(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5717103号
(P5717103)
(45)発行日　平成27年5月13日(2015.5.13)　(24)登録日　平成27年3月27日(2015.3.27)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 220Z
G06F 17/30 330C

請求項の数 8　(全 16 頁)

(21)出願番号　特願2012-143175(P2012-143175)
(22)出願日　平成24年6月26日(2012.6.26)
(65)公開番号　特開2014-6802(P2014-6802A)
(43)公開日　平成26年1月16日(2014.1.16)
審査請求日　平成26年6月9日(2014.6.9)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(73)特許権者　304021417
国立大学法人東京工業大学
東京都目黒区大岡山２丁目１２番１号
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　東中　竜一郎
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　松尾　義博
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】文書間関係推定装置、方法、及びプログラム

(57)【特許請求の範囲】
【請求項１】
時間情報及びユーザ情報が各々付与された複数の文書からなる文書群における全ての文書間の各々について、文書間の内容の関連性と、文書間の応答関係、及び文書間における何れか一方の文書が同一のユーザ情報が付与された文書のうちの先頭または最後の文書であるか否かを示す特徴を抽出する特徴抽出手段と、
前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記特徴に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推定する関係推定手段と、
を含む文書間関係推定装置。
【請求項２】
文書群の各文書について、前記文書が他の文書と関係がある場合、前記文書と関係がある前記他の文書が１つである制約を生成する制約生成手段を更に含み、
前記関係推定手段は、前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記制約生成手段によって生成された前記制約と、前記特徴及び前記制約に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推定する請求項１記載の文書間関係推定装置。
【請求項３】
前記特徴は、文書間の特徴を示す述語であって、

前記関係推定手段は、
前記推定モデルとして、前記特徴を示す述語又は前記推定される前記文書間の関係を示す潜在述語を用いて記述された各論理式を用いて構築されるマルコフ論理ネットワーク（ＭＬＮ：Markov Logic Network）における各論理式の重みを用いて、前記マルコフ論理ネットワークにより、前記制約生成手段によって生成された前記文書間に対する前記制約を記述した論理式を満足し、かつ、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴を示す述語に対して尤もらしい、前記文書間の関係を示す潜在述語を推定する請求項２記載の文書間関係推定装置。
【請求項４】
前記文書間の特徴は、前記文書間のうちの先の文書が後の文書の返信先であること、後の文書が先の文書のユーザに対して返信していること、前記文書のユーザ情報が同じであること、先の文書が同じユーザ情報が付与された文書のうちの最後の文書であること、及び先の文書が同じユーザ情報が付与された文書のうちの最初の文書であることの少なくとも一つを含む請求項１～請求項３の何れか１項記載の文書間関係推定装置。
【請求項５】
時間情報及びユーザ情報が各々付与された複数の学習用文書からなる学習用文書群における全ての学習用文書間の各々について、前記特徴を抽出する学習用特徴抽出手段と、
前記学習用文書群における全ての学習用文書間の各々について予め定められた文書間の関係と、前記学習用特徴抽出手段によって前記学習用文書群における全ての学習用文書間の各々について抽出された前記学習用文書間に対する前記特徴とに基づいて、前記推定モデルを学習する学習手段と、
を更に含む請求項１～請求項４の何れか１項記載の文書間関係推定装置。
【請求項６】
前記学習用文書群における各学習用文書について、前記学習用文書が他の学習用文書と関係がある場合、前記学習用文書と関係がある前記他の学習用文書が１つである制約を生成する学習用制約生成手段を更に含み、
前記学習手段は、前記学習用文書群における全ての学習用文書間の各々について予め定められた文書間の関係と、前記学習用特徴抽出手段によって前記学習用文書群における全ての学習用文書間の各々について抽出された前記学習用文書間に対する前記特徴と、前記学習用制約生成手段によって生成された各学習用文書に対する前記制約とに基づいて、前記推定モデルを学習する請求項５記載の文書間関係推定装置。
【請求項７】
特徴抽出手段によって、時間情報及びユーザ情報が各々付与された複数の文書からなる文書群における全ての文書間の各々について、文書間の内容の関連性と、文書間の応答関係、及び文書間における何れか一方の文書が同一のユーザ情報が付与された文書のうちの先頭または最後の文書であるか否かを示す特徴を抽出し、
関係推定手段によって、前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記特徴に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推定する
文書間関係推定方法。
【請求項８】
コンピュータを、請求項１～請求項６の何れか１項記載の文書間関係推定装置を構成する各手段として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書間関係推定装置、方法、及びプログラムに係り、特に、文書間の関係を推定する文書間関係推定装置、方法、及びプログラムに関する。
【背景技術】

【０００２】
インターネット上では様々なユーザが発言を投稿しており、たとえば、掲示板やマイクロブログサービスでは、ユーザが日々発言を投稿し、やりとりをしている。このようなデータは非常に膨大であり、構造化されていないため、効率的に閲覧することが難しい。そこで、発言間の関係を同定し、構造化する手法が提案されている。たとえば、非特許文献１では、ＱＡサイトに投稿された発言間の関係性をマルコフロジックネットワーク（ＭＬＮ）と呼ばれる教師あり学習の手法で関係づけを行っている。関係性としては、「類似」（発言同士が同様の内容を保持している）や「包含」（片方の発言がもう片方の内容を完全に含み、新たな内容も含んでいる）などである。
【０００３】
非特許文献１では、発言の特徴、および、発言間の特徴から、発言間の関係を同定している。ここで用いられているのは発言の内容に基づく特徴である。具体的には、特徴として、どちらが長いか、連続する発言かどうか、発言間の投稿間隔、反意語となる単語対が発言間にあるか、同じURLを含むかどうか、同じ固有表現を含むか否か、括弧で囲まれた同じ表現を含むかどうか、異なる固有名詞を含むかどうか、単語のコサイン類似度、名詞の包含度を用い、関係性の正解データから、関係性推定のモデルをＭＬＮによって学習している。ＭＬＮについては、非特許文献２に詳述されている。ＭＬＮは、確率的に推論を行う仕組みとして、近年注目されている。
【０００４】
ＭＬＮでは重み付きの述語を扱うことができ、このため、かならず成り立つわけではないような関係も論理的な関係と同時に扱うことがでる。現実的な、おおよそ成り立つ関係について、重みを学習によって決定し、推定に役立てることが可能な学習手法である。
【先行技術文献】
【非特許文献】
【０００５】
【非特許文献１】Hikaru Yokono; Takaaki Hasegawa; Genichiro Kikui; Manabu Okumura, Identification of relations between answers with global constraints for Community-based Question Answering services, Proc. IJCNLP, 2011年.
【非特許文献２】吉川克正、浅原正幸、松本裕治、「Markov Logic による日本語述語項構造解析」、情報処理学会研究報告（NL-199）、2010年.
【発明の概要】
【発明が解決しようとする課題】
【０００６】
インターネット上の発言には、発言対象のユーザ名などが入っていたり、掲示板やマイクロブログサービスによっては、発言者のユーザＩＤや発言同士が応答関係にあるかどうかといった関係性がシステムによって付与されていたりすることが多い。このような情報は、関係性の同定に有効だと考えられるが、従来は用いられておらず、関係性の推定性能が低くなる、という問題があった。
【０００７】
本発明は、上記の事情を鑑みてなされたもので、文書間の関係を精度よく推定することができる文書間関係推定装置、方法、及びプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００８】
上記の目的を達成するために本発明に係る文書間関係推定装置は、時間情報及びユーザ情報が各々付与された複数の文書からなる文書群における全ての文書間の各々について、文書間の内容の関連性と、文書間の応答関係、及び文書間における何れか一方の文書が同一のユーザ情報が付与された文書のうちの先頭または最後の文書であるか否かを示す特徴を抽出する特徴抽出手段と、前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記特徴に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推

定する関係推定手段と、を含んで構成されている。
【０００９】
本発明に係る文書間関係推定装置は、文書群の各文書について、前記文書が他の文書と関係がある場合、前記文書と関係がある前記他の文書が１つである制約を生成する制約生成手段を更に含み、前記関係推定手段は、前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記制約生成手段によって生成された前記制約と、前記特徴及び前記制約に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推定するようにすることができる。
【００１０】
本発明に係る文書間関係推定方法は、特徴抽出手段によって、時間情報及びユーザ情報が各々付与された複数の文書からなる文書群における全ての文書間の各々について、文書間の内容の関連性と、文書間の応答関係、及び文書間における何れか一方の文書が同一のユーザ情報が付与された文書のうちの先頭または最後の文書であるか否かを示す特徴を抽出し、関係推定手段によって、前記文書群における全ての文書間の各々について、前記特徴抽出手段によって抽出された前記文書間に対する前記特徴と、前記特徴に基づいて文書間の関係を推定するための予め学習された推定モデルとに基づいて、前記文書間の関係を推定する。
【００１１】
本発明に係るプログラムは、コンピュータを、上記の文書間関係推定装置の各手段として機能させるためのプログラムである。
【発明の効果】
【００１２】
以上説明したように、本発明の文書間関係推定装置、方法、及びプログラムによれば、文書間の各々について、文書間の内容の関連性と、文書間の応答関係、及び何れか一方の文書が同一のユーザ情報が付与された文書のうちの先頭または最後の文書であるか否かを示す特徴を抽出し、予め学習された推定モデルに基づいて、文書間の関係を推定することにより、文書間の関係を精度よく推定することができる、という効果が得られる。
【図面の簡単な説明】
【００１３】
【図１】本発明の実施の形態に係る発言間関係推定装置の構成を示す概略図である。
【図２】入力されるツイート集合の一例を示す図である。
【図３】本発明の実施の形態に係る発言間関係推定装置におけるモデル学習処理ルーチンの内容を示すフローチャートである。
【図４】本発明の実施の形態に係る発言間関係推定装置における発言間関係推定処理ルーチンの内容を示すフローチャートである。
【図５】実験結果を示す図である。
【発明を実施するための形態】
【００１４】
以下、図面を参照して本発明の実施の形態を詳細に説明する。
【００１５】
＜システム構成＞
図１に示すように、本発明の実施の形態に係る発言間関係推定装置１００は、発言を示すテキストデータの集合が入力され、各発言間の関係を出力する。１つの発言は１つ以上の文からなるテキストデータである。この発言間関係推定装置１００は、ＣＰＵと、ＲＡＭと、後述するモデル学習処理ルーチン及び発言間関係推定処理ルーチンを実行するためのプログラムを記憶したＲＯＭとを備えたコンピュータで構成され、機能的には次に示すように構成されている。図１に示すように、発言間関係推定装置１００は、入力部１０と、演算部２０と、出力部３０とを備えている。
【００１６】

入力部１０は、入力された発言の集合を受け付ける。本実施の形態の例では、発言の集合として、マイクロブログサービスの一つであるツイッター（Ｒ）の発言集合を用いる。ツイッター（Ｒ）では、ユーザが日々膨大な数の発言（ツイートと呼ばれる）を行っており、構造化が望まれるデータの一つである。ここでは、ツイートの集合をtogetter（Ｒ）（http://togetter.com/）と呼ばれるサービスから収集した。togetter （Ｒ）は、個人が自身のお気に入りのツイートを「まとめ」として登録することのできるサービスで、一定のトピックに関係するツイートが雑多に集められている。

【００１７】

また、学習データとして入力された発言の集合と共に、入力部１０は、入力された発言間の関係を受け付ける。入力される発言間の関係は、手動で付与されたものであり、使用する関係性は、例えば、「関係あり」である。

【００１８】

また、使用する関係性として詳細なものを用いてもよく、例えば、「賛成」及び「反対」、並びに、「矛盾」、「類似」、「演繹」及び「帰納」である。賛成は同意を表し、反対は、同意していないことを表す。矛盾は、異なる内容を述べていることを表し、類似は、発言に同意をして、類似した内容を述べていることを表す。演繹は、発言の内容をもとにして、推論による内容を発言したり、新たな情報を加え議論の展開することを表し、帰納は、発言の内容をまとめた内容を発言していることを表す。

【００１９】

本実施の形態の具体例として、togetter（Ｒ）における14のまとめページから、14のツイート集合を収集した。また、実験のため、これらの集合をそのまま用いるのではなく、まとめのトピックとは直接関係のないと思われるツイートは削除し、リプライ先が含められていないツイートがあった場合には、twitter（Ｒ）から改めて取得するという処理を行った。図２は、あるツイート集合からの抜粋である。IDはツイートのID、ユーザ名はツイートをしたユーザの名前、返信先は、ツイッターによって付与される返信先情報であり、in_reply_toから取得できる。本文はツイートの発言そのものであり、リンク先IDは、in_reply_to からでは判定できないが、内容から確認できる、発言対象を表すＩＤである。賛否は、関係性のある発言との間の関係としての「賛成」「反対」のいずれかであり、関係は、関係性のある発言との間の関係としての「矛盾」「類似」「演繹」「帰納」のいずれかである。リンク先ID、賛否、関係は、学習データとして人手で付与した発言間の関係の一例である。

【００２０】

また、入力部１０は、上記の関係性がツイート間にあるかないか、あるとしたらどの関係かを推定するために入力された、発言の集合を受け付ける。

【００２１】

演算部２０は、発言集合データベース２１、発言間関係データベース２２、特徴量生成部２５、モデル学習部２６、モデル記憶部２７、入力発言集合データベース２８、特徴量生成部２９、及び関係同定部３１を備えている。なお、関係同定部３１が、関係推定手段の一例である。

【００２２】

発言集合データベース２１は、入力部１０により受け付けた学習データとしての発言の集合を記憶する。発言間関係データベース２２は、入力部１０により受け付けた学習データとしての発言間の関係を記憶する。

【００２３】

特徴量生成部２５は、形態素解析部２５１、固有表現抽出部２５２、発言間特徴生成部２５３、及び発言間制約生成部２５４を備え、発言集合データベース２１に記憶されている発言の全ペアについてペア間の特徴を生成する。本実施の形態の例では、発言間の特徴は一階述語論理として表現される。特徴量は、述語として表される。

【００２４】

形態素解析部２５１は、各発言に対して形態素解析を行う。形態素解析の手法は、従来

既知の手法を用いればよく、入力を形態素に分割できるものであれば何でもよい。固有表現抽出部２５２は、各発言から固有表現を抽出する。固有表現抽出の手法は、従来既知の手法を用いればよく、入力について、固有表現を抽出できるものであれば何でも良い。本実施の形態の例では、どちらについてもCaboCha を用いている。

【００２５】

発言間特徴生成部２５３は、ツイートの発言間（ツイートiとツイートj）の各々について、後述する「in_reply_to」、「reply」、「sameuser」、「latestutt」、及び「firstutt」を含む述語を、発言間の特徴（素性とも呼ぶ）として生成する。

【００２６】

「in_reply_to」：ツイートjにリプライ先のツイートのID＝iが指定されている場合、特徴として、述語in_reply_to(i、j)を生成する。たとえば、先の例における、ツイート2_userB_1 はin_reply_to にID 1_userA_1が指定されているため、述語in_reply_to(1(1_userA_1)、2 (2_userB_１))が特徴として生成される。

【００２７】

「reply」：ツイートjが@...という形でツイートiのユーザー名に言及している場合、reply(i、j)を生成する。たとえば、ツイートｊ＝44_userB_10の中で@で言及されているユーザのツイートのIDを用いて、述語reply(29(29_userX_17)、44(44_userB_10))、述語reply(33(33_ userX _19)、44(44_userB_10))といった特徴を生成する。

【００２８】

「sameuser」：ツイートjとツイートiが同ユーザの場合、述語sameuser(i、j)を生成する。たとえば、2_userB_1と4_userB_2のように、同じユーザのツイート同士であれば、述語sameuser(2(2_userB_1)、4(4_userB_2))を生成する。

【００２９】

「latestutt」：ツイートjとツイートiの間にツイートiと同じユーザのツイートがない場合、つまり、相手の一番最近のツイートに対してリンクがある場合、述語latestutt(i、j)を生成する。たとえば、ツイート2のユーザの発言が、ツイート3の時点で最も最近のそのユーザの発言である場合、述語latestuut(2(2_userB_1)、3(3_userC_1))を生成する。すなわち、会話中で誰かの発言に対して応答を行う場合、相手の新しい発言を無視して、以前の発言に対して応答することは少ないだろう、という状況を表す意図で、ツイートが発言された時点で、応答相手の最新の発言に対しての応答かどうかを表す述語を生成する。

【００３０】

「firstutt」：ツイートjがリンクをもつツイートiが所定の区間内でそのユーザの最初の発言である場合、述語firstutt(i、j)を生成する。firstuttは、相手のユーザの最初のツイートとの組みであることを表す述語である。たとえば、ツイートiのＩＤが「i_ユーザ名_1」となっている場合、 述語firstutt(i、j)を生成する。質問や問題提起など、会話の発端となるようなツイートが広く様々なユーザから言及される場合がtogetter（Ｒ）上で多く見られるため、togetter（Ｒ）のまとめ上でそれぞれのユーザの初めての発言はより多くのユーザーから応答されやすいだろう、という意図でこの述語を生成する。なお、所定の区間内というのは、応答関係を判断するツイート集合内、例えば、togetter（Ｒ）のまとめ全体の範囲を表わす。

【００３１】

なお、述語「in_reply_to」が、先の文書が後の文書の返信先であることを示す特徴の一例であり、述語「reply」が、後の文書が先の文書のユーザに対して返信していることを示す特徴の一例であり、述語「sameuser」が、文書のユーザ情報が同じであることを示す特徴の一例である。また、述語「latestutt」が、先の文書が同じユーザ情報が付与された文書のうちの最後の文書であることを示す特徴の一例であり、述語「firstutt」が、先の文書が同じユーザ情報が付与された文書のうちの最初の文書であることを示す特徴の一例である。

【００３２】

また、発言間特徴生成部２５３は、従来法で素性として用いられた、発言ペアのどちらが長いか、発言ペアが連続する発言かどうか、発言間の投稿間隔、反意語となる単語対が発言間にあるか、発言ペアが同じURLを含むかどうか、発言ペアが同じ固有表現を含むか否か、発言ペアが括弧で囲まれた同じ表現を含むかどうか、発言ペアが異なる固有名詞を含むかどうか、発言間における単語ベクトルのコサイン類似度、発言間の名詞の包含度の各々を表わす述語を、発言間の特徴として、発言間の各々について生成する。

【００３３】

固有表現は、人名、製品名、施設名、地名、時間表現、数値表現を扱っており、個々の固有表現ごとの素性ではなく、固有表現の種類毎に固有表現を含むかどうかを表わす述語を生成する。反意語に関する述語は、予め準備した反意語リストに従って、一方が、もう一方の反意語を含むかを表す述語を生成する。たとえば、好評に対して悪評と不評、重んじるに対して軽んじる、夏至に対して冬至、などが反意語のリストである。発言間の間隔については離散値の特徴であり、出現順の差が、5以下、10以下、15以下、20以下、30以下、40以下、50以下、それ以上であるかどうかを表わす述語を生成する。名詞の包含度はツイートi、j間の両方向に対して定義され、ツイートiのツイートjに対する包含度は、（共通する名詞の異なり数）/（ツイートiに出現した名詞の数）であり、発言間の間隔と同様に、離散値の特徴として述語を生成する。単語ベクトルのコサイン類似度は、それぞれのツイートについて単語unigram、単語bigramそれぞれのベクトルを作成し、そのベクトル間のコサイン類似度を計算し、発言間の間隔と同様に、離散値（例えば、０．１刻みの離散値）の特徴として述語を生成する。

【００３４】

発言間制約生成部２５４は、発言間の制約を示す論理式を生成する。ツイッターでは、一つの発言が短く、複数の発言に一度に応答することは少ない。よって、発言が他の発言と関係がある場合、関係がある他の発言が１つである、という制約を表す論理式を、各発言について生成する。例えば、以下に示す論理式が生成される。

【００３５】

```
for Id i
  if tweet(i) : |Id k:has_aa_relation(k,i)|<=1;
```

上記のように制約の論理式が記述される。これで、has_aa_relation(k,i)を満たすk が最大一つとなる。

【００３６】

また、発言間制約生成部２５４は、上記の制約に加え、基本的な制約として、推移律と呼ばれる、ツイートi とjにある関係R が成り立ち、ツイートjとkに同じくRが成り立つ場合、ツイートiとkにも同様の関係Rが成り立つという制約を生成する。たとえば、あるツイートに類似している２つのツイートj、k は類似しているという制約を生成する。

【００３７】

例えば、以下に示す論理式である。

【００３８】

```
aa_relation(i,j,"類似") & aa_relation(j,k,"類似") => aa_relation(i,k,"類似")
```

上記の論理式は、iとjが類似、jとkが類似ならばiとkも類似であることを表わす。

【００３９】

上記の制約を表わす論理式があれば、他のツイート間の関係から別のツイート間の関係を推定することができる。

【００４０】

また、推移律ではない制約として、以下に示すような論理式を生成してもよい。

【００４１】

```
tweet(i)& !tweet(j) => !has_aa_relation(i,j)
```

上記の論理式は、ツイートとツイートでない定数の間には応答関係は存在しないことを表わす。

【００４２】

上記の制約を表わす論理式があれば、現実的に意味をもたない状況を出力から除くことができる。

【００４３】

モデル学習部２６は、特徴量生成部２５が出力した特徴（観測可能な述語の集合）及び発言間関係データベース２２に記憶された発言間の関係（ラベル）を示す潜在述語の集合から、各述語及び潜在述語を用いて記述された各論理式の重みを学習する。学習にはＭＬＮを用いる。ＭＬＮでは、ラベル間に対して予め記述した制約を表わす論理式、及び特徴量とラベルとの間に対して予め記述した制約を表わす論理式を満たしつつ、観測可能な述語について行われた推論結果が、正解データとして与えられた述語の集合に近づくように、各論理式の重みが決定される。ここで学習された論理式の重みの集合がモデルとなり、重みの集合を表わすモデルが、モデル記憶部２７に記憶される。なお、ラベルは、発言間関係データベース２２に記憶された発言間の関係である。

【００４４】

ＭＬＮの学習について簡単に説明する。ＭＬＮは、述語の集合Ｘに対してある確率を与える。この確率を最大化する述語の集合（潜在述語を含む）が、ＭＬＮにおける推論結果となる。ＭＬＮは、実際には一階述語論理式の集合として表される。論理式には、違反を許容する重み付きの論理式と、違反を許容しない論理式を混在させることができ、制約を表わす論理式は、違反を許容しない論理式として記述される。

【００４５】

述語の集合Ｘに対する確率は、述語を論理式中の変数に代入することで成立する(真となる)論理すべてについて重みの和をとり、exponentialを取り、正規化したものである。例えば、以下の（１）式で表される。

【００４６】

【数１】

$$p(X)=\frac{1}{Z}\exp\left(\sum_i w_i n_i(X)\right) \quad \cdots (1)$$

【００４７】

ここで、Zは正規化のための定数であり、w_iがi番目の論理式に対応する重みである。n_iは、ある述語の集合X上で真をとる論理式、例えば、tweet(i) & tweet(j) & in_reply_to(i,j) のそれぞれについて、カウントが１増える。論理式の一部のみが真でもカウントはされない。この場合、論理式 tweet(i) & tweet(j) & in_reply_to(i,j) が真になる回数n_iは、X中に、この論理式が真になる i と j の組が存在する数と同じとなる。

【００４８】

なお、モデル学習部２６は、素性として生成された観測可能な述語の集合、発言間関係データベース２２に記憶された発言間の関係を表わす潜在述語（ラベル）を用いた論理式を、以下のように生成しておく。

【００４９】

例えば、ツイートi、j、kがあるとき、以下のような述語が観測可能な述語として与えられる。

【００５０】

tweet(i)　（i はツイートである）
tweet(j) tweet(k) in_reply_to(i、j)　（jがi に対してリプライしている）
in_reply_to(j、k)　（kがjに対してリプライしている）
has_span(i、j、"1-5")　（iとｊの間隔が1~5の間）

【００５１】

加えて、in_reply_to 属性が付いていて、has_aa_relation が成り立っているならば（

真であれば）、tweet(i) & tweet(j) & in_reply_to(i,j) => has_aa_relation(i,j) という論理式を用意する。

【００５２】

また、上記のような論理式は、以下に説明するように、述語に対する論理式のテンプレートに従って、重みと共に用意される。

【００５３】

例えば、ツイートiとツイートjの間にin_reply_to属性がついていて、i、j間にhas_aa_relation が成り立っているならば（真であれば）、w_inreplyto という重みを得ることを示す以下のような記述が、人手によって与えられる。

【００５４】

if tweet(i)&tweet(j)&in_reply_to(i、j)　add[has_aa_relation(i、j)]*w_inreplyto

【００５５】

これらは「tweet(i) & tweet(j) & in_reply_to(i,j)」ならば「has_aa_relation(i,j)」という意味の論理式がすべてのiとjにあてはまるツイートに対して用意される。また、それぞれの重みw_inreplytoがＭＬＮに含まれる。

【００５６】

また、ＭＬＮの学習では、has_aa_relation(i、j)、および、has_aa_relation(j、k)となったとすると、テンプレートで記述した論理式のうち、成立した論理式の重みをw_inreplyto*2 +w_hasspanのように計算し、確率を計算する。（w_inreplytoはi、jとj、kで二度真になっているため）このような、素性として与えられていない述語は潜在述語とよばれ、これが、いわゆる分類問題で言う出力するラベルにあたる。

【００５７】

本実施の形態の例では、正例（正解データの潜在述語を含む述語集合）と負例（それ以外の述語集合）間のマージンを最大化する学習を行って、各述語に対する論理式の重みを求める。なお、尤度を最大化する学習を行ってもよい。

【００５８】

入力発言集合データベース２８は、入力部１０により受け付けた推定対象データとしての発言の集合を記憶する。入力されたツイート集合、具体的には、ツイッターから取得できる情報である、ID、発言内容、in_reply_to属性、及びユーザ名からなるデータの集合が、入力発言集合データベース２８に記憶される。

【００５９】

特徴量生成部２９は、上記の特徴量生成部２５と同様に、形態素解析部２５１、固有表現抽出部２５２、発言間特徴生成部２５３、及び発言間制約生成部２５４を備え、入力発言集合データベース２８に記憶されている発言の全ペアについてペア間の特徴を生成する。

【００６０】

関係同定部３１は、入力発言集合データベース２８に記憶されている発言のペアの各々について、生成された特徴(観察可能な述語集合)に対して、モデル記憶部２７に記憶されたモデルを用いて、尤もらしい関係性ラベル（潜在述語）の集合を得る。具体的には、上記（１）式中の尤度を最大化する述語集合を得る。これらが推定結果であり、同定された関係である。

【００６１】

関係同定部３１により推定された潜在述語が表わす発言の各ペアの関係性を、出力部３０により出力する。

【００６２】

＜発言間関係推定装置の作用＞

次に、本実施の形態に係る発言間関係推定装置１００の作用について説明する。まず、発言の集合と、各発言間に対して手動で付与した発言間の関係を示すラベルの集合とが発言間関係推定装置１００に入力されると、発言間関係推定装置１００によって、入力された、発言の集合が、発言集合データベース２１へ格納され、入力された、発言間の関係を

示すラベルが、発言間関係データベース２２に格納される。
【００６３】
そして、発言間関係推定装置１００によって、図３に示すモデル学習処理ルーチンが実行される。
【００６４】
まず、ステップＳ１０１において、発言集合データベース２１に格納された各発言に対して、形態素解析処理を行う。次のステップＳ１０２では、発言集合データベース２１に格納された各発言から、固有表現を抽出する。
【００６５】
そして、ステップＳ１０３において、発言集合データベース２１に格納された発言の全ペアの各々に対して、発言間の特徴である述語を生成する。次のステップＳ１０４では、予め定められた制約を表わす論理式を生成する。
【００６６】
ステップＳ１０５において、発言間関係データベース２２に格納された発言の各ペアの関係を示すラベルを用いて、潜在述語を生成し、発言集合データベース２１に格納された発言の全ペアの各々に対して観察可能な述語の集合及び潜在述語からなる学習データを生成する。
【００６７】
そして、ステップＳ１０６において、上記ステップＳ１０５で生成した学習データに基づいて、上記ステップＳ１０３で生成した制約を表わす論理式を満足し、かつ、上記（１）式で表される確率が最大となる、発言の各ペアの関係を示す潜在述語及び各述語を用いた各論理式の重みを学習する。次のステップＳ１０７では、上記ステップＳ１０６で学習された各論理式の重みを、モデル記憶部２７に格納して、モデル学習処理ルーチンを終了する。
【００６８】
そして、発言間の関係を推定する推定対象の発言集合が発言間関係推定装置１００に入力されると、発言間関係推定装置１００によって、入力された発言集合が、入力発言集合データベース２８へ格納される。
【００６９】
そして、発言間関係推定装置１００によって、図４に示す発言間関係推定処理ルーチンが実行される。
【００７０】
ステップＳ１１１において、入力発言集合データベース２８に格納された各発言に対して、形態素解析処理を行う。次のステップＳ１１２では、入力発言集合データベース２８に格納された各発言から、固有表現を抽出する。
【００７１】
そして、ステップＳ１１３において、入力発言集合データベース２８に格納された発言の全ペアの各々に対して、発言間の特徴である述語を生成する。次のステップＳ１１４では、予め定められた制約を表わす論理式を生成する。
【００７２】
ステップＳ１１５において、発言の全ペアの各々に対して、上記ステップＳ１１３で生成された発言間の特徴である述語の集合、上記ステップＳ１１４で生成された制約の論理式、及び学習されたモデル（各論理式の重み）に基づいて、当該ペアに対する潜在述語を推定することにより、当該ペアの関係を推定する。
【００７３】
そして、ステップＳ１１６において、上記ステップＳ１１５で推定された各ペアの関係を出力部３０により出力して、発言間関係推定処理ルーチンを終了する。
【００７４】
＜実施例＞
学習データを元にＭＬＮによってモデルを学習し、交差検定によって、上記の実施の形

態で説明した手法の評価を行った。ここでは、三つの条件で比較した。３つの条件は、ツイッター向け素性有り＆制約有り、ツイッター向け素性無し＆制約あり、ツイッター向け素性無し＆ツイッター向け制約無しである。ツイッター向け素性とは、上記で説明した「in_reply_to」、「reply」、「sameuser」、「latestutt」、及び「firstutt」を含む述語である。ツイッター向け制約とは、上記の実施の形態で説明した、発言は関係を持つ先が一つであるという制約である。なお、ここでは、発言間に関係があるかどうか（ツイート間について、has_aa_relationが真か）を推定することを行う。その他の潜在ラベルについても同様の推定は可能である。

【００７５】

図５に実験結果を示す。ここで、Recallは再現率を表し、関係性があるツイートペアのうち、いくつを正しく関係があると判定できたかを表す。Precisionは適合率を表し、学習モデルに基づいて関係性があると推定したツイートペアのうち、いくつが実際に正しかったかを表す。F1はRecallとPrecisionの調和平均である。この値が高ければ、正確に漏れなく関係があることを推定できていると言える。F1値に着目すると、ツイッター向けの素性および制約を入れることで、性能が改善されていることが分かる。これにより、本実施の形態で提案する発言間の特徴の有効性が示された。

【００７６】

以上説明したように、本実施の形態に係る発言間関係推定装置によれば、発言（ツイート）間の各々について、発言間の内容の関連性を示す述語と、発言間の応答関係を示す述語と、先の発言が同一のユーザ情報が付与された発言のうちの最初の発言であることを示す述語と、先の発言が同一のユーザ情報が付与された発言のうちの最新の発言であることを示す述語とを、発言間の特徴として抽出すると共に、発言と関係がある他の発言が１つである制約を示す論理式を生成し、抽出された発言間の特徴と、生成された制約の論理式と、予め学習されたモデルとに基づいて、マルコフ論理ネットワークにより、発言間の関係を示す潜在述語を推定することにより、発言間の関係を精度よく推定することができる。

【００７７】

また、発言対象のユーザ名、発言者のユーザＩＤ、システムが付与する発言同士の応答関係を、発言間の関係性の同定に用いることで、関係性の同定精度を向上させる。発言間の関係性の同定精度が改善し、インターネット上の膨大な発言を高精度で構造化できるようになる。発言が高精度に構造化できれば、膨大な情報から効率的に内容を閲覧したり、情報を抽出したりすることが可能となる。

【００７８】

なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

【００７９】

例えば、発言以外の文書データ間の関係を推定することに、本発明を適用してもよい。

【００８０】

また、潜在述語として、関係があることを示す潜在述語を用いて、発言間に関係があるか否かを推定する場合を例に説明したが、詳細な関係性を示す潜在述語を用いてもよい。この場合には、例えば、「賛成」及び「反対」、並びに、「矛盾」、「類似」、「演繹」及び「帰納」の各々を示す潜在述語を追加して、発言間の関係性を推定するようにしてもよい。

【００８１】

また、本願明細書中において、プログラムが予めインストールされている実施形態として説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。

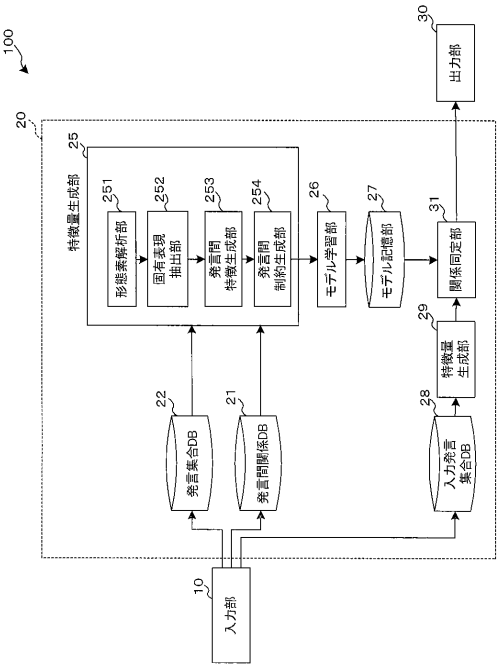
【符号の説明】

【００８２】

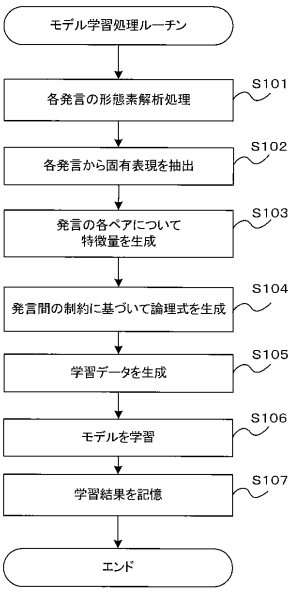
１０　入力部

２０　演算部
２１　発言集合データベース
２２　発言間関係データベース
２５　特徴量生成部
２６　モデル学習部
２７　モデル記憶部
２８　入力発言集合データベース
２９　特徴量生成部
３０　出力部
３１　関係同定部
１００　発言間関係推定装置
２５１　形態素解析部
２５２　固有表現抽出部
２５３　発言間特徴生成部
２５４　発言間制約生成部

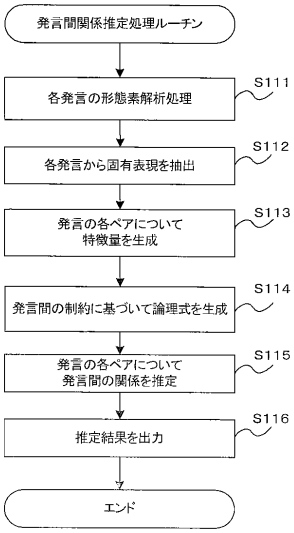
【図１】



【図３】



【図４】


発言間関係推定処理ルーチン
各発言の形態素解析処理
S111
各発言から固有表現を抽出
S112
発言の各ペアについて
特徴量を生成
S113
発言間の制約に基づいて論理式を生成
S114
発言の各ペアについて
発言間の関係を推定
S115
推定結果を出力
S116
エンド

【図２】

ID	ユーザー名	返信先	本文	リンク先ID	賛否関係
1_user A_1	user A		個人の書庫をデータ化するサービスなんじゃないかと。でも書籍のデータもこの会社に蓄積しますよね。 RT @ user E 安い！本１冊 100 円でスキャンして電子化するサービス http://bit.ly/cEwJw6		
2_user B_1	user B	1_user A_1	@ user A PDFをお渡しした後にスキャン漏れや、届いてないなどのトラブル回避のため、短時間は保管します。でも、権利関係の問題が生じるので基本的に蓄積はしません。 RT 個人の書庫をデータ化するサービスなんじゃないかと。でも書籍のデータもこの会社に蓄積しますよね。	1_user A_1	反対矛盾
3_user C_1	user C		OCR で文字も読み取っていたら iPhone でも読みやすそう。 RT @ user E 安い！本１冊 100 円でスキャンして電子化するサービス http://www.bookscan.co.jp/		
4_user B_2	user B		読みやすいですね。OCR も100%の認識率ではないですが、読めます。RT@user C ：OCR で文字も読み取っていたら iPhone でも読みやすそう。RT @ user E 安い！本１冊 100 円でスキャンして	3_user C_1	賛成演繹
5_user D_1	user D		これ、商売としてなりたつの？RT @ user E :安い！本１冊 100 円でスキャンして電子化するサービス http://bit.ly/cEwJw6		

【図５】

	Recall	Precision	F1
ツイッター向け素性有り＆制約有り	0.047	0.043	**0.045**
ツイッター向け素性なし＆制約有り	0.035	0.033	0.034
ツイッター　向け素性無し＆ツイッター向け制約無し	1.000	0.020	0.039

フロントページの続き

(72)発明者　森田　一
東京都目黒区大岡山２－１２－１　国立大学法人東京工業大学内
(72)発明者　奥村　学
東京都目黒区大岡山２－１２－１　国立大学法人東京工業大学内

審査官　野崎　大進

(56)参考文献　特開２０１１－１８０９８８（ＪＰ，Ａ）
特開２００７－２８７１３４（ＪＰ，Ａ）
米国特許出願公開第２００７／０２３３４６５（ＵＳ，Ａ１）
徳永 泰浩 他，チャット対話における発言間の継続関係と応答関係の同定，自然言語処理，日本，言語処理学会，２００５年　1月１０日，Vol.12,No.1，pp.79-105.
吉川 克正 他，機械学習手法による結合推論を利用した時間的順序関係推定，第７３回人工知能基本問題研究会資料，日本，社団法人人工知能学会，２００９年　3月　6日，SIG-FPAI-A804-12，pp.61-67.
宮部 泰成 他，文書横断文間関係の特定，言語処理学会第１２回年次大会発表論文集，日本，言語処理学会，２００６年　3月１３日，pp.496-499.
小西 卓哉 他，統計的言語特性を考慮した評判情報のトピックモデリング，第３回データ工学と情報マネジメントに関するフォーラム　論文集，日本，電子情報通信学会データ工学専門委員会，２０１１年　7月２７日，Vol.DEIM2011,No.A8-2，pp.1-8.

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩＩ）