(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)公開番号
特開2023-165410
(P2023-165410A)

(43)公開日　令和5年11月15日(2023.11.15)

(51)国際特許分類		FI		テーマコード（参考）
Ｇ０６Ｆ	*40/279 (2020.01)*	Ｇ０６Ｆ	40/279	５Ｂ１７５
Ｇ０６Ｆ	*16/383 (2019.01)*	Ｇ０６Ｆ	16/383	

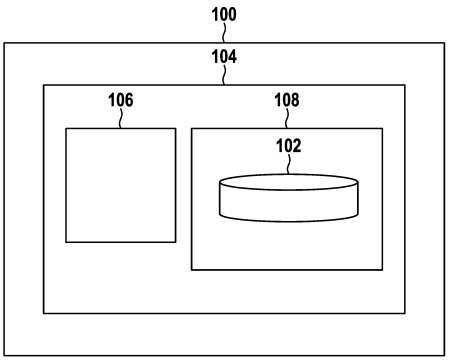
審査請求　未請求　請求項の数　14　　ＯＬ　外国語出願　（全15頁）

(21)出願番号　特願2023-75489(P2023-75489)
(22)出願日　令和5年5月1日(2023.5.1)
(31)優先権主張番号　10 2022 204 296.4
(32)優先日　令和4年5月2日(2022.5.2)
(33)優先権主張国・地域又は機関
ドイツ(DE)

(71)出願人　390023711
ローベルト ボツシユ ゲゼルシヤフト ミツト ベシユレンクテル ハフツング
ＲＯＢＥＲＴ ＢＯＳＣＨ ＧＭＢＨ
ドイツ連邦共和国 シユツツトガルト （番地なし）
Ｓｔｕｔｔｇａｒｔ，Ｇｅｒｍａｎｙ
(74)代理人　100114890
弁理士　アインゼル・フェリックス＝ライインハルト
(74)代理人　100098501
弁理士　森田 拓
(74)代理人　100116403
弁理士　前川 純一
(74)代理人　100134315
最終頁に続く

(54)【発明の名称】　知識データベースへのデータ投入のために非構造化テキストに属する時間表現を処理するための装置及び方法

(57)【要約】
【課題】知識データベース（１０２）へのデータ投入のために特に非構造化テキストに属する時間表現を処理するための方法及び装置に関する。
【解決手段】本方法及び本装置においては、テキスト内の時間表現が決定され、テキストに依存して時間表現の型が決定され、時間表現及び型が、時間表現の非文脈依存表現における時間表現の値の予測にマッピングされる。
【選択図】図１



【特許請求の範囲】
【請求項１】
知識データベース（１０２）へのデータ投入のために特に非構造化テキストに属する時間表現を処理するためのコンピュータ実装された方法において、
テキスト内の時間表現が決定され（３０４）、
前記テキストに依存して前記時間表現の型が決定され（３０６）、
前記時間表現及び前記型が、前記時間表現の非文脈依存表現における前記時間表現の値の予測にマッピングされる（３０８）
ことを特徴とする方法。
【請求項２】
前記時間表現及び前記型を前記予測にマッピングするようにモデルが構成されており、前記時間表現及び前記型が、前記モデルによって前記予測にマッピングされる（３０８）、
請求項１に記載の方法。
【請求項３】
前記時間表現を含むテキストが準備され（３００）、
前記テキストに対して基準年月日が決定又は設定され（３０２）、
前記非文脈依存表現に依存してかつ前記基準年月日に依存して、時間特性が決定され（３１０）、
前記時間特性が前記知識データベースに投入される（３１２）、
請求項１又は２に記載の方法。
【請求項４】
時間表現が準備され（２０２）、
非文脈依存表現が準備され（２０４）、
前記時間表現を前記非文脈依存表現に対する予測にマッピングするように、前記モデルが前記時間表現と前記非文脈依存表現とに依存してトレーニングされる（２０８）、
請求項２又は３に記載の方法。
【請求項５】
前記非文脈依存表現は、複数のスロットを含み、前記時間表現は、前記スロットの内容によって文脈に依存せずに表され、前記複数のスロットのうちの少なくとも１つがマスキングされ（２０６）、
前記少なくとも１つのマスキングされたスロットのマスキングされた内容に対する予測に前記時間表現をマッピングするように、前記モデルが前記時間表現に依存してトレーニングされる（２０８）、
請求項４に記載の方法。
【請求項６】
少なくとも１つのスロットがマスキングされておらず、かつ、少なくとも１つのスロットがマスキングされている第１のトレーニングフェーズにおいて、前記時間表現と、マスキングされていない前記少なくとも１つのスロットのマスキングされていない内容とを、マスキングされている前記少なくとも１つのスロットのマスキングされている内容に対する予測にマッピングするように、前記モデルがトレーニングされる（２０８）、
請求項５に記載の方法。
【請求項７】
前記第１のトレーニングフェーズにおいてトレーニングされた前記モデルが、前記スロットが特にすべてマスキングされている第２のトレーニングフェーズにおいて、前記時間表現を前記スロットの内容に対する予測にマッピングするようにトレーニングされる（２０８）、
請求項６に記載の方法。
【請求項８】
前記モデルは、複数の反復においてトレーニングされ、第１の反復中、第２の反復中よ

りも、特に前記第１の反復後に実施される第２の反復中よりも、少ないスロットがマスキングされる（２０６）、
請求項４乃至７のいずれか一項に記載の方法。
【請求項９】
前記モデルにより、予測として、数字若しくは数値若しくは演算子、特に算術演算子、又は、文字若しくは単語構成部分、又は、予め定められた語彙からの単語若しくはワイルドカードが決定される（２０８，３０８）、
請求項１乃至８のいずれか一項に記載の方法。
【請求項１０】
前記モデルにより１つのスロットのマスキングされた内容ごとに、語彙に関する確率分布が決定され、前記確率分布に依存して前記スロットの内容が決定される（２０８，３０８）、
請求項９に記載の方法。
【請求項１１】
特にルールに基づいて前記テキストに依存して、前記テキストから前記時間表現を含む抜粋テキストが決定され（３０４）、
前記予測が前記モデルにより、前記時間表現を含む前記抜粋テキストに依存して決定される（３０８）、
請求項１乃至１０のいずれか一項に記載の方法。
【請求項１２】
特にルールに基づいて前記テキストに依存して、前記テキストから前記時間表現を含む抜粋テキストが決定され（２０２）、前記モデルが前記時間表現を含む前記抜粋テキストに依存してトレーニングされる、
請求項１乃至１１のいずれか一項に記載の方法。
【請求項１３】
知識データベース（１０２）へのデータ投入のために特に非構造化テキストに属する時間表現を処理するための装置（１００）であって、
当該装置（１００）は、請求項１乃至１２のいずれか一項に記載の方法を実施するように構成された計算装置を含む、ことを特徴とする装置（１００）。
【請求項１４】
コンピュータプログラムであって、当該コンピュータプログラムは、コンピュータ可読命令を含み、前記コンピュータ可読命令がコンピュータにより実行されるときに、当該コンピュータにより請求項１乃至１２のいずれか一項に記載の方法が実施される、ことを特徴とするコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、知識データベースへのデータ投入のために非構造化テキストに属する時間表現を処理するための装置及び方法に関する。
【背景技術】
【０００２】
知識データベースへのデータ投入のために、非構造化テキストに属する時間表現の抽出及び正規化を想定することができる。
【０００３】
たとえば、Chang & Manning著による２０１２年刊の“SUTime: A library for recognizing and normalizing time expressions”（https://aclanthology.org/L12-1122/）、又は、Stroetgen & Gertz著による２０１３年刊の“Multilingual and cross-domain temporal tagging”（https://link.springer.com/article/10.1007/s10579-012-9179-y）に記載されているように、両方のサブタスクを一緒に解決することができる。

【０００４】
Bethard著による２０１３年刊の“A Synchronous Context Free Grammar for Time Normalization”（https://aclanthology.org/D13-1078/）、又は、Lee等著による２０１４年刊の“Context-dependent Semantic Parsing for Time Expressions”（https://aclanthology.org/P14-1135/）に記載されているように、両方のサブタスクを互いに切り離して処理することができる。
【０００５】
これらの正規化アプローチ、及び、Ning等著による２０１８年刊の“CogCompTime: A Tool for Understanding Time in Natural Language”(https://aclanthology.org/D18-2013/)、及び、Ding等著による２０２１年刊の“Automatic rule generation for time expression normalization”(https://aclanthology.org/2021.findings-emnlp.269)に記載されている正規化アプローチは、ルール体系に基づくものである。
【先行技術文献】
【非特許文献】
【０００６】
【非特許文献１】Chang & Manning著、２０１２年刊、“SUTime: A library for recognizing and normalizing time expressions”（https://aclanthology.org/L12-1122/）
【非特許文献２】Stroetgen & Gertz著、２０１３年刊、“Multilingual and cross-domain temporal tagging”（https://link.springer.com/article/10.1007/s10579-012-9179-y）
【非特許文献３】Bethard著、２０１３年刊、“A Synchronous Context Free Grammar for Time Normalization”（https://aclanthology.org/D13-1078/）
【非特許文献４】Lee等著、２０１４年刊、“Context-dependent Semantic Parsing for Time Expressions”（https://aclanthology.org/P14-1135/）
【非特許文献５】Ning等著、２０１８年刊、“CogCompTime: A Tool for Understanding Time in Natural Language”（https://aclanthology.org/D18-2013/）
【非特許文献６】Ding等著、２０２１年刊、“Automatic rule generation for time expression normalization”（https://aclanthology.org/2021.findings-emnlp.269）
【発明の概要】
【課題を解決するための手段】
【０００７】
発明の開示
知識データベースへのデータ投入のために非構造化テキストに属する時間表現を正規化する、コンピュータ実装された独立請求項１による方法は、ディープラーニングに基づく時間表現正規化方法を含む。
【０００８】
ルールに基づく正規化に対するディープラーニングに基づく正規化方法の利点は、新しい言語及びテキストジャンルへの転用が可能であることであり、このような転用は、さもなければ新たなルールの作成により多大な労力をかけることでしか担保することができないであろう。
【０００９】
決められたルールによる正規化とは異なり、ディープラーニングに基づく正規化方法によれば、スタティックなルール体系によって担保されない時間表現も正規化することができる。即ち、ディープラーニングに基づく正規化方法は、テキスト中の障害源に関して、ルールに基づくアプローチよりもフレキシブルであり、そのような障害源とは、たとえば比較的軽微な正書法上の誤りや未知の単語などであるが、それらの意味は、文脈から推定できるものである。

【００１０】
ディープラーニングに基づく正規化方法は、明示的にトレーニング済みである必要がない種々の抽出モデルと組み合わせて使用することができる。例示的な抽出モデルにおいて、ルールに基づくアプローチを用いて、又は、ニューラルネットワークを用いて、抽出を行うことができる。抽出をユーザによって行うことができ、正規化すべき表現に対する境界をユーザ自身が設定する。
【００１１】
知識データベースへのデータ投入のために、特に非構造化テキストに属する時間表現を処理するための方法は、テキスト内の時間表現が決定され、テキストに依存して時間表現の型が決定され、さらに時間表現及び型が、時間表現の非文脈依存表現における時間表現の値の予測にマッピングされるように構成されている。時間表現及び／又は型は、たとえば、ルールに基づいて、又は、そのためにトレーニングされたニューラルネットワークを用いて、テキストに依存して決定される。
【００１２】
好ましくは、時間表現及び型を予測にマッピングするようにモデルが構成されており、時間表現及び型が、このモデルによって予測にマッピングされる。時間表現の値は、たとえば、モデルとしてこのためにトレーニングされた変換器によって決定され、変換器は、時間表現及び型を予測にマッピングするように構成されており、即ち、そのようにトレーニングされている。変換器の代わりに、このためにトレーニングされた他のモデル、たとえば、ディープラーニングモデル、特に、リカレントニューラルネットワーク又は畳み込みニューラルネットワークを設けることもでき、これは、時間表現及び型を予測にマッピングするように構成されており、即ち、そのようにトレーニングされている。
【００１３】
好ましくは、時間表現を含むテキストが準備され、このテキストに対して基準年月日が決定又は設定され、非文脈依存表現に依存してかつ基準年月日に依存して時間特性が決定され、さらに、この時間特性が知識データベースに投入される。これにより、知識データベースに、テキストにより予め与えられた時間的文脈において結び付けられている時間特性が投入される。
【００１４】
好ましくは、時間表現が準備され、非文脈依存表現が準備され、さらに時間表現を非文脈依存表現に対する予測にマッピングするように、モデルが時間表現と非文脈依存表現とに依存してトレーニングされる。結果として得られる変換器は、テキスト中の障害源に関してフレキシブルであり、そのような障害源とは、たとえば正書法上の誤りや未知の単語などであるが、それらの意味は、文脈から推定できるものである。結果として得られるモデル、たとえば、変換器又はディープラーニングモデル、及び、非文脈依存表現の決定は、複数の言語に適用することができる。
【００１５】
非文脈依存表現は、好ましくは複数のスロットを含み、時間表現は、スロットの内容によって文脈に依存せずに表され、複数のスロットのうちの少なくとも１つがマスキングされ、さらに少なくとも１つのマスキングされたスロットのマスキングされた内容に対する予測に時間表現をマッピングするように、モデルが時間表現に依存してトレーニングされる。これにより、文脈依存表現に比較して、必要とされるトレーニングデータが少なくなるので、モデルが非常に効率的にトレーニングされる。
【００１６】
好ましくは、少なくとも１つのスロットがマスキングされておらず、かつ、少なくとも１つのスロットがマスキングされている第１のトレーニングフェーズにおいて、時間表現と、マスキングされていない少なくとも１つのスロットのマスキングされていない内容とを、マスキングされている少なくとも１つのスロットのマスキングされている内容に対する予測にマッピングするように、モデルがトレーニングされる。これにより、モデルが非常に効率的にトレーニングされる。

【００１７】
好ましくは、第１のトレーニングフェーズにおいてトレーニングされたモデルが、スロットが特にすべてマスキングされている第２のトレーニングフェーズにおいて、時間表現をスロットの内容に対する予測にマッピングするようにトレーニングされる。これによって、モデルが非常に効率的にトレーニングされ、非文脈依存表現の完全な予測を学習する。
【００１８】
好ましくは、モデルは、複数の反復においてトレーニングされ、第１の反復中、第２の反復中よりも、特に第１の反復後に実施される第２の反復中よりも、少ないスロットがマスキングされる。これにより、たとえば様々な言語において、ごく僅かなトレーニングデータしか存在しない場合に特に、モデルが非常に効率的にトレーニングされる。
【００１９】
好ましくは、モデルによって、非文脈依存表現のスロットに対する予測として、数字若しくは数値若しくは演算子、特に算術演算子、又は、文字若しくは単語構成部分、又は、予め定められた語彙からの単語若しくはワイルドカードが決定される。それぞれ異なる非文脈依存表現のために、それぞれ異なる語彙を使用することができる。語彙は、所望の非文脈依存表現に従って予め定められる。これにより、モデルによる予測が非常に信頼できるものとなる。
【００２０】
モデルによって、好ましくは１つのスロットのマスキングされた内容ごとに、語彙に関する確率分布が決定され、この確率分布に依存してスロットの内容が決定される。これにより、モデルが非常に効率的にトレーニングされ、継続的な解析又は代替的な予測を確率分布に基づき行うことができる。
【００２１】
ここで想定されることは、特に、ルールに基づいてテキストに依存して、テキストから時間表現を含む抜粋テキストが決定され、さらに予測がモデルにより、時間表現を含む抜粋テキストに依存して決定されるようにすることである。抜粋テキストは、たとえば１つの文である。予測は、その文脈における時間表現に基づいている。これによって、予測が付加的に改善される。
【００２２】
ここで想定されることは、特に、ルールに基づいてテキストに依存して、テキストから時間表現を含む抜粋テキストが決定され、モデルが、時間表現を含む抜粋テキストに依存してトレーニングされるようにすることである。
【００２３】
知識データベースへのデータ投入のために特に非構造化テキストに属する時間表現を処理するための装置は、上述の方法を実施するように構成された計算装置を含む。この装置は、上述の方法に相応する利点を有する。
【００２４】
コンピュータ可読命令を含み、それらの命令がコンピュータによって実行されるときに、コンピュータによって上述の方法が実施されるコンピュータプログラムは、それ相応の利点を有する。
【００２５】
以下の説明及び図面から、さらに他の有利な実施形態を見て取ることができる。
【図面の簡単な説明】
【００２６】
【図１】知識データベースへのデータ投入のために非構造化テキストに属する時間表現を処理するための装置を示す概略図である。
【図２】知識データベースへのデータ投入のために非構造化テキストに属する時間表現を処理するための方法におけるトレーニングステップを示す図である。
【図３】知識データベースへのデータ投入のために非構造化テキストに属する時間表現を

処理するための方法における推論ステップを示す図である。
【発明を実施するための形態】
【００２７】
図１には、知識データベース１０２へのデータ投入のために特に非構造化テキストに属する時間表現を処理するための装置１００が概略的に示されている。
【００２８】
時間表現の例はたとえば、年月日表記、たとえば２０２２年１月、及び、時刻表記、たとえば１３時、並びに、期間、たとえば5ヶ月間である。
【００２９】
知識データベース１０２へのデータ投入が想定されている適用事例の例は、知識グラフ、英語ではknowledge base、又は、時間特性を有するデータベースの作成若しくは拡張である。たとえば、時間表現に基づきエントリに対する有効期間をそれらに設定することができ、それらが表すイベントを特定の時点に対応づけることができる。
【００３０】
装置１００は、計算装置１０４を含み、この計算装置１０４は、知識データベース１０２へのデータ投入のために非構造化テキストに属する時間表現を処理するための、以下において説明する方法を実施するように構成されている。
【００３１】
計算装置１０４はたとえば、少なくとも１つのプロセッサ１０６及び少なくとも１つのメモリ１０８を含む。少なくとも１つのメモリ１０８は、この実施例においては、知識データベース１０２を格納するように構成されている。少なくとも１つのメモリ１０８は、この実施例においては、命令、たとえばコンピュータプログラムを格納するように構成されており、それらの命令が少なくとも１つのプロセッサ１０６によって実行されるときに、以下において説明する方法が実施される。
【００３２】
この方法について、以下においては、時間表現の抽出及びそれに続く正規化の例を挙げて説明する。抽出及び正規化は、「時間タグづけ」の一部である。
【００３３】
抽出は、この実施例においては、型の決定を含む。あるテキストが与えられたならば、抽出の役割は、時間表現の境界を見出すことである。そのような境界は、通常、いわゆるキャラクタオフセットであり、これらは、２つの数字によりテキスト内の表現の一義的な境界を規定する。たとえば、１１キャラクタ長の表現の場合には、キャラクタ番号２０において開始し、キャラクタ番号３０（これを含む）において終了する。
【００３４】
抽出は、たとえば、ルールに基づいて実行される。たとえば、ある表現が予め定められたルールに従って時間表現のためのパターンと一致している場合には、その表現は、時間表現として識別される。
【００３５】
抽出は、たとえばディープラーニングのためのアプローチを用いて、特に、系列ラベリングタスク又はSeq2Seq問題を解決することによって、実行される。
【００３６】
Laparra等著による２０１８年刊の“From Characters to Time Intervals: New Paradigms for Evaluation and Neural Parsing of Time Normalizations”（https://aclanthology.org/Q18-1025/）、及び、Lange等著による２０２０年刊の“Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text”（https://aclanthology.org/2020.repl4nlp-1.14/）には、系列ラベリングタスクの解決による抽出について記載されている。
【００３７】
この方法について、時間表現を含むテキストの例を挙げて説明する。
【００３８】

型の決定は、たとえばHeidelTimeによって実行される。HeidelTimeについてはたとえば、Stroetgen & Gertz著による２０１５年刊の“A Baseline Temporal Tagger for all Languages”（https://aclanthology.org/D15-1063/）に記載されている。

【００３９】

時間表現の型の決定はたとえば、４つのクラス（DATE, TIME, DURATION, SET）のうちの１つにおいて行われる。他のクラス、より多いクラス又はより少ないクラスを想定するものとしてもよい。

【００４０】

この実施例においては、時間表現の型の決定は、時間表現の抽出と同時に行われる。時間表現が抽出され、次いで、型が決定されるようにすることも想定される。

【００４１】

抽出及び型の決定の出力として、この実施例においては、時間表現ごとにテキスト内のその表現のアノテーション境界及び型が決定される。

【００４２】

例示的な抜粋テキスト“Der Eta Aquariiden Meteorschauer wird im Mai erscheinen.”（「みずがめ座η流星群は、５月に出現する」）の場合には、たとえば時間表現Mai（５月）が抽出され、型DATEによって型が決定される。

【００４３】

この抜粋テキストにおいて、この実施例においては、時間表現MaiがHeidelTimeによって抽出され、型DATEによって抜粋テキスト自体にアノテーションが付される。結果として得られるアノテーションが付された抜粋テキストは、“Der Eta Aquariiden Meteorschauer wird im 〈TIMEX3 type=“DATE”〉Mai〈/TIMEX3〉 erscheinen.”となる。

【００４４】

正規化の過程における時間表現は、Maiの例においては、時間表現の非文脈依存表現によって補われる。非文脈依存表現は、この実施例においては、TIMEX 3表現内の時間表現と共に、抜粋テキスト内の時間表現の代わりに使用される。この実施例の場合には、その結果として以下の抜粋テキストが生じる：

“Der Eta Aquariiden Meteorschauer wird im 〈TIMEX3 type=“DATE” value=“UNDEF-year-05”〉Mai〈/TIMEX3〉 erscheinen.”

【００４５】

TIMEX 3についてはたとえば、James Pustejovsky等著による“The specification language TimeML”、（The language of time: a reader, pages 545-557, Oxford University Press）に記載されている。

【００４６】

この実施例では、HeidelTimeにおいて、テキスト内の文字Mのポジション：START = 42とテキスト内の文字iのポジション：END = 44とを有する時間表現Maiに、時間表現の非文脈依存表現kuReprであるkuRepr = “UNDEF-year-05”が割り当てられる。

【００４７】

非文脈依存表現KuReprに関して、この実施例においては、TIMEX 3表現における１つの値が設けられている。

【００４８】

TIMEX 3表現を含む抜粋テキストは、この実施例においては、正規化のための入力量として使用される。

【００４９】

正規化は、この実施例においては、Masked Language Modelingを含む。このことが意味することは、正規化がMasked Language Modelingタスクとしてモデリングされる、ということである。Masked Language Modelingは、たとえばBERTのような言語モデルによる変換器を含む。

【００５０】
BERTについてはたとえば、Devlin等著による２０１９年刊の“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”（https://aclanthology.org/N19-1423/）に記載されている。
【００５１】
変換器の代わりに、このためにトレーニングされた他のモデル、たとえば、ディープラーニングモデル、特に、リカレントニューラルネットワーク又は畳み込みニューラルネットワークを設けることもできる。
【００５２】
Masked Language Modelingタスクの場合、非文脈依存表現における特定の領域がマスクによってマスキングされ、マスクを満たすように変換器がトレーニングされる。
【００５３】
変換器のトレーニングにおいて、この変換器は、時間表現に関して、その非文脈依存表現に対する予測を、時間表現に依存して決定するようにトレーニングされる。
【００５４】
変換器による非文脈依存表現の予測のために、この実施例においては、複数のスロットが使用される。
【００５５】
TIMEX 3表現の値における非文脈依存表現kuReprは、この実施例においては、スロットの内容によって表される。
【００５６】
１つのスロットの内容は、この実施例においては、非文脈依存表現の一部又はワイルドカードを含む。複数のスロットは、この実施例においては、予め定められた不変の長さを有する。複数のスロットは、この実施例においては、同等の長さを有する。それぞれ異なる長さのスロットを定義するものとしてもよい。これらのスロットは、この実施例においては、それぞれ１つの異なるセマンティックな意味を有する。
【００５７】
年月日表記のためにたとえば、２つの数字を有する日表記、及び、２つの数字を有する月表記のためにそれぞれ１つのスロット、並びに、年表記のために２つのスロットが設けられる。年表記のためにはたとえば、１の位及び１０の位の年のための２つの数字を有するスロットと、１００の位と１０００の位の年のための２つの数字を有するスロットとが設けられる。この実施例においては、スロットの順序が定義されている。スロットの内容は、時間表現に関するその意味に従って、一義的に対応づけられている。１日未満の時間表記は、たとえば、３つの他のスロットに格納され、即ち、時間表記のための２つの数字を有するスロットと、分表記のための２つの数字を有するスロットと、秒表記のための２つの数字を有するスロットとに格納される。ここで想定されることは、少なくとも１つのさらに別のスロットが設けられているようにすることである。少なくとも１つのさらに別のスロットの内容は、関数呼び出しとすることができる。関数呼び出しは、「プラス」又は「マイナス」といった演算子によって表すことができる。少なくとも１つのさらに別のスロットの内容は、関数呼び出しのための引数とすることができる。少なくとも１つのさらに別のスロットの内容は、付加情報を含み得る。付加的な情報は、たとえば、年月日表記を紀元前若しくは紀元後とする目印、又は、少なくとも１つのさらに別のスロットを期間若しくは祝日とする目印である。たとえば、内容ＢＣ又はＡＤは、紀元前又は紀元後の年表記である。たとえば４つのスロットの内容は、明後日のためのthis-day-plus-2又は昨日のためのthis-day-minus-1である。たとえば、スロットの内容は祝日を表す。他の表現、たとえばUNDEF、のためにはスロットは設けられていない。なぜならば、この表現は、「次の」又は「最後の」といった特定の単語を含む時間表現が処理される場合に、暗黙的に採用されるからである。
【００５８】
スロットの内容及び非文脈依存表現の一部への予め定められた対応づけによって、非文

脈依存表現を再構成することができる。
【００５９】
この実施例においては、使用されていないスロットには、ワイルドカードが設けられる。使用されないスロットは、非文脈依存表現に値が含まれていないこと又は非文脈依存表現に値が含まれないことを変換器が予測するスロットである。
【００６０】
この実施例においては、１１個のスロットが用いられる。ここで想定されることは、他の個数のスロットを使用することである。この変換器は、時間表現に依存してスロットの内容を予測するように設計されている。この実施例においては、変換器は、時間表現を含む抜粋テキストに依存して、スロットの内容を予測するように設計されている。この実施例においては、変換器は、TIMEX 3表現を含む抜粋テキストに依存して、スロットの内容を予測するように設計されている。
【００６１】
トレーニングにおいては、１つ又は複数のスロットがマスクによってマスキングされる。変換器は、マスキングされた１つのスロット又はマスキングされた複数のスロットに対して、可能な限り一致する予測を決定するようにトレーニングされる。
【００６２】
TIMEX 3表現を用いた抜粋テキストは、この実施例においては、変換器に供給され、個々のマスクに対して想定される挿入内容に関する確率分布が、出力として計算される。この実施例においては、変換器は、重み、たとえば、重みを有する人工ニューラルネットワークを含む。この実施例においては、変換器の重みを適応させる目的で逆伝播が用いられる。
【００６３】
この実施例においては、変換器による時間表現の正規化のために、その時間表現の非文脈依存表現が予測される。ここで想定されることは、変換器を用いて、TIMEX 3表現における具体的な値、即ち、非文脈依存表現kuReprを決定することである。この実施例においては、Mai（５月）という時間表現に対して、たとえば月表記のためのスロットの内容０５が予測され、その際に他のスロットの内容はワイルドカードである。この場合においては、内容０５は、ルールに基づいて非文脈依存表現 kuRepr = “UNDEF-year-05”にマッピングされる。
【００６４】
時間表現に対して予測された非文脈依存表現は、この実施例においては、正規化された表現を成している。
【００６５】
知識データベースへのデータ投入は、この実施例においては、２つの部分的側面を含む：
【００６６】
１．時間表現がテキスト内の相対的表現である場合、この表現は、予め定められた基準年月日に結び付けられる。たとえば、「昨日」という時間表現は、予め定められた基準年月日を用いて意味に即して決定される。
【００６７】
２．時間表現が、たとえばクリスマスイブといったようなテキストに属する暗黙的表現である場合、この表現は、これが１２月２４日の晩であるというセマンティック知識によって解明される。
【００６８】
知識データベースは、この実施例においては、エンティティとリレーションとを含み、これらのリレーションによって、エンティティが互いに好ましくはペアとして結合されている。知識データベースの一例は、エンティティをノードとして、リレーションをノード間のエッジとして表す知識グラフである。知識データベースは、１つの実施例においては、１００個以上のエンティティを含む。知識データベースは、１つの実施例においては、

１０００個以上のエンティティを含む。知識データベースは、１つの実施例においては、１００００個以上のエンティティを含む。知識データベースは、１つの実施例においては、１００個以上のリレーションを含む。知識データベースは、１つの実施例においては、１０００個以上のリレーションを含む。知識データベースは、１つの実施例においては、１００００個以上のリレーションを含む。

【００６９】

テキストは、この実施例においては、少なくとも１つの時間表現を含む。テキストは、１つの実施例においては、１００個以上の時間表現を含む。テキストは、１つの実施例においては、１０００個以上の時間表現を含む。テキストは、１つの実施例においては、１００００個以上の時間表現を含む。テキストは、この実施例においては、１つの基準年月日を含む。テキストは、１つの実施例においては、１００個以上の基準年月日を含む。テキストは、１つの実施例においては、１０００個以上の基準年月日を含む。テキストは、１つの実施例においては、１００００個以上の基準年月日を含む。

【００７０】

知識データベース１０２へのデータ投入のために、特に非構造化テキストに属する時間表現を処理するための方法は、トレーニングステップと推論ステップとを含む。トレーニングステップの目的は、テキストに属する１つの時間表現に対して、又は、テキストに属する複数の時間表現に対して、それらの時間表現の非文脈依存表現を予測するように、変換器をトレーニングすることである。推論ステップの目的は、テキストに属する１つの時間表現に対して、又は、テキストに属する複数の時間表現に対して、それらの時間表現の非文脈依存表現を予測するように、変換器を使用することである。

【００７１】

知識データベースへのデータ投入の目的は、テキスト内の時間表現に対する推論ステップによって決定される１つ又は複数の非文脈非依存表現を、１つ又は複数のエンティティあるいは１つ又は複数のリレーションあるいはそれらの属性を知識データベースに追加するために用いるということである。

【００７２】

トレーニングステップは、推論ステップとは切り離して実施することができる。推論ステップを特に、すでにトレーニング済みの変換器を用いて、トレーニングステップとは切り離して実施することができる。

【００７３】

図２には、トレーニングステップが示されている。

【００７４】

ステップ２０２において、時間表現が準備される。

【００７５】

１つの実施例においては、抜粋テキスト、たとえば１つの文が、時間表現を含むテキストから決定される。

【００７６】

時間表現は、たとえばルールに基づいてテキストにおいて決定される。

【００７７】

ステップ２０４において、非文脈依存表現kuReprが準備される。

【００７８】

非文脈依存表現kuReprは、１つの実施例においては、複数のスロットを含む。時間表現は、スロットの内容により文脈に依存することなく表される。

【００７９】

１つの実施例においては、時間表現のためのTIMEX 3表現と、その表現の非文脈依存表現とを含む抜粋テキストが準備される。

【００８０】

ステップ２０６において、非文脈依存表現kuReprのスロットのうちの少なくとも１つがマスキングされる。

【００８１】
変換器は、１つの実施例においては、複数の反復においてトレーニングされる。第１の反復中、第２の反復中よりも少ないスロットがマスキングされる。第２の反復は、１つの実施例においては、第１の反復の後に実施される。
【００８２】
ステップ２０８において、非文脈依存表現kuReprと時間表現とが、変換器によって非文脈依存表現kuReprに対する予測にマッピングされる。１つの実施例においては、時間表現を含む抜粋テキストが、変換器によって予測にマッピングされる。１つの実施例においては、時間表現のためのTIMEX 3表現を含む抜粋テキストが、変換器によって予測にマッピングされる。
【００８３】
ここで想定されることは、第１のトレーニングフェーズにおいては、少なくとも１つのスロットがマスキングされておらず、少なくとも１つのスロットがマスキングされていることである。第１のトレーニングフェーズ中、時間表現と、マスキングされていない少なくとも１つのスロットの内容とが、変換器によって、マスキングされている少なくとも１つのスロットの内容に対する予測にマッピングされる。
【００８４】
ここで想定されることは、第２のトレーニングフェーズ中、複数のスロットが特にすべてマスキングされていることである。第２のトレーニングフェーズ中、時間表現が、第１のトレーニングフェーズ中にトレーニングされた変換器によって、スロットの内容に対する予測にマッピングされる。
【００８５】
変換器によって、１つの実施例においては、予測として、数字若しくは数値若しくは演算子、特に算術演算子、又は、文字若しくは単語構成部分、又は、予め定められた語彙からの単語若しくはワイルドカードが決定される。
【００８６】
抜粋テキスト“Der Eta Aquariiden Meteorschauer wird im 〈TIMEX3 type=“DATE” value=“UNDEF-year-05”〉Mai〈/TIMEX3〉 erscheinen.”は、この実施例においては、非文脈依存表現kuRepr = “UNDEF-year-05”を含む。変換器は、この実施例においては、そのために定義されているスロットについて、内容として数字０５をこのスロットのために設けられた変換器の出力によって予測するように、トレーニングされる。
【００８７】
以下においては、さらに他の予測の例を挙げておく。
【００８８】
抜粋テキスト「．．．明後日．．．」は、たとえば非文脈依存表現“UNDEF-this-day-PLUS-2”によって表され、この表現に対して、変換器によって、第１のスロットのために設けられた変換器出力においては、内容として単語“this”が予測され、第２のスロットのために設けられた変換器出力においては、内容として“day”が、第３のスロットのために設けられた変換器出力においては、内容として演算子PLUSが、第４のスロットのために設けられた変換器出力においては、内容として数字２が予測される。
【００８９】
変換器によって１つの実施例においては、１つのスロットのマスキングされた内容ごとに、語彙に関する確率分布が決定され、この確率分布に依存してスロットの内容が決定される。
【００９０】
非文脈依存表現が複数のスロットを含む場合、非文脈依存表現及び時間表現が変換器によって、少なくとも１つのマスキングされたスロットの内容に対する予測にマッピングされる。
【００９１】

ステップ２０２乃至２０８は、１つのトレーニング期間において、この実施例ではテキストに属する多数の時間表現に対して実施される。

【００９２】

ステップ２１０において、変換器の少なくとも１つの重みが決定される。この重みはたとえば、非文脈依存表現に対する予測と非文脈依存表現との偏差に依存して決定される。非文脈依存表現が複数のスロットを含む場合、複数のスロットのうちの少なくとも１つのスロットのマスキングされた内容に対する予測と、この少なくとも１つのスロットのマスキングされた内容との偏差に依存して、重みが決定される。重みは、この実施例においては、テキストに属する多数の時間表現に対して決定される多数の偏差に依存して、逆伝播により決定される。

【００９３】

ステップ２１０は、この実施例においては、多数のトレーニング期間にわたり実行される。

【００９４】

図３には、推論ステップが示されている。推論のために変換器は、時間表現及び型を、時間表現の非文脈依存表現に対する予測にマッピングするように構成されている。

【００９５】

ステップ３００において、時間表現を含むテキストが準備される。

【００９６】

ステップ３０２において、テキストに対して基準年月日が決定又は設定される。たとえば基準年月日は、テキストから特にルールに基づいて決定される。

【００９７】

ステップ３０４において、テキスト内の時間表現が、特にルールに基づいて決定される。

【００９８】

１つの実施例においては、特にルールに基づいてテキストに依存して、時間表現を含むテキストから抜粋テキストが決定される。

【００９９】

ステップ３０６において、時間表現の型がテキストに依存して、特にルールに基づいて決定される。

【０１００】

ステップ３０８において、時間表現及び型が、変換器によって非文脈依存表現に対する予測にマッピングされる。

【０１０１】

変換器によって、１つの実施例においては、予測として、数字若しくは数値若しくは演算子、特に算術演算子、又は、文字若しくは単語構成部分、又は、予め定められた語彙からの単語若しくはワイルドカードが決定される。

【０１０２】

変換器によって１つの実施例においては、スロットごとに語彙に関する確率分布が決定され、この確率分布に依存してそのスロットの内容が決定される。

【０１０３】

１つの実施例においては、変換器による予測は、時間表現を含む抜粋テキストに依存して決定される。

【０１０４】

ステップ３１０において、非文脈依存表現に依存してかつ基準年月日に依存して、時間特性が決定される。

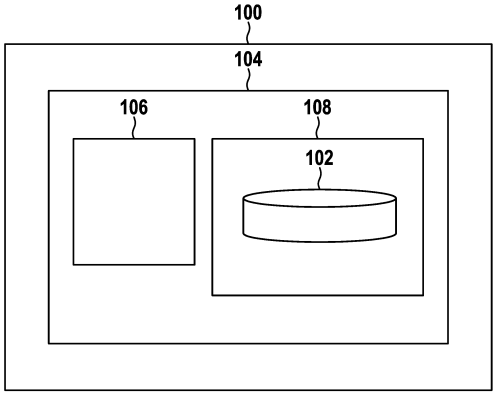
【０１０５】

ステップ３１２において、知識データベースに時間特性が投入される。

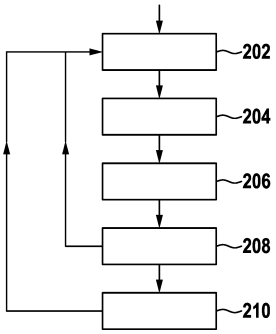
【０１０６】

次いで、ステップ３０４が実施される。

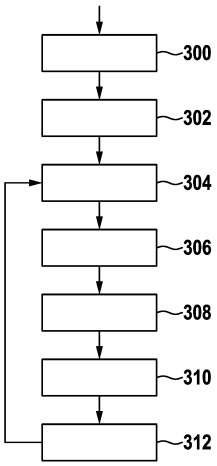
【０１０７】
ステップ３０４乃至３１２は、この実施例においては、テキスト内に存在する時間表現が処理されるまで実施される。
【図面】
【図１】



【図２】



【図３】



【外国語明細書】
20231654100000005.pdf

フロントページの続き

弁理士　永島 秀郎
(74)代理人　100162880
弁理士　上島 類
(72)発明者　ルーカス ランゲ
ドイツ連邦共和国 プフォルツハイム プフリューガーシュトラーセ ４７
(72)発明者　ヤニク シュトレートゲン
ドイツ連邦共和国 カールスルーエ ブライテ シュトラーセ ３２
(72)発明者　ハイケ アーデル－ヴー
ドイツ連邦共和国 レニンゲン ヒンターリートシュトラーセ ４／１

Ｆターム（参考）　5B175　FB02