に組み立てることができる、ACLプラットフォームI学習コンテンツオーサリング／アセンブリ構成要素を示す。

【図20】ACLプラットフォームIのパイロットの結果を示す。このとおり、10人の生徒全員のピアグレードは満点であった。

【発明を実施するための形態】

【0016】

背景技術

自然言語コンテンツの自動機械処理の使用例は世界中に無数にある。自然言語の機械処理に現在利用可能なほぼすべての技術は、各単語の意味表現に基づくパターンを含む単語パターンのニューラル言語モデルに依存している。このアプローチは、テキストをその完全な構成コンテキストで理解できない。人間は、テキストをその完全な構成コンテキストで理解する。誤解を避けるために、非構造化コンテンツとは、処理のためにテキストに転写されたビデオまたはオーディオデータを指す。

【0017】

従来、「単語の袋」法は、自然言語処理(NLP)タスクの一般的なアプローチであった。単語の袋モデルは、分類を必要とするNLP問題でよく使用される。インターネット検索などのNLPタスクの大部分は分類問題である。単語の袋(Bow)モデルでは、テキスト(文またはドキュメントなど)は、文法および語順を無視して、単語の集まりとして表現される。ドキュメントのコレクションにわたって、各単語の出現頻度が分類子をトレーニングするための特徴として使用される。次に、分類子を新たなドキュメントに適用して、ドキュメントを一連のカテゴリに分類したり、トピックを識別したり、検索クエリー(例えば、検索エンジン用の検索クエリー)との一致を決定したりする。この方法はドキュメント内で出現する単語の頻度により焦点を当てているため、Bowモデルは単語がどこでどのように使用されたかというコンテキストを保持せず、処理中のドキュメントの意味を理解しない。

【0018】

一般的な「単語の袋」法の限界を克服するために、いくつかの大規模な事前トレーニング済み言語モデルが導入されている。2つのカテゴリの統計手法が、言語モデルの構築のために使用されてきた。1つは、隠れマルコフモデルなどの確率的アプローチを使用し、もう1つは、ニューラルネットワークを使用する。どちらも、予測のために自然言語テキストの大規模なコーパスに依存している。単語の袋モデルでは、使用方法に関係なく、異

10

20

30

40

50

なる単語は異なる表現を持つ。これにより、類似した単語に類似した表現を作成しようとする単語埋め込みと呼ばれる新たな手法が開発された。「単語は一緒にいる人によって知られる (A word is known by the company it keeps)」という有名な格言が、単語埋め込みの動機である。単語埋め込みの支持者は、これが単語の意味およびその使用法を捉えることに似ていると考えている。

【0019】

単語埋め込みには、主にそのような埋め込みのコーパス依存の性質と、単語が使用される異なるセンスを処理できないことに関係するいくつかの重大な制限がある。さらに、類似した単語を決定するためにコーパスに基づいて学習された単語表現は、完全なテキストの人間らしい意味表現にほとんど似ていない。

【0020】

言語モデルは、単語のシーケンスの確率を予測することを学習する。最も確率の高い次の単語が、モデルによって出力として選択される。これらのモデルは、非常に大量の生のテキストでトレーニングされる。先行する単語または他の単語特性を与えられた単語のシーケンスの確率は、明らかにテキストの意味を表すものではない。これらのモデルは、モデルのトレーニングに使用されるデータのすべてのバイアスを保持する。例えば、MarcusおよびDavid (2021) は、より強力な言語モデルの1つであるGPT3が、提供された文章の続きとして、以下の太字の斜体のテキストを出力することを発見した。

【0021】

「You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So you drink it. You are now dead.」

【0022】

また、MicrosoftがチャットボットTayのリリースでそれを発見したように、言語モデルに依存するチャットボットなどのNLPアプリケーションが、人種差別的または性差別的なコメントを吐き出すことがあることもよく知られた事実である。また、GPT2に「3人のイスラム教徒」と入力すると、彼らがテロリストまたは犯罪者として描写されているテキストが表示されることが多いことも指摘されている。これらの例は、トレーニングデータセットによって導入されるバイアスを示している。

【0023】

現在の言語モデルは、あるアプリケーションから次のアプリケーションに知識を「学習」または「転送」する能力において、普遍的とは程遠いものである。これらのモデルには、完全なコンテキストを保持できないこと、説明可能性が欠如していること、ニューラル言語モデルと組み合わせるタスク固有の大規模なデータが必要であること、およびデータからのバイアスなど、同じ重大な制限がある。

【0024】

最近、これらのモデルの制限のいくつかに対処するために、外部知識をモデルに統合する傾向がある。この研究では、大規模な知識強化LMを事前にトレーニングするために、何らかの形態の外部知識を統合することが一般的に提唱されている。これらの強化により、言語モデルのパフォーマンスは向上するが、自然言語テキストの説明可能な意味表現という根本的な問題にはまったく対処されていない。BenderおよびKoller (2017) は、言語モデルは原理的には意味の学習にはつながらないと主張している。彼らは、「意味」を人間らしい自然言語理解であると適切に定義し、これらの言語モデルが意味の理解に近いと主張するのは不誠実であると主張している。

【0025】

10

20

30

40

50

自然言語による談話では、必要に応じて修辞表現を使用して著者のアイデアおよび知識を構成することを伴う。良い談話は、首尾一貫しており、談話の構造とコンテンツに関して広く受け入れられている原則と規則に従う。より高いレベルでは、ニュース記事、ブログ、研究論文、法的ドキュメント、電子メール、ソーシャルメディアの投稿など、さまざまな種類の談話モデルがある。これらのそれぞれには、その分野の専門家が理解できる明確に定義された談話モデルがあり、例えば、弁護士は法的契約書の基礎となる談話モデルを理解し、私たちのほとんどは電子メールのモデルを理解し、ニュース記事は談話モデルに従っている。人間らしい意味表現モデルは、構成の基礎となる談話モデルを認識する必要がある。さらに、そのような表現は説明可能でなければならない。これは、現在のどの方法でも実現できない。

10

【0026】

機械による自然言語理解の必要性を示す例として、ほとんどの知識労働者が日常的に行う一般的なタスクである「調査」について考えるとする。今日において、世界にはおよそ10億人の知識労働者が存在し、その数は増加していると推定されている。知識労働者の例は、学生、教師、学術研究者、アナリスト、コンサルタント、科学者、弁護士、マーケティング担当者などを含む。知識労働者は、あらゆる分野に存在する。今日彼らが行っている調査または知識獲得の作業は、インターネット上の検索エンジンの使用、および社内またはプライベートドキュメントコレクションに対する検索技術の使用を含む。検索エンジンは、これまでアクセスできなかった情報へのアクセスを提供することで大きな価値を提供してきた。しかし、それらはアクセスを提供するだけであり、それが調査の出発点にすぎない。しかし、現在の検索エンジンは、インターネット上であろうとプライベートドキュメントリポジトリ上であろうと、アクセスを提供する場合でも、上で概説した制限のために、かなりの偽陽性を生成する。

20

【0027】

ユーザがGoogleなどの検索エンジンを使用してクエリーを実行すると、通常、多数のヒットの中から1～10件の項目が結果として表示される。今日の検索エンジンは依然としてドキュメント内のテキストを完全に理解できないため、結果セットには偽陽性と偽陰性の両方が存在する。クエリーに基づいて、1ページに10項目を表示する設定では、関連性のない結果が1ページ目という早さから表示され得、クエリーの性質に関係なく、それ以降のページにも確実に表示される。同様に、関連性の高い結果が、クエリーに対して10～20ページ目のようにリストのかなり後ろの方に表示される。今日の検索技術による結果を処理する場合、ユーザはまず各項目を開いて、どの検索結果が自分のクエリーに本当に関連しているかを決定する必要がある。

30

【0028】

実験では、MITの人気のオープンコースウェアのコンテンツコレクションで「データサイエンスと倫理」の検索が実行された。結果では、1ページ目の10件の結果のうちの5件を含め、80%以上で関連性がなかった。同様に、図1は、一般的な検索エンジンを使用してランダムに10個のクエリーを分析した結果を示す。選択されたクエリーは、1単語、2単語および3単語を有し、一部は、非常に曖昧な単語を有した。各クエリーについて、人間の専門家が検索エンジンからの結果セット全体の1,785件の記事すべてをレビューし、記事を関連性のあるものと関連性のないものに分類した。分析では、ホームページ、インデックスページ、オーディオ、ビデオページ、書籍、ショッピングページは関連性がないとみなされた。図1は、ブラウザのデフォルトページ表示による関連性のある結果と関連性のない結果を示す。10件のクエリーにわたって、Googleは関連性のある結果を59%しか返さなかった。以上のように、後のページには関連する結果がたくさんあり、その逆も同様である。ユーザは、関連性を決定するために、まずすべてのドキュメントを読む必要がある。

40

【0029】

検索エンジンの結果の不正確さのレベルは、最初の問題にすぎない。この工程は、調査タスクの「アクセス」工程である。現在の技術は、残りの工程を支援するものではない。

50

知識労働者が関連ドキュメントを識別して入手すると、次の工程は、効果的な手動処理を行うためにそれらを整理することである。これは、サブトピックに整理する形式をとることができる。次に、人間が各ドキュメントを読み、各ドキュメント内の知識を理解して統合し、すべてのドキュメントにわたる総合的な理解を深める必要がある。さらに、知識の獲得は一度きりではないため、研究成果はトピックに関する最新の知識で継続的に更新する必要がある。現在の検索技術では、非構造化コンテンツからの完全な調査または知識獲得プロセスを自動的に支援することはできない。

【0030】

テキスト概要は、ドキュメントのコンテンツを完全に理解する必要があるもう1つのNLUタスクである。これは、上記の調査ワークフローにおける統合タスクになる。

10

【0031】

インターネット上やプライベートコーパスに蓄積されるテキストの量は増加しており、それを有用なものにするためにはまとめる必要がある。当然ながら、そのニーズとは、意味と一貫性を失うことなく、テキストを自動的にまとめることである。自動テキスト概要とは、元のテキストドキュメントの意味を維持しながら、人間の助けを借りずに簡潔で流暢な概要を作成するタスクである。

【0032】

テキスト概要には、抽出型と抽象型の2種類がある。抽出テキスト概要においては、概要はコンテンツの元の文章が保持される。抽象概要の場合、人間が自身のグローバル知識を使用してどのようにドキュメントをまとめるかを再現しようと試みる。現在の技術では、抽出概要は部分的にしか成功しておらず、抽象概要は非常に不十分である。効果的な抽出概要は、知識労働者にとって非常に役立つ。

20

【0033】

トピックモデリングは、もう1つの非常に一般的なNLUタスクである。現在の技術は、ドキュメントのコンテンツをその完全な構成コンテキストで理解するように設計されないため、ドキュメント内のトピックを識別するには不十分である。一般的なデジタル出版サイトでは、ユーザまたは購読者が関心のある記事およびコンテンツを見つけやすくしたいと考えていた。組織は、コンテンツドキュメントを分類するトピックオントロジーを定義した。一般的な言語モデルでは、正しい分類はわずか20%しか達成されなかった。このモデルは説明不可能であり、非常に大規模なコーパスでトレーニングされている。明らかに、そのトレーニングデータはサイトのコンテンツを分類するのに十分ではない可能性があり、そのモデルはトピックを識別する前に各ドキュメントの完全なコンテンツを理解することに重点を置いていない。

30

【0034】

非構造化コンテンツを完全に理解する必要がある他の例示的なユースケースとしては、非構造化コンテンツのタグ付け、盗用または重複を検出するために複数のドキュメントを相互に比較すること、インテリジェントな自然言語出力の生成を自動化することなどがある。

【0035】

したがって、完全な構成コンテキストで自然言語コンテンツを自動的に解釈し、データがほとんどまたはまったくなくても動作し、説明可能で扱いやすいプラットフォームが明らかに必要である。この方法およびシステムは、テキストの談話以外の関連知識も統合できる必要がある。上記の調査または知識獲得の例では、プラットフォームはすべての調査タスクを自動化するか、知識労働者を大幅に支援できる必要がある。また、トピックモデリングの例では、プラットフォームは、構成性を含むコンテンツを完全に理解した上で、各ドキュメントのトピックの識別を改善できる必要がある。

40

【0036】

本発明の概要

このセクションでは、知識ベースの計算言語学[CL]を使用した深層学習に基づく自然言語理解プラットフォーム[NLUプラットフォーム]について説明する。このプラッ

50

トフォームは、構成言語モデルとも呼ばれる。プラットフォームは完全なコンテキストと追跡可能性を保持し、動作するために必ずしもデータを必要としない。したがって、本発明は深層言語学習に依存しているが、動作するためにデータでトレーニングする必要がある従来の機械学習のように、データからの学習を必要としない。利用可能なデータがあれば、プラットフォームはそれを活用することができ、プラットフォームに必要なのは、現代の機械学習技術に必要なデータのほんの一部だけである。これは、この開示の後半で提供されるいくつかの結果セットから観察される。

【0037】

NLUプラットフォームは、自然言語の意味表現を生成する。この表現により、ゼロショット学習またはフューショット学習が可能になる。NLUプラットフォームの使用法は、人間らしい理解を必要とする自動化された調査のための例示的なNLUタスクについて示される。この開示では、「ドキュメント」という単語が頻繁に使用されているが、その定義は、処理のためにテキストに転写されたビデオまたはオーディオコンテンツも含む。この開示では、「ドキュメント」と「コンテンツ項目」という単語が同じ意味で使用される。

10

【0038】

一般的に、テキストが長くなると、理解と、ひいては表現の作業が、より複雑になる。テキストが長くなるほど、人間による理解が複雑になり、任意の機械表現もその複雑さを継承する。理解は「層」という観点から概念化され、ドキュメントの主要なアイデアの高レベルの理解から始まり、追加の詳細または層の理解へと広がる。ユーザは、ある論説の本質、ある論説の明確なテーマ、ある特定の観点からの論説、記事で表現された個人の視点などを理解したいと考える可能性がある。これは、ユーザが談話を「理解」するために柔軟で階層化された表現が必要であることを示唆している。

20

【0039】

自然言語の普遍的な意味表現は、テキストを、著者が自分のアイデアを表現するために使用したコンセプトと談話関係に戻って分解することを伴う。これには、テキストをリバーシエンジニアリングして、その基本的なアイデア、ならびに完全なドキュメントにつながる文および段落を形成するために結合されたフレーズおよび構成要素に戻す必要がある。そして、人間と同じようにそれを完全に理解するには、予備知識／グローバル知識を統合する必要がある。

30

【0040】

より具体的には、機械表現は次のニーズに答えたりサポートしたりできる必要がある。

【0041】

・ドキュメントの主要なアイデアは何であるか

【0042】

・ドキュメント内の他のアイデアは何であるか。

【0043】

・そのアイデアは他のアイデアおよび推論によってどのようにサポートされているか。

【0044】

・ドキュメント全体を読まなくても済むように、ドキュメントのコンテンツの概要を作成するために使用され得るか。

40

【0045】

・ドキュメント内の特定のアイデアに対する関心に基づいて、ドキュメントのアイデア固有の概要を作成するために使用され得るか。

【0046】

・読者が自分の興味とペースに基づいて層を剥がすことができるように、階層化された理解表現を作成するために使用され得るか。

【0047】

・多層的な表現やグラフなどの視覚的に効果的な表示を通じて認知負荷を軽減するために使用され得るか。

50

【0048】

・人間が自然に行うのと同じように、ドキュメントの外部のグローバル知識によって表現を強化することはできるか。

【0049】

・あらゆるトピックに関する知識コレクションを動的に作成するために使用され得るか。

【0050】

・複数のドキュメントまたは知識コレクションにわたって理解を集約／統合するために使用され得るか。

【0051】

・ドキュメントコレクションまたは知識コレクション全体の主要なトピックを識別するために使用され得るか。

【0052】

・自然言語応答を生成するために使用され得るか。

【0053】

図2は、NLUプラットフォーム200を高レベルで示す。図2は、本発明の好ましい実施形態による、NLUプラットフォームの主要な構成要素とそれらの関係の概要を示す。ユーザ205は、継続的に調査を実施したり知識を発見したりするために、NLUプラットフォームアプリケーション210と対話する。アプリケーション210は、ユーザ205が関心を持つ1つ以上のトピックをNLUプラットフォーム215に転送する。NLUプラットフォーム200は、右側の図2に示すすべてのドキュメントレベル250およびドキュメントコレクションレベルの構成要素をカプセル化する。NLUプラットフォーム200は、インターネット220上の一般的な検索エンジンのいずれかを使用して、トピック（複数可）のコンテンツを探す。NLUプラットフォーム200はまた、パブリックインターネット220、プライベートまたは内部のドキュメントリポジトリ225、およびPubMed、USPTOなどのパブリックドキュメントリポジトリ228からのコンテンツを集約することもできる。次に、NLUプラットフォーム200は結果を処理して、ドキュメントレベル250およびドキュメントコレクションレベル280の構成要素として表示される出力を生成する。保存された知識240は、トピックまたはコンセプトレベルでユーザが以前に作成した知識を含有する。グローバル知識260は、コンセプトおよび関係性についてすべてのユーザが利用できる知識を含有する。談話モデル290は、とりわけ、電子メール、契約書、調査ドキュメントなど、異なる談話種類用の談話構造を含有する。保存された知識とグローバル知識はどちらも、処理中のドキュメントに明示的には存在しない。例えば、「このドキュメントをゼロックスでコピー（Xerox）できるか」は、ドキュメントをコピーすることを暗に意味する。

【0054】

図3は、NLUプラットフォームの処理フローを示す。NLUプラットフォーム200がNLUプラットフォームアプリケーション210からユーザのトピック（複数可）またはクエリーを受信すると、クエリーは、クエリーノーマライザー302によって正規化される。クエリーノーマライザー302は、クエリーを調べてコンセプトを抽出し、ユーザのクエリーに存在する場合は、クエリー内の修辭表現で表現されたユーザの意図も抽出する。例えば、クエリーが「コーヒーは血圧を上昇させるか。」である場合、クエリーノーマライザー302は、「コーヒー」、「血圧」、および「上昇させる」を、「コーヒー」と「血圧」との間の特定の修辭関係として抽出する。クエリーが「コーヒーと癌」である別の例を考える。この場合、クエリーノーマライザー302は、クエリーをユーザの関心のある複合コンセプトとして正規化し、デフォルトで2つの単純なコンセプト（「コーヒー」と「癌」）の間のあらゆる種類の修辭関係を見つける。

【0055】

ドキュメントアグリゲータ305は、ユーザ205のために、アクセスできるソースに対してクエリーを実行する。例示的な実施形態では、ドキュメントアグリゲータ305は

10

20

30

40

50

、インターネット 220、パブリックドキュメントリポジトリ 228 およびプライベートドキュメントリポジトリ 225 にアクセスできる。インターネット 220 では、ドキュメントアグリゲータは、Google などの検索エンジンを使用する。ドキュメントアグリゲータ 305 は、結果をドキュメントプリプロセッサ 307 に渡し、ドキュメントプリプロセッサ 307 は、各ドキュメントに対して一連の前処理工程を実行する。このような前処理アクションの例は、ドキュメントコンテンツの一部ではない任意の HTML タグ、任意の余分な情報のドキュメントコンテンツを削除またはクリーンアップすることを含む。ドキュメントプリプロセッサ 307 は、ドキュメント構造を検査してその談話モデルを決定する談話モデル決定 300 に、クリーンなドキュメントを提供する。談話モデル決定 300 は、談話モデル 290 から使用可能な談話モデル構造をロードする。談話モデルの一般的な例は、電子メール、記事、ブログ、および調査レポートである。管理者は、プログラミングなしで抽象的な談話モデル構造を使用して、新たな談話モデルを追加することができる。新たな談話モデルで抽象的な談話モデル構造の変更が必要な場合は、談話モデル決定 300 の構成要素が強化される。談話モデル構造は、このセクションの後半でより詳細に説明される。

10

【0056】

NLU プラットフォーム 200 が談話モデルを決定すると、ドキュメントパーサー 310 を使用してドキュメントが解析される。ドキュメント解析では、談話モデルを入力として受け取り、フォーマット固有のネイティブドキュメント解析を実行し、データをフォーマットに依存しない表現に変換する。フォーマットの例は、HTML、Docx、PDF、Pptx および eBook を含む。解析は、ヘッダ、フッタ、脚注、段落、箇条書きリスト、ナビゲーションリンクおよびグラフィックを含むネイティブ要素の明示的な解析を含む。解析後、ドキュメントパーサー 310 は、すべての要素の絶対的な追跡可能性を維持するのに十分な属性を持つ階層グラフにコンテンツを整理する。

20

【0057】

基本的な処理が完了すると、ドキュメントは繰り返し表層言語パーサー 320 を通過する。繰り返し表層言語パーサー 430 は、表層 NLP タスクを実行して、ドキュメントの階層グラフ表現から大量のテキストのすべての機械属性および言語属性を評価する。段落、文、節を識別し、文を単語によってトークン化する。セクション（サブセクション）、段落（子段落）、文、単語、品詞（POS）タグ、題目、句、言語属性を決定する（Wikipedia の Grammatical Category のページで拡張された分類法を参照すること）。すべての属性は、パイプラインの後続の工程で果たす役割を持つ。これらの属性は、後続の工程で言語規則の構成要素を形成する。

30

繰り返し表層言語パーサー 320 は、すぐに使用できる表層 NLP パーサーによって生成される一般的なエラーのために繰り返しである。これは、一般的な品詞タグ付けプログラムからの品詞タグの一部の一般的なエラーを防ぐために行われる。これらの規則は、単語自体、単語の前後の単語、節やその属性、文の種類や文の属性と相関している。このような規則はメタデータとして保存され、ユーザまたは管理者がプログラミングなしで強化され得る。

【0058】

40

言語簡略化ツール 330 は、言語規則および構造規則（さまざまなレベルにおける）を使用してさらなる簡略化を実行し、解析された出力の精度を向上させる。言語簡略化ツール 430 には、外部知識も組み込まれており、実際のコンセプトを検出し（軽微なセンスの曖昧さ解消を通じて）、単語を集約して文の長さを全体的に簡略化する。情報と関係を維持しながら文の長さをさらに正規化するための機械的な副文検出（および個別の処理）も、この構成要素によって実行される。簡略化ツール 330 は、一般的なエラーを防ぐために、機械的な置換を実行してパーサーによく知られた単語を提示し、意味的な置換を実行してパーサーに従属節または句全体の 1～2 語を提示する。置換は知識ネット（Knowledge Net）と言語属性（表層解析から）を使用して識別され、並列境界は拡張され（例えば、法と憲法上の権利＝法的権利と憲法上の権利）、従属節は簡略化される（

50

例えば、住宅を所有する一戸建て世帯＝一戸建て住宅所有者）。

【0059】

繰り返し深層パーサー340は、文のより深い解析を実行する。繰り返し深層パーサー340は、POSタグ付けプログラムによる誤判定が原因で解析ツリーに生じた特定のエラーを修正するために、繰り返しである。修正後、プロセスが再開される。深層パーサー440はまた、コンセプトとその共参照も決定する。共参照関係は代名詞的な関係（例えば、ドキュメント内で前または後に使用されている名詞を参照するための代名詞の使用）である場合もあれば、コピーを指すために「xerox」を使用するなどの実用的な関係である場合もある。コンセプト間のより深い共参照関係および他の修辞関係を可能にするために、深層パーサー340はまた、追加のコンテキストの現実世界の知識も統合し、例えば、Civil Rights Movementは、「の間に（during）」、「の後に（after）」などの前置詞が付いている場合は「期間」も意味する。Wikipediaによると、この期間は1954年～1968年である。また、このドキュメント全体にわたる参照（直接的または暗示的）により、Rightsがここでは動詞または形容詞ではなく名詞であるという事実もわかる。この構成要素は、構造内の修飾語および制約語（存在する場合）を含む文の構造も、あれば、決定する（例えば、Civil Rights Movement during Kennedy Administration）。

10

【0060】

繰り返し意味エンコーダ350は、先行する構成要素の出力をすべて使用して、人間らしい意味表現層を作成する。これは、アイデアの検出、アイデア内およびアイデア間の修辞関係を含む。繰り返し意味エンコーダ350はまた、アイデア／修辞表現、中心テーマ、および主要なトピックの重要性などの重要な言語メトリックも決定する。繰り返し意味エンコーダ350によって実行される意味エンコーディングプロセスは、階層、アイデア、議論／修辞表現、文、節、コンセプトを組み合わせる結合された意味表現ハイパーグラフ（MR）を出力する。MRは完全に可逆であり、ドキュメント全体をMRから生成することも、その逆も可能である。意味エンコーダ450はまた、属性（重要性、層、概要など）も導出し、より選択的な分析と専門的な理解のニーズに対応する。

20

【0061】

グラフは、人間が情報を整理して表示する方法を反映する効果的な方法であることが広く認められている。構成性を維持した談話のグラフベースの表現により、効果的に上記のテキスト理解レベルをコンピュータで代理できるようになる。意味表現グラフ〔MR〕は、上記の多層の意味の定義と一致する任意のドキュメントまたはドキュメントコレクションの意味表現を提供する。

30

【0062】

MRグラフは、知識グラフ、コンセプトグラフ、AMRグラフなど、文献内の他のグラフ形態と区別する必要がある。これらすべてのグラフ表現には重複があるが、MRグラフは、構成性〔文、句、節〕、コンセプト、およびコンセプト間の修辞的つながりの観点から、表面的な形態で完全なドキュメントを表す。MRは表面的な修辞句を抽象的な修辞関係の種類にマッピングするが、逆変換は可能である。

40

【0063】

MRは、任意のドキュメントの多層意味表現である。意味とは、談話で説明されているコンセプトとコンセプト間の関係を理解することであると定義される。当然ながら、人間の心の中では、コンセプトおよび相互関係についての既存の知識がこれに補足される。NLUプラットフォーム200における意味の階層化表現が最初に提示され、その後、ドキュメントの理解に関連するドキュメント外部の知識を含むように拡張される。

【0064】

概念的には、ドキュメントは、著者が読者に伝えようとしている主題に関するアイデアのセットとして定義され得る。次に、アイデアは文の形態で表現され、各文は修辞的なつながりを使用して、ある特定のコンセプト間の関係を表す。したがって、単純な文は、主

50

語と目的語が動詞によって結び付けられていると考えることができる。

【0065】

このアイデアは数学的に以下のように表すことができる。

(数1)

ドキュメント $D = \{I\}$ であり、式中、 I は、アイデアまたはテーマのセットであり

、

ここで、

$I = \{S\}$ であり、式中、 S は、文のセットであり、

$S = \{C, R\}$ であり、式中、 C は、コンセプトのセットであり、 R は、修辞表現のセットである。

10

【0066】

NLUプラットフォーム200は、上記の定義を計算的に列挙してMRグラフを生成する。前述のように、生成されたMRグラフは多層構造になっており、徐々に意味を明らかにすることができる。「アイデア」、「テーマ」、「トピック」は、開示の残りの部分では互換的に使用される。

【0067】

多くの文 $\{S_1, \dots, S_n\}$ を含む仮想ドキュメント D を考える。最高レベルでは、NLUプラットフォーム200MRは、ドキュメント内のアイデアのグラフである。図4は、これを概念的に示す。短いドキュメントの場合、ドキュメントは、1つのアイデアのみを反映し得る。ドキュメントが長くなるほど、アイデアの数も多くなる。しかし、ほとんどのドキュメントが、メインテーマといくつかのサブテーマを有する可能性がある。場合によっては、アナリストレポートなど、サブテーマは相互に独立する場合がある。

20

【0068】

概念的には、アイデアまたはテーマを反映する文は一緒にまとめられ、それぞれの新たなテーマは理想的には新たな文のセットとして現れるはずである。すべての文が同じドキュメントからのものであるため、すべての文は関連しているということを覚えておくこと。しかし、ドキュメント内で表現されている主要なサブテーマを検出することは興味深いことである。

【0069】

NLUプラットフォーム200は、単語、節、文、段落、セクションおよびドキュメント全体のレベルでドキュメントの言語構造／属性を詳細に調べることで、アイデアを検出する。ドキュメントのすべての部分を分析して、それらがドキュメントの他の部分と修辭的および言語的にどのように関連しているかを決定する。

30

【0070】

各アイデアは談話の中で複数の文を含み、アイデアを紹介する主文がそのアイデアの見出し文になる。アイデアは、文または主文の主要なコンセプトを通じて表現され得る。

【0071】

図5は、図4の意味表現をアイデア1の拡張表現とともに示す。拡張表現は、アイデア1がドキュメント内の1番目の段落と2番目の段落からの複数の文を含むことを示す。NLUプラットフォーム200は、単語、節、文、段落、セクション、ドキュメント全体のレベルでの包括的な言語属性のセットと、談話モデル要素とを使用して、アイデアまたはテーマの始まりを検出する。談話モデル要素の例は、「新たな段落」マーカーである。

40

【0072】

図5は、アイデア1が段落P1のS1、S2、S3、段落2のS1などを含むことを示す。図5はまた、文間の修辭関係も示す(例えば、R12)。R12は、文S1(P1)とS2(P1)との関係を示す。このような関係の例は、詳細化、時間的などである。簡単に言えば、R12が「詳細化」として識別される場合、S2はS1(P1)の詳細化であると判断される。NLUプラットフォーム200は、このような修辭関係の完全なセットを検出する機能を有する。

【0073】

50

次に、MRは、ドキュメントのコンセプトレベルの表現に拡張される。NLUプラットフォーム200のコンセプトグラフは、コンセプトレベルでのドキュメントの完全な認知表現である。NLUプラットフォーム200は、文を拡張して、コンセプトとそれらの関係の形態で文の構成構造を明らかにする。文同士の関係は、実質的には文の主語同士の関係でもある。例えば、代名詞で始まる文は、関連する文の中の人物を指す。コンセプトは名詞または句であり得、場合によっては名詞または動詞句の正規化された形態であることもある。これはノードとエッジからなる有向グラフであり、ノードはコンセプトであり、エッジは1つのコンセプトを別のコンセプトとに結び付けるテキストの句である。

【0074】

NLUプラットフォーム200のコンセプトグラフ(GG)は、形式的には以下のよう
に定義され得る。

(数2)

$GG = \{V, E\}$

式中、GG=コンセプトグラフ、

$V = GG$ 内の頂点のセット $\{A, B, C, D, E, \dots\}$ 、および

E = 表面レベルの修辞句を含むエッジのセット、

$E_{ij} = i$ 番目と j 番目の頂点間のエッジ

R = 表面レベルの修辞句(E)の抽象的な関係の種類

【0075】

図6は、コンセプトノード $\{C_1, \dots, C_8\}$ とエッジ E を持つ有向グラフの例を示す。 E 内の各エッジは E_{ij} で示され、 i はソースであり、 j はターゲットである。 E_{ij} は、抽象関係の種類(R_{ij})にマッピングされ、 R_{ij} は、抽象関係の種類 R のセットに属する。エッジ E には、重みが関連付けられることがあり、これは、ドキュメントの全体的な意味を解釈する際の意味的重要性を示すことができる。NLUプラットフォーム200は、エッジがマッピングされる抽象関係の種類 R に基づいてデフォルトの重みを使用する。わかりやすくするために、図6は、エッジの重み(RW)を示さない。

【0076】

図7は、ニュース記事710、研究論文720、電子メール730、法的契約書740の4つの典型的な談話モデルを示す。示されているように、談話モデルは階層構造を有する。談話モデルの階層内の任意のノードはオプションであり、[...]のノードは任意であり、他のノードは通常、常に存在する。しかし、構造的には、管理者は、談話モデルに名前を付けるための一意のラベルを使用して、談話モデル290内に任意の数の談話モデルを作成することができる。同様に、管理者は、階層に任意の深さレベルを作成することができる。例えば、法的契約書DM 740は、「条項1テスト」というタイトルのノードと、「[サブ条項]」というタイトルの子ノードとを有する。管理者は、必要に応じてサブ条項の子ノードを作成することができる。NLUプラットフォーム200は、ドキュメントに最も近い談話モデルを見つけようとする。

【0077】

図8は、移民執行に関連するニュース記事のサンプルを示す。説明と参照を容易にするために、各文には $S_1, S_2, S_3, \dots, S_n$ というラベルが付けられている。

【0078】

図9a~bは、図1で報告された10個のクエリーの結果に対するNLUプラットフォームの分析を示す。検索エンジンの結果は、NLUプラットフォーム200を使用して処理された。NLUプラットフォームの出力は、図9a~bに示される。NLUプラットフォーム200を使用すると、関連ドキュメントの検出が大幅に改善される。図9aは、NLUプラットフォーム200対検索エンジンの精度メトリックを示す。関連項目の検出精度は、59%から88%に、29%と大幅に向上した。F1スコアも74%から89%に向上した。NLUプラットフォーム200による関連性に基づく項目のランキングと、Googleの検索結果における項目の順序に基づいて、平均二乗誤差[MSE]統計も示されている。図9bは、これをグラフで示す。これは、[Googleページ-NLUプ

10

20

30

40

50

プラットフォームページ]として計算されたページ差異の頻度分布を示す。結果は驚くべきものであり、NLUプラットフォーム200は、関連性をはるかに正確に検出することができる。

【0079】

図10は、NLUプラットフォーム200によってドキュメントのクエリートピックとの関連性を決定するプロセスを示す。コンセプト1010とドキュメント1015は、NLUプラットフォーム200に入力され、3段階のプロセスを経てそのドキュメントのクエリーとの関連性が決定される。コンセプト検出（ドキュメントのMRグラフでのコンセプトの出現）の後、複雑なコンセプト（複数のトピック）については、文の重要性（主要なアイデア、主要なアイデアの文のセット）とコンセプトの重要性（主題）の他に、NLUプラットフォーム200はコンテキストの強さも決定する。コンテキストの強さは、ドキュメントのコンテキスト内でコンセプトが互いにどれだけ不可欠であるかを決定する。例えば、複合コンセプトの各単語は、互いのコンテキストに直接関係することなくドキュメント内のさまざまな場所に出現しているか、それとも、互いに直接的なコンテキスト関係で出現しているか、である。

10

【0080】

図11は、図8のドキュメントの1番目の文のコンセプトレベルの表現を示す。図面のノードは文中のコンセプトを表し、エッジは表現された修辭関係を反映している。

【0081】

図12は、図8に示されたドキュメントの概要を示す。NLUプラットフォーム200は、27の文のうちの7つがドキュメントの本質を反映していることを見つけた。ユーザが図12の概要に含まれるドキュメントから7つの別個のアイデア／テーマのいずれかを拡張したい場合、NLUプラットフォーム200は、そのアイデアの一部である文を使用して、そのアイデアを拡張する。拡張は、ユーザが関心のある特定の修辭関係のセット（例えば、「因果関係」）によってフィルタリングされ得る。概要は、ユーザが関心のある粒度レベルになる。同様に、概要は、ユーザが関心のある修辭関係の種類のみを反映できる。例えば、ユーザはドキュメントの主要なアイデアのみに関心があり、「詳細」には関心がない可能性がある。NLUプラットフォーム200はまた、ユーザが関心のある任意のドキュメントのトピック用の概要も生成する。例えば、ユーザは、ドキュメントの主要なトピックに関心がある場合もあれば、ドキュメント内の他のトピックのいずれかに関心がある場合もある。NLUプラットフォーム200によって生成された概要は、ユーザの設定によって異なる。

20

30

【0082】

図13は、NLUプラットフォーム200によって生成された電気自動車知識コレクションを示す。目次（COP26、政府、企業、中小企業、個人、司法）は、関連性があると判断したすべてのドキュメントのコンテンツの完全な理解に基づいて、NLUプラットフォーム200によって生成される。各目次トピックは、NLUプラットフォーム200がそのトピックに分類したコンテンツ項目の数を有する。コンテンツ項目は、複数の目次トピックに分類され得る。このような場合、異なるトピック用の概要は異なることに注意を払うべきである。任意の目次トピックを選択またはクリックすると、NLUプラットフォームアプリケーション310にそのトピックに関連する記事が表示される。ルートトピック「気候変動対策」を選択すると、NLUプラットフォームアプリケーション210に「気候変動対策」に関連するすべての記事が表示される。

40

【0083】

図14は、継続的な知識の発見および更新の例を示す。これは、特定のノードに対して自動検出がオンになっていることを示す。次に、NLUプラットフォーム200は、そのノードとそのすべての子ノードに関連する新たなコンテンツ項目を自動的に検索して見つける。新たに見つかったコンテンツ項目は、NLUプラットフォーム200によって処理され、通常の処理工程を使用して現在の「知識コレクション」に統合される。本発明の実施形態では、新たに見つかった知識をスーパーユーザまたは著者が確認し、新たに見つ

50

かった項目のうちどれを統合すべきかを決定できるようにステージングすることができる。

【0084】

図15は、7つのドキュメントの例示的なコレクションについて、NLUプラットフォーム200によって生成されたトピックまたはタグを示す。トピックは、コンテンツ項目全体に対して作成することも、コンテンツ項目のさまざまなセクションに対して個別に作成することもできる。図16は、NLUプラットフォーム200のゼロショット学習機能の図解である。NLUプラットフォーム200をこれらの7つのドキュメントに適用し、手動で生成されたタグに正しく一致するドキュメント内のトピックを検出できるかどうかを決定した。使用例は、顧客／見込み客がキーワードで検索したときに、そのようなドキュメントの発見可能性を最大化することであった。図16はまた、NLUプラットフォームがトレーニングデータなしで80%の一致を見つけ、現代の言語モデルよりも桁違いで優れていることも示す。タスク固有のトレーニングが行われていない主要な言語モデルでは、同じデータ上のタグの20%しか見つけられなかった。

10

【0085】

図16は、NLUプラットフォームによって生成された、統合された意味表現グラフをより詳細に示す。これは、NLUプラットフォームによってインテリジェントに集約され、図17に示す統合された意味表現グラフを生成する3つのコンテンツ項目からの意味表現を示す。図面では簡潔にするために、3つの意味表現グラフをそれぞれ表示しない。図17はまた、グローバル知識ストアからのグローバル知識の使用も示す。これは、「バイデンの移民政策」が「バイデンの公的扶助規則に関する大統領令」とどのように関連しているかを見れば明らかである。

20

【0086】

図17は、NLUプラットフォーム200がその処理に使用できるグローバル知識入力の構造を示す。入力のフォーマットは柔軟であり、任意のコンセプトのペア間の関係を反映する。関係の列は、拡張可能な関係の種類セットからの修辞関係の種類を反映する。関係の種類を変更するには任意のプログラミングは必要なく、孤立制御などのデータ整合性の原則によって管理される。方向を示す手がかり句（前または後）を、その関係に関連付けることができる。センスの列は、指定された関係においてコンセプトが参照されているセンスを反映する。センスは、拡張可能な属性である。NLUプラットフォーム200の任意のコンセプト（または単語）用のセンスリストは、プログラミングなしで拡張され得る。グローバル知識は、すべての知識コレクション、特定の知識コレクションまたはドメイン、あるいはコンテンツ項目のみに適用できるように指定され得る。必要に応じて、NLUプラットフォーム200は、自動知識獲得システムを呼び出したり、コンセプトに関するグローバル知識を含有するグローバル知識ストアにアクセスしたりすることもできる。

30

【0087】

図17と同様に、図18では、NLUプラットフォームアプリケーション310のユーザがセンスの曖昧さ解消規則を入力できる。これらの規則は、単語の言語特性、その単語を含有する文の中でその単語の前後に出現する単語のセット、それらの単語の品詞タグ、手がかり句、固有名詞や明確な名詞などの内容語の組み合わせになる。NLUプラットフォーム200はまた、センスの曖昧さを解消するために、自動分類子が存在する場合はそれと呼び出すこともできる。このような自動分類子は、特定の状況で単語がどのような意味で使用されているかを自動的に識別する。グローバル知識の場合と同様に、スコープは規則がどの程度広く適用されるか（すべての知識コレクション、特定の知識コレクションまたはドメイン、またはドキュメントのみ）を示す。

40

【0088】

別の実施形態では、本発明のNLUプラットフォームによってアジャイル継続的学習プロセスで実行される自動化された自然言語理解タスクの例示的な方法および使用法を以下に説明する。

50

【0089】

アジャイル継続的学習プラットフォーム（ACLプラットフォーム）：

アジャイル継続的学習プラットフォーム用の効果的なフレームワークは、ユーザに以下のコア要素の要件を提供するプロセスを含む。

【0090】

1．評価やマイクロレデンシャルの有無にかかわらず、継続的に更新される知識ポータルから学習者が自己学習できる機能

【0091】

2．ある機関のコンテンツライブラリ内外から新たな学習コンテンツを迅速に組み立て、整理し、統合し、常に最新の状態に保つ機能（迅速な学習コンテンツ組み立て）。

【0092】

3．ループリックに対してスコアリングされた自由回答の自動化された説明可能な評価（スケーラブルなオープンテキストの回答評価）。

【0093】

4．スキルベース、キャリア、需要主導型の学習ジャーニーのパーソナライゼーション。

【0094】

本発明の自動化されたNLUプラットフォームのACLプラットフォームアプリケーションは、図13～14にも示されているように、その継続的な知識ポータル内の任意のトピックに関する継続的な知識コレクションを自動的に作成する。

【0095】

本発明のACLプラットフォームアプリケーションは、トピックのコレクションに関する自己学習を容易にする1つ以上のポータルを含む。例えば、サードパーティ機関が、ポータルを評価および資格情報と組み合わせることで、生涯学習者にポータルを提供することができる。教員またはコース設計者が、学習者がアクセスできるコンテンツを管理できる。教員は、自分の知識や教授法を伝えるために、ACLコンテンツに適切と思われる注釈を付けることができる。

【0096】

ACLプラットフォームポータルはまた、機関の学習配信インフラストラクチャおよび学習管理システムと統合することもできる。

【0097】

迅速な学習コンテンツ収集（コース作成）

ACLプラットフォームI学習コンテンツオーサリング／アセンブリ構成要素（図19）により、任意の機関が、新たなトピックまたは特定のスキルに関連するトピックに関する学習コンテンツを迅速に組み立てることができる。ACLプラットフォームは、その自然言語理解エンジンを使用して、コンテンツをコンセプトレベルまで分解する。あらゆるフォーマットの既存の学習コンテンツを取り込むことができる。また、他のプライベート学習コンテンツコレクションまたはパブリックインターネットなどの外部リポジトリからの新たなコンテンツでコンテンツを拡大することもできる。

【0098】

ある大手のグローバル教育機関では、知識（G y a n）を用いて既存の膨大なコンテンツを粒度の細かいレベルの学習単位に迅速に分解し、外部コンテンツで補完することで、オンデマンドでの新たな学習モジュールの作成を可能にしている。機関は、新たな学習ニーズを満たすよう、粒度の細かい学習単位を組み合わせることができる。コンテンツは必要に応じて個々の学習者に合わせてパーソナライズされ得る。パーソナライゼーションは、個々の学習者の学習の好みに基づいて行うことができる。

【0099】

このプラットフォームはまた、学習の概要を構築し、希望する範囲に基づいて評価を作成し、最終的なコースまたは学習コンテンツを一般的な学習管理ソリューション（LMS）と統合するインテリジェンスも有する。

10

20

30

40

50

【0100】

スケーラブルな自由テキスト回答の評価

人間によるエッセイおよび自由テキスト回答〔ORA〕のグレーディングは、時間がかかり、矛盾およびバイアスが生じやすくなる。エッセイの自動グレーディング〔AES〕に関する既存の調査は、コンテキスト認識の欠如、大量のトレーニングデータの必要性、データ駆動型のバイアス、およびそれらのブラックボックスの性質のために批判されてきた。さらに、MOOCのようなオンラインコースでは、ピアグレーディングが一般的に行われている。ピアグレーディングは、グレーディング方法として重大な欠点があり、その最大の欠点は、グレードインフレを自然に引き起こすということである。本発明のACLプラットフォームは、これらの課題に効果的に対処する。ACLプラットフォームは、エッセイまたはORAをそのコンテキストで評価し、ラベル付けされたエッセイのスパースデータセットからインテリジェンスを収集し、その推論またはロジックが透明で扱いやすいため、理解しやすい方法で改善することができる。さらに、ACLプラットフォームは、コース固有の要件を反映するように迅速に構成され得る。

10

【0101】

著名な世界的に評価の高い高等教育機関では、現在、オンラインコースの自由テキスト回答の評価にピアグレーディングを使用している。すでにグレーディングされたコースからの課題が、本発明のACLプラットフォームの評価のために選択された。選択された課題については、ピアグレーダーが互いに満点をつけていた。

【0102】

20

ACLプラットフォームは、教授が指定したルーブリックを使用して構成された。ACLプラットフォームが構成された後、10人の学生の提出した課題はすべてACLプラットフォームを通じて評価された。

【0103】

ACLプラットフォームによって提供されるアウトプットグレーディングを取得し、それを人間の基準と比較するために、3人のティーチングアシスタント(TA)に同じエッセイを手動でグレーディングするよう依頼した。図20は、パイロットの結果を示す。見られるように、10人の生徒全員のピアグレードは満点であった。TAは同じエッセイに対して大きなばらつきを有した。ACLプラットフォームの各評価に対して、ACLプラットフォームが最終スコア／グレードにどのように到達したかを説明する詳細な推論レポートが教授／機関に提供された。すべての場合で、教授とTAはACLプラットフォームの評価に同意した。

30

【0104】

パーソナライズされた学習パスウェイ

従来の4年制大学の学位に関係なく、仕事に必要なスキルを持つ人を雇用したいという雇用主の意欲が高まっている。大手の雇用主は教育機関と協力して、必要な人材と労働力を育成するための学習プログラムを作成している。

【0105】

ACLプラットフォームを使用すると、個人に合わせたパーソナライズされた学習パスウェイを自動的に作成できる。これは、世界的な高等教育機関と、失業している成人向けのキャリアガイダンスに重点を置くスタートアップ企業によって使用されている。

40

【0106】

パーソナライズされた学習パスウェイを決定するプロセスは、個々のユーザが履歴書またはプロフィールをアップロードして入力することから始まる。ACLプラットフォームは、履歴書から個人のスキルを抽出する。ACLプラットフォームは、履歴書のさまざまなセクション(明示的に記載されている場合は教育、経験、目的、およびスキル)を処理する。ACLプラットフォームでは、スキルに加えて、各スキルの熟練度レベルも決定する。ACLプラットフォームは、ONET、EMSI Burning Glass、または採用企業に固有の独自のスキル分類法など、あらゆるスキル分類法と統合され得る。

【0107】

50

A C Lプラットフォームは、履歴書に記載されている個人のスキルと希望する職業との間のスキルギャップを識別する。履歴書から抽出されたスキルは、O N E Tなどの選択されたスキル分類法を使用してA C Lプラットフォームによって正規化される。次に、正規化されたスキルを、希望する職業と比較し、その特定の職業における個人のスキルギャップを識別する。

【0108】

ドキュメントの発見可能性タグ

ある機関が提供するコンテンツの発見可能性を向上させるには、通常、タグを手動で作成する、および／またはコンテンツ作成構造（例えば、トピックの階層的なセット）を手動で作成する必要がある。新たなコンテンツが到着したときに、このタグ付けプロセスを先取りし続けることも、同様に困難である。高品質のタグを自動的に生成すると、このプロセスが大幅に効率化されるだけでなく、より大規模でのコンテンツのタグ付けも可能になる。

10

【0109】

本発明のA C Lプラットフォームは、顧客／見込み客による検索での発見可能性を最大化するために、ドキュメントを最もよく説明するタグを生成することをプラットフォームに要求するパイロットプロジェクトで評価された。

【0110】

図15に示すように、7つの記事からなるサンプルサイズの場合、A C Lプラットフォームはトレーニングデータなしで80%のタグとの一致率であり、これは、最新のA Iベースの機械学習システム（L M）よりも桁違いに優れていた。タスク固有のトレーニングを行わない現在使用されている主要なL Mでは、本発明のA C Lプラットフォームアプリケーションと比較して、同じデータでタグの20%しか正確に検出されなかった。

20

【0111】

要するに、本発明は、自然言語理解（理解力）のためのシステムおよび方法を対象とし、このシステムは、1つ以上のテキストドキュメントから自然言語コンテンツの人間らしい自然言語理解のタスクを自動的に実行するように構成されたプロセッサを含み、このシステムは、構成言語モデルを使用してドキュメントの意味の機械表現を作成する自然言語理解エンジンを有する。エンジンは、ドキュメント全体をその構成構造に分解して、複数の意味の層を反映することにより、ドキュメントの機械意味表現を作成し、また、ドキュメントのために識別された談話モデルに従ってドキュメントを解析し、ドキュメントの主要なアイデアを識別し、ドキュメント内の新たなアイデアの始まりを識別し、アイデアによってドキュメントをサブドキュメントに分解し、サブドキュメントを構成要素に分解してドキュメント全体の意味表現を作成することにより、可逆的な方法で機械意味表現を生成し、コンピューティングシステムは、1つ以上の仮想サーバーまたは専用サーバー、あるいは同様のコンピューティングデバイスを含み得、そして実行可能な命令でプログラムされ得る。

30

【0112】

本発明の方法は、1つ以上のテキストドキュメントから自然言語コンテンツの人間らしい自然言語理解機能を自動的に実行し、この方法では、構成言語モデルを使用してドキュメントの意味の機械表現を生成し、次に、ドキュメント全体をその構成構造に分解して、複数の意味の層を反映することにより、ドキュメントの機械意味表現を生成し、そして、ドキュメントに対して識別された談話モデルに従ってドキュメントを解析し、ドキュメントの主要なアイデアを識別し、ドキュメント内の新たなアイデアの始まりを識別し、アイデアによってドキュメントをサブドキュメントに分解し、サブドキュメントをその構成要素に分解することによってドキュメント全体の意味表現を作成することにより、ドキュメントをその構成構造に分解することによって、可逆的な方法で機械意味表現を生成する。

40

【0113】

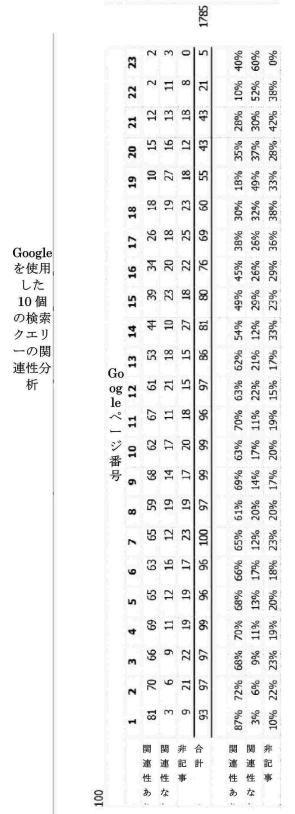
構成言語モデルは、その意味表現を作成するために、統計的機械学習、または単語埋め込みのような統計的に導出された分布意味論を利用しない。自然言語理解エンジンは、計

50

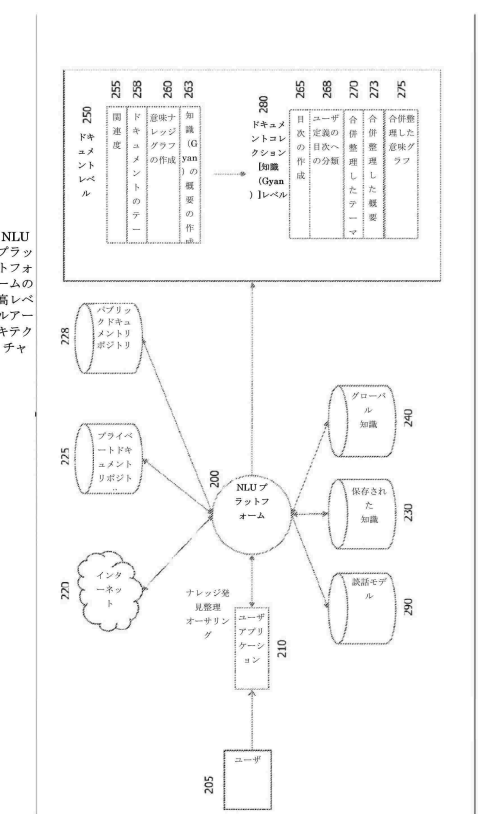
算言語学、言語の構成性および修辞構造を利用し、いずれのトレーニングまたはトレーニングデータもなしで1つ以上のドキュメントを処理できる。自然言語理解エンジンは、完全に自然言語で動作し、ドキュメントの任意の部分を実数値のベクトルエンコーディングに変換しない。

【図面】

【図 1】



【図 2】



10

20

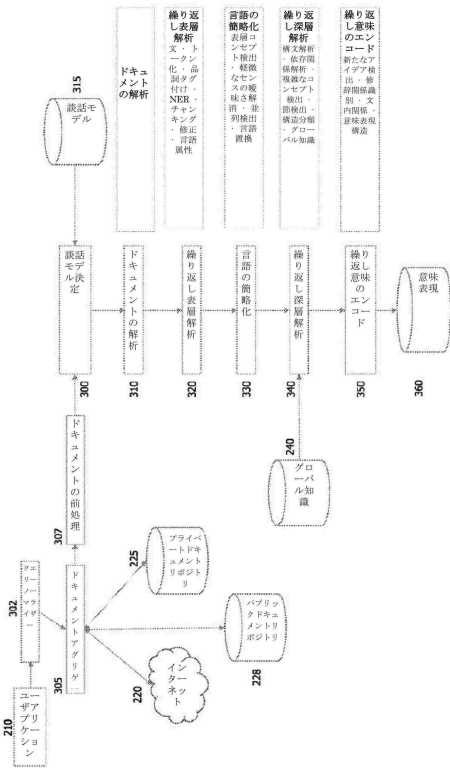
30

40

50

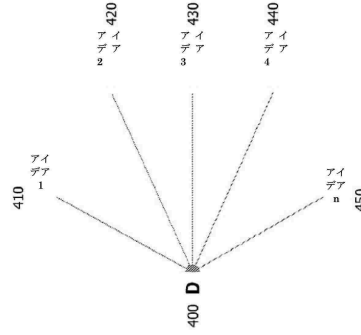
【図 3】

NLU
プラットフォーム
パイプライン
200



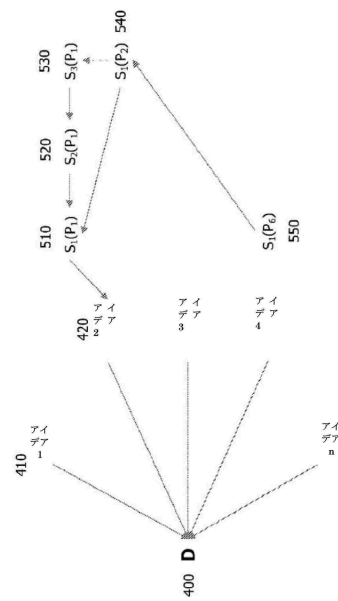
【図 4】

意味
表現
一層 1



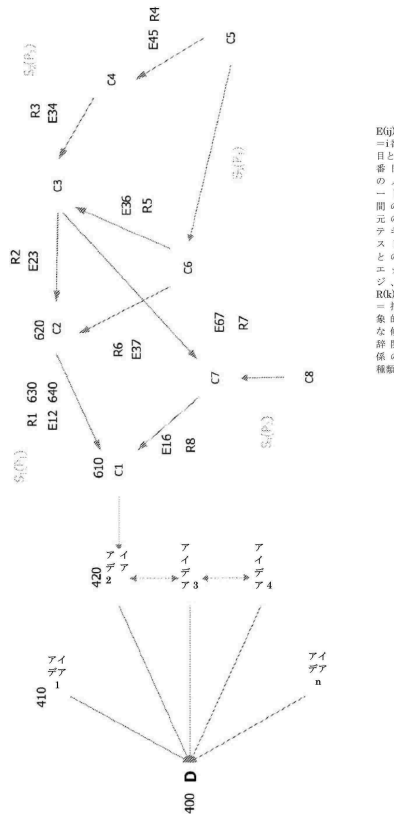
【図 5】

拡張
された意味表現
一層 2

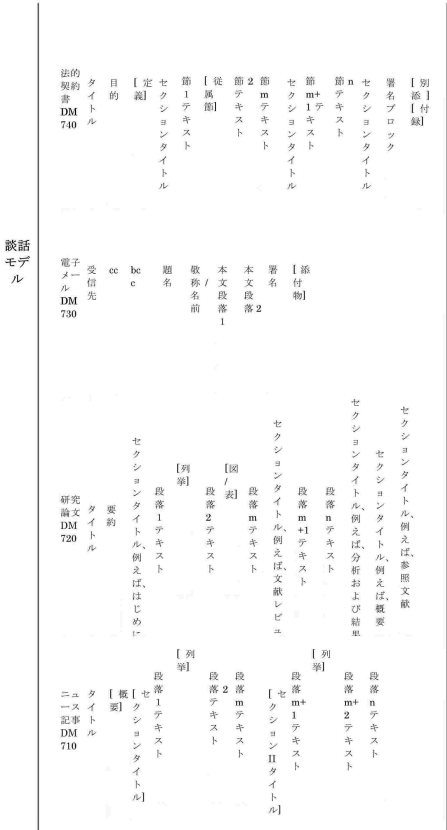


【図 6】

コン
セプト
レベル
の意味表
現グラフ



【図 7】



【図 8 A】

記事例 [S1]ソルトレイクシティ教育委員会が火曜日に再確認した決議によると、ソルトレイクシティ学区は、学生の憲法上および法的権利を保護し、学校が移民執行措置によって混乱しないようにするために「合法的に行使できるあらゆる手段」を講じる。

[S2] 委員会は、学区が入学手続きの一環として、生徒の移民ステータスやその親または保護者のステータスについて問い合わせることはできないし、問い合わせないことを規定する 2017 年学校安全決議を支持することを決議した。

[S3] この決議案は主に、3 月に亡くなった元理事長ヘザー・ベネット氏によって起草された。委員会は、新たな決議を検討するのではなく、その主著者および委員会リーダーとしてのベネット氏の貢献を称えるために、以前のバージョンを再確認することを決定した。

[S4] 全国的に移民法執行が強化され、学生や家族に影響が及ぶ中、この委員会の行動は、全国の地方教育委員会が同様の政策を支持している行動に加わるものである。

[S5] 委員会はまた、児童保護決議を採択し、「米国とメキシコの国境における移民や亡命希望者、特に未成年者の扱いについて構成員が深い懸念を抱いている」と指摘した。

[S6] 委員会は、米国が国境を安全に守る権利を認めているが、「子供たちの生活向上に尽力する者として、委員会は、安全を求める人々を阻止することを目的とした移民政策の手段として、子どもたちに取り返しのつかない害と生涯にわたるトラウマを引き起こす可能性のある家族の分離を利用することを遺憾に思う」と一部で述べている。

[S7] 理事のネイト・サラザール氏は、歴史的に見て、ソルトレイク郡は歓迎的なコミュニティであったと言った。ソルトレイクの学校は移民や難民の生徒が多く、少数派が多数派を占めているが、その多様性を今後も称賛し続けることが重要だと彼は言った。「家族や子供、学生、そしてもちろんソルトレイクの支援団体から聞く現在の恐怖は現実のものであり、子供たちが学校に通い、そのような恐怖の中で自宅で生活していることで、学校に通い、学校に出席し、教育を受ける機会を得るという重要な仕事が確実に奪われている」とサラザール氏は言った。

[S8]彼はさらに、「私は教育委員会のメンバーとして、私たちの最優先事項は子供たちだと心から信じている。この問題に関する不安を和らげるために、学区や委員会として何かできることがあれば、それを実行する必要があると思う」と言った。

【図 8 B】

[S9]2017 年学校安全決議に対する投票は、すべての生徒が「民族、人種、宗教、国籍、移民のステータス」に関係なく、また嫌がらせや脅迫を受けずに」学校に通い、学校で利用できる機会と利益を享受できる平等な機会を提供するという学区の取り組みを再確認するものである。

[S10] 学校安全決議では、当局が連邦または州の裁判官または治安判事から発行された搜索令状を持っていない限り、まれな場合を除き、移民執行の目的で学校への立ち入りは許可されないと規定されている。

[S11]「連邦法で定義されている非常に限定的でまれな『緊急事態』においては、地区は有効な搜索令状なしでアクセスを許可することが法的に義務付けられる。しかし、緊急事態とは一般的に、逃走中の犯罪容疑者を警察が追跡している場合や、犯罪の証拠が破壊されようとしている場合などであるため、生徒が学校活動に従事している間に緊急事態が発生する可能性は極めて低い」と方針には記されている。

[S12]学区の方針では、職員が学校の敷地内に入ろうとしたり、学区の監督下にある生徒と連絡を取ろうとしたり、学区が移民執行の目的で生徒に関する情報の要請を受けたりした場合は、直ちに政策および法務サービスのエグゼクティブディレクターに連絡するよう職員に指示されている。

[S13] 決議では、「政策・法務サービス担当の事務局長が審査するまで、いかなる要請も認められない」と規定されている。

[S14]学校、交通ルート、学区の敷地内、または学校行事中の執行活動は、「学習環境を著しく混乱させ、不当な搜索や押収から自由であり、無料の公教育を受けるという生徒の憲法上の権利を著しく侵害することになる」と決議では述べている。

[S15] 決議では、家族が「子供が学校にいる間に強制送還されるのではないか、あるいは移民当局が学校に介入するのではないかという恐れから、子供を学校に通わせることに懸念を表明している」と指摘している。

【図 8 C】

[S16]活動家組織「移民団結」のエイミー・ドミンゲス氏は委員会の行動に感謝の意を表した。しかし、彼女は委員会メンバーに対し、拘留、国外追放、家族の離散の影響を受けた学生を支援するための緊急対応チームも設置するよう求めた。

[S17]ドミンゲス氏は、連邦移民執行に関する包括的かつ独立した無党派の情報源である TRAC 移民局を引用し、7 月以降ユタ州で 1,382 件の退去命令が出されたと言った。

[S18]「私たちは、これが移民の学生に引き起こしているストレスを認識し、家族を引き離す慣行が私たちの偉大な州に迫っていることを思い出さなければならない」とドミンゲス氏は言った。安全学校決議では、「今まで、学区は連邦移民当局から何の接触も受けていない」と記されている。

[S19]米国民関税執行局は、公立学校を「敏感な場所」、つまり通常の業務の過程で移民執行措置が行われるべきではない場所と位置づけていると決議は述べている。しかし、ドミンゲス氏は、ICE は営業時間外や週末にも学区の敷地内に車両を駐留させていると言った。

[S20]「これは移民コミュニティに恐怖心を煽り、強制退去を恐れる学生たちが授業に出席できないようにするためのものだ」と私たちは考えている」とドミンゲス氏は言った。学区は移民コミュニティに対し、これは学区の許可なく行われ、連邦移民当局と協力していないことを明確にする必要がある、と彼女は言った。

[S21] サラザール氏は、委員会の行動は「こうした慣行がさまざまな形で推進され、実行されている連邦上層部の人々の心や考えを大きく変えることはないだろう。もし私たちが、ここにいる子供たちや家族に、学校や教育者、学区が自分たちに寄り添って立っていることを知ってもらい影響を与えることができれば、それは非常に価値のあることだと思う。これは学生や地域社会の日々の生活の質に大きな影響を与える」と言った。

[S22] 国内の他の地域では、学区が学校敷地内での移民の権利を保護する計画を展開している。

10

20

30

40

50

【図 8 D】

[S23]デトロイト・フリー・プレスによると、デトロイト公立学校コミュニティ地区の新しい方針では、連邦執行機関などが捜索令状なしで地区の敷地に入ることは許可されなくなり、学校は生徒の移民ステータスに関する情報を収集しなくなる。

[S24]マイアミ、ニューヨーク市、デモイン、シカゴの各地区も、令状がない限り連邦執行官に協力しないという方針を採用している。

【図 9 A】

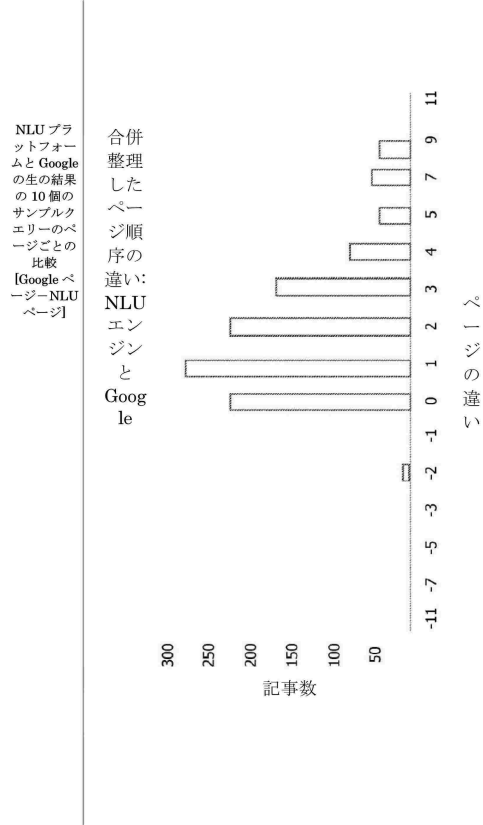
NLUプラットフォームとGoogleの10個のサンプルエリートの生の結果の比較

差分 (N-H)	29%	15%	-28793%	-29%	-41%
Google 生データ (H)	59%	74%	30168%	41%	41%
NLU エン ジン (N)	88%	89%	1375%	12%	1%
メ トリ ック	P	F1	MS E	MA E	MB E

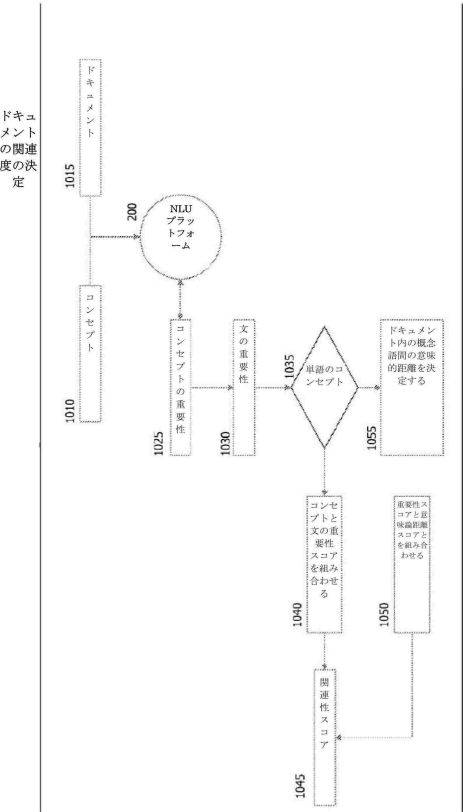
10

20

【図 9 B】



【図 1 0】

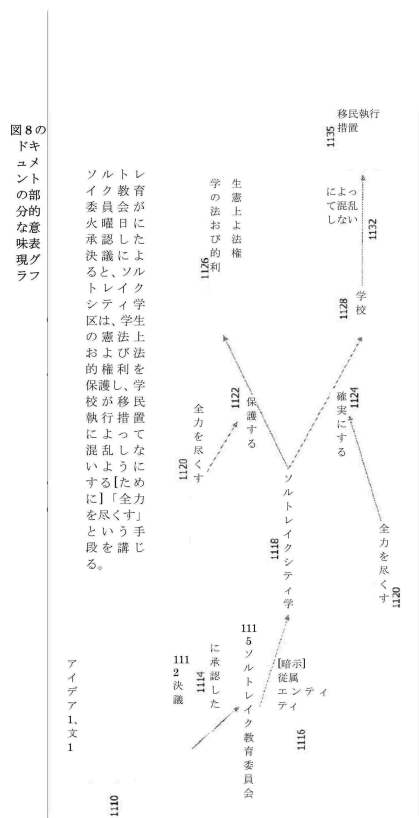


30

40

50

【図 1 1】



【圖 12】

令状がなければ移民執行のために学校に立ち入ることはできない、とソルトレイクシティ教育委員会が決議

ソルトレイクシティ教育委員会は、学区が「生徒の機密情報を保護するために、法的権限の範囲内であらゆる手段を講じる」ことを誓約した。

ドキュメント概要の例

全国的に移民法執行が強化され、学生や家族に影響が及ぶ中、この委員会の行動は、全国の地方教育委員会が同様の政策を支持している行動に加わるものである。

理事のネイト・サラザール氏は、歴史的に見て、ソルトレイク郡は歓迎的なコミュニティであったと言った。

学校安全決議では、当局が連邦または州の裁判官または治安判事から発行された搜索令状を持っていない限り、まれな場合を除き、移民執行の目的で学校への立ち入りは許可されないと規定されている。

学区の方針では、職員が学校の敷地内に入ろうとしたり、学区の監督下にある生徒と連絡を取ろうとしたり、学区が移民執行の目的で生徒に関する情報の要請を受けたりした場合は、直ちに政策および法務サービスのエグゼクティブディレクターに連絡するよう職員に指示されている。

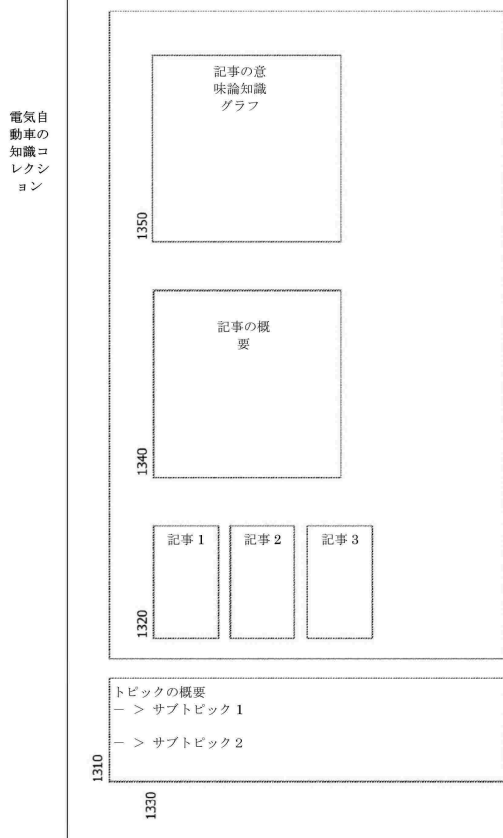
活動家組織「移民団結」のエイミー・ドミンゲス氏は委員会の行動に感謝の意を表した。

国内の他の地域では、学区が学校敷地内での移民の権利を保護する計画を展開している。

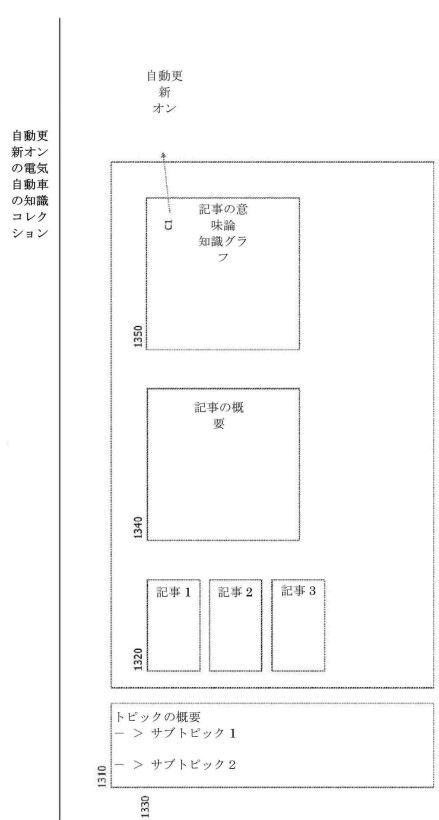
10

20

【図 13】



【圖 14】



30

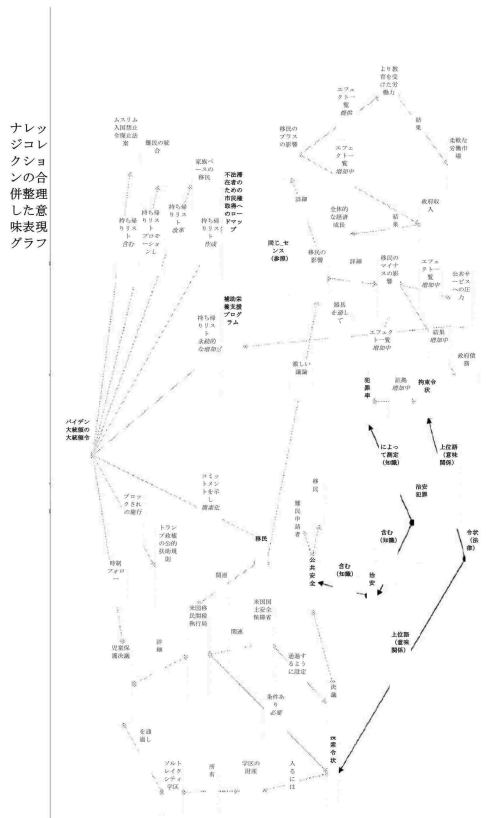
40

50

【図 15】

ドキュメント番号	手動で識別されたタグ	知識 (Gyan) の結果	
	重要性の高いタグ	正しく識別された	一致率%
ドキュメント 1	12	10	83%
ドキュメント 2	3	2	67%
ドキュメント 3	5	4	80%
ドキュメント 4	4	4	100%
ドキュメント 5	4	3	75%
ドキュメント 6	11	9	82%
ドキュメント 7	4	4	100%
総計	43	36	84%

【 図 1 6 】



【圖 17】

グローバル知識入力	#コンセプト1	セシ1	関係	手がかり句	方向	コンセプト2	セシ2	スコアの詳細	スコープの詳細
	米国スタイル	1	と同じ		任意	米国スタイル	1	グローバル	グローバル
	米国のスタイル	1	と同じ		任意	米国のファッション	1	グローバル	グローバル
	野球	1	の、種別である		この	スポーツ	1	グローバル	グローバル
	フットボール	1	の、種別である		この	スポーツ	1	グローバル	グローバル
	米国の宗教	1	の、属性である		この	宗教	1	グローバル	グローバル
	2397は無宗教	1	の、属性である		この	米国の宗教	1	グローバル	グローバル
	ジーンズ	1	の、種別である		この	服装スタイル	1	グローバル	グローバル
	米国のブランド	1	の、属性である		この	米国のブランド	1	グローバル	グローバル
	カルフレーション	1	の、種別である		この	米国のブランド	1	グローバル	グローバル
	ファッション観	1	の、属性である		この	米国のファッション	1	グローバル	グローバル
	米国の料理	1	と同じ		任意	米国の食品	1	グローバル	グローバル
	テラス・メヌス	1	の、種別である		この	米国の食品	1	グローバル	グローバル
	米国の音楽	1	の、種別である		この	州の芸術文化	1	グローバル	グローバル
	感謝祭	1	の、種別である		この	米国の祝日	1	グローバル	グローバル
	成長率	1	の、属性である		この	人口	1	グローバル	グローバル
	成長	1	の、属性である		この	人口	1	グローバル	グローバル
	低下	1	の、属性である		この	成長率	1	グローバル	グローバル
	長寿化	1	の、属性である		この	人口	1	グローバル	グローバル
	最も急速に成長している人口	1	の、属性である		この	最盛期	1	グローバル	グローバル
	AHA メンバース	1	の、属性である		この	AHA メンバー	1	ドキュメント	ドキュメント
	ボットライト	1	の、属性である		この	AHA	1	ドキュメント	ドキュメント
	人口規模の減少	1	の、属性である		この	人口	1	グローバル	グローバル
	移民	1	の、属性である		この	人口	1	グローバル	グローバル
	移住者の流出	1	の、属性である		この	人口	1	グローバル	グローバル
	最も急速に成長している人口	1	の、属性である		この	人口	1	グローバル	グローバル
	社会的な不満	1	の、属性である		この	エストニア右派民主主義者の台頭	1	グローバル	グローバル
	公開討論	1	の、属性である		この	エストニア	1	グローバル	グローバル

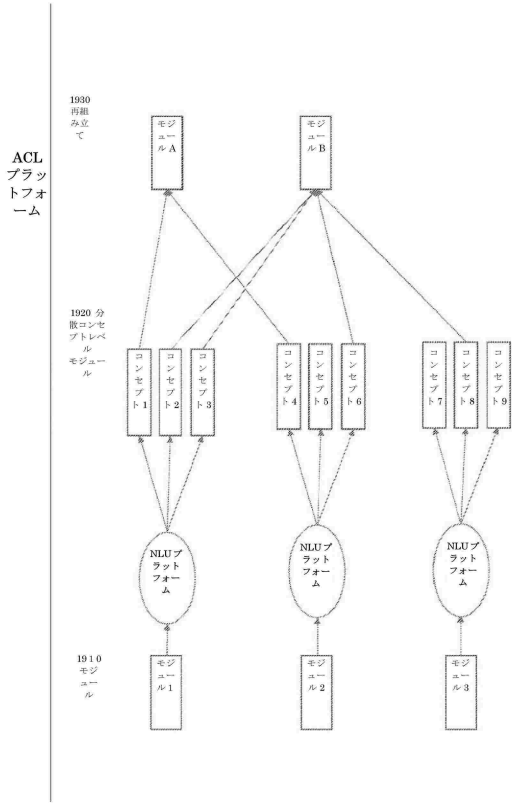
【圖 18】

[illegible]

グローバル知識
手動入力

センズ
の暖房
さ解消
の手動
規則入
力

【図 19】



【図 20】

ACLプラットフォーム

生後	知識 (Gyan) のスコア	ピアダ レーダ ーのスコア	ライ ン ア ン シ ス タ ン ト の ス コ ア [範 囲]
1	11	15	8-11
2	15	15	13-15
3	15	15	13-14
4	12	15	11-15
5	13	15	10-14
6	14	15	13-15
7	15	15	10-14
8	12	15	12-15
9	13	15	11-13
10	14	15	13-15

10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 23/66123

A. CLASSIFICATION OF SUBJECT MATTER

IPC - INV. G06F 40/30, G10L 15/18 (2023.01)
ADD.

CPC - INV. G06F 40/30, G10L 15/18

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)

See Search History document

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

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2009/0327878 A1 (Grandison et al.) 31 December 2009 (31.12.2009), entire document	1-48, 50-61
A	US 9,811,579 B1 (Olson) 07 November 2017 (07.11.2017), entire document	1-48, 50-61
A	US 2008/0010311 A1 (Kon et al.) 10 January 2008 (10.01.2008), entire document	1-48, 50-61
A	US 2013/0138696 A1 (The Institute for System Programming of the Russian Academy of Sciences), 30 May 2013 (30.05.2013), entire document	1-48, 50-61
A	US 2003/0217335 A1 (Chung et al.) 20 November 2003 (20.11.2003), entire document	1-48, 50-61

☐ 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

"D" document cited by the applicant in the international application

"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

16 June 2023 (16.06.2023)

Date of mailing of the international search report

JUL 18 2023

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Kari Rodriguez

Telephone No. PCT Helpdesk: 571-272-4300

Form PCT/ISA/210 (second sheet) (July 2022)

フロントページの続き

,NA,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,D
E,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,ME,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,S
M,TR),OA(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,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CV,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,L
A,LC,LK,LR,LS,LU,LY,MA,MD,MG,MK,MN,MU,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,P
H,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,ZA,ZM,ZW
チズヒル タワーズ イー 2 9 0 2