

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5398007号
(P5398007)

(45) 発行日 平成26年1月29日 (2014. 1. 29)

(24) 登録日 平成25年11月1日 (2013. 11. 1)

(51) Int. Cl. F 1
G 0 6 F 17/27 (2006.01) G 0 6 F 17/27 Z

請求項の数 12 (全 39 頁)

(21) 出願番号	特願2010-42938 (P2010-42938)	(73) 特許権者	301022471
(22) 出願日	平成22年2月26日 (2010. 2. 26)		独立行政法人情報通信研究機構
(65) 公開番号	特開2011-180746 (P2011-180746A)		東京都小金井市貫井北町4-2-1
(43) 公開日	平成23年9月15日 (2011. 9. 15)	(74) 代理人	100115749
審査請求日	平成25年1月22日 (2013. 1. 22)		弁理士 谷川 英和
		(74) 代理人	100121223
			弁理士 森本 悟道
		(72) 発明者	土田 正明
			東京都小金井市貫井北町4-2-1 独立 行政法人情報通信研究機構内
		(72) 発明者	ステイン デ サーガ
			東京都小金井市貫井北町4-2-1 独立 行政法人情報通信研究機構内

最終頁に続く

(54) 【発明の名称】 関係情報拡張装置、関係情報拡張方法、及びプログラム

(57) 【特許請求の範囲】

【請求項1】

意味的关系を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部と、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的关系を有する確からしさを示すスコアであり、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアを取得するスコア取得部と、前記スコア取得部が取得した経由スコアを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該経由スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備えた関係情報拡張装置。

【請求項2】

意味的关系を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補

情報が記憶される拡張関係候補情報記憶部と、
前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、
前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を
生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情
報生成部と、

前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的關係を有する確
からしさを示すスコアであり、前記拡張関係候補情報に含まれる2以上の言語表現の共起
に関するスコアである共起スコアと、前記拡張関係候補情報が意味的關係を有する確から
しさを示すスコアであり、拡張関係候補情報がより多くの関係情報から得られるものであ
るほど、より高い値となるスコアである経由スコアとを取得するスコア取得部と、

前記スコア取得部が取得した共起スコアと経由スコアとを用いて、前記拡張関係候補情報
記憶部で記憶されている拡張関係候補情報のうち、当該共起スコアが高く、かつ、当該経
由スコアが高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備えた関係
情報拡張装置。

【請求項3】

意味的關係を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、
類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、
少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補
情報が記憶される拡張関係候補情報記憶部と、

前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、
前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を
生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情
報生成部と、

前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的關係を有する確
からしさを示すスコアであり、前記拡張関係候補情報に含まれる2以上の言語表現の共起
に関するスコアである共起スコアを取得するスコア取得部と、

前記スコア取得部が取得した共起スコアを用いて、前記拡張関係候補情報記憶部で記憶さ
れている拡張関係候補情報のうち、当該共起スコアの高い拡張関係候補情報である拡張関
係情報を選択する選択部と、を備え、

前記スコア取得部は、前記拡張関係候補情報に含まれる2以上の言語表現と、当該拡張関
係候補情報の生成時に用いられた関係情報の意味的關係と同じ種類の意味的關係を有する
各関係情報に含まれる2以上の言語表現に対して共起の高い言語表現である共起言語表現
とが共起する方が、前記拡張関係候補情報に含まれる2以上の言語表現のみが共起するよ
りも高い値となる共起スコアを取得する、関係情報拡張装置。

【請求項4】

前記関係情報は、当該関係情報が有する2以上の言語表現の意味的關係の種類を識別する
情報である種類識別情報をも有するものであり、

前記拡張関係候補情報生成部は、拡張関係候補情報の生成に用いる関係情報が有する種類
識別情報を有する拡張関係候補情報を生成し、

種類識別情報と、当該種類識別情報に対応する、当該種類識別情報で識別される意味的関
係の種類に対応する1以上の共起言語表現とを有する対応情報が1以上記憶される対応情
報記憶部をさらに備え、

前記スコア取得部は、前記拡張関係候補情報に含まれる2以上の言語表現と、当該拡張関
係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、前記拡張
関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコア
を取得する、請求項3記載の関係情報拡張装置。

【請求項5】

前記スコア取得部は、2以上の言語表現の組に含まれる当該2以上の言語表現と共起する
言語表現を少なくとも素性として用い、当該素性の値及び2以上の言語表現の組に対する
意味的關係の有無を教師データとする機械学習を行い、前記拡張関係候補情報に含まれる

10

20

30

40

50

2以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得する、請求項3記載の関係情報拡張装置。

【請求項6】

前記経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、当該拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアである、請求項1または請求項2記載の関係情報拡張装置。

【請求項7】

関係情報及び拡張関係候補情報は、第1の言語表現と第2の言語表現とである2個の言語表現を有するものであり、

前記スコア取得部は、ある拡張関係候補情報について、当該拡張関係候補情報と第2の言語表現が一致する各関係情報の第1の言語表現と、当該拡張関係候補情報の第1の言語表現との類似度の和である第1の計算値と、当該拡張関係候補情報と第1の言語表現が一致する各関係情報の第2の言語表現と、当該拡張関係候補情報の第2の言語表現との類似度の和である第2の計算値と、前記関係情報記憶部で記憶されている各関係情報と当該拡張関係候補情報との第1の言語表現同士の類似度及び第2の言語表現同士の類似度の積の和である第3の計算値とのうち、任意の1以上の計算値を引数とする増加関数の値である経由スコアを取得する、請求項6記載の関係情報拡張装置。

【請求項8】

意味的关系を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、

前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部と、

前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的关系を有する確からしさを示すスコアを取得するスコア取得部と、

前記スコア取得部が取得したスコアを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備え、

前記関係情報は、当該関係情報が有する2以上の言語表現の意味的关系の種類を識別する情報である種類識別情報をも有するものであり、

前記類似情報記憶部では、種類識別情報と、当該種類識別情報に対応する類似情報とが記憶されており、

前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を置換する際に、当該関係情報が有する種類識別情報に対応する類似情報を用いて置換を行う、関係情報拡張装置。

【請求項9】

意味的关系を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、拡張関係候補情報生成部と、スコア取得部と、選択部とを用いて処理される関係情報拡張方法であって、

前記拡張関係候補情報生成部が、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成ステップと、

前記スコア取得部が、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が

10

20

30

40

50

意味的關係を有する確からしさを示すスコアであり、拡張關係候補情報がより多くの關係情報から得られるものであるほど、より高い値となるスコアである經由スコアを取得するスコア取得ステップと、

前記選択部が、前記スコア取得ステップで取得した經由スコアを用いて、前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報のうち、当該經由スコアの高い拡張關係候補情報である拡張關係情報を選択する選択ステップと、を備えた關係情報拡張方法。

【請求項 10】

意味的關係を有する 2 以上の言語表現を有する關係情報が記憶される關係情報記憶部と、類似する 2 以上の言語表現を有する類似情報が 2 以上記憶される類似情報記憶部と、少なくとも 1 個の言語表現が類似する言語表現に置換された關係情報である拡張關係候補情報が記憶される拡張關係候補情報記憶部と、拡張關係候補情報生成部と、スコア取得部と、選択部とを用いて処理される關係情報拡張方法であって、

前記拡張關係候補情報生成部が、前記關係情報記憶部で記憶されている關係情報に含まれる少なくとも 1 個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張關係候補情報を生成し、当該拡張關係候補情報を前記拡張關係候補情報記憶部に蓄積する拡張關係候補情報生成ステップと、

前記スコア取得部が、前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報が意味的關係を有する確からしさを示すスコアであり、前記拡張關係候補情報に含まれる 2 以上の言語表現の共起に関するスコアである共起スコアと、前記拡張關係候補情報が意味的關係を有する確からしさを示すスコアであり、拡張關係候補情報がより多くの關係情報から得られるものであるほど、より高い値となるスコアである經由スコアとを取得するスコア取得ステップと、

前記選択部が、前記スコア取得ステップで取得した共起スコアと經由スコアとを用いて、前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報のうち、当該共起スコアが高く、かつ、当該經由スコアが高い拡張關係候補情報である拡張關係情報を選択する選択ステップと、を備えた關係情報拡張方法。

【請求項 11】

コンピュータを、

意味的關係を有する 2 以上の言語表現を有する關係情報が記憶される關係情報記憶部で記憶されている關係情報に含まれる少なくとも 1 個の言語表現を、類似する 2 以上の言語表現を有する類似情報が 2 以上記憶される類似情報記憶部で記憶されている類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張關係候補情報を生成し、当該拡張關係候補情報を、拡張關係候補情報が記憶される拡張關係候補情報記憶部に蓄積する拡張關係候補情報生成部、

前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報が意味的關係を有する確からしさを示すスコアであり、拡張關係候補情報がより多くの關係情報から得られるものであるほど、より高い値となるスコアである經由スコアを取得するスコア取得部、

前記スコア取得部が取得した經由スコアを用いて、前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報のうち、当該經由スコアの高い拡張關係候補情報である拡張關係情報を選択する選択部として機能させるためのプログラム。

【請求項 12】

コンピュータを、

意味的關係を有する 2 以上の言語表現を有する關係情報が記憶される關係情報記憶部で記憶されている關係情報に含まれる少なくとも 1 個の言語表現を、類似する 2 以上の言語表現を有する類似情報が 2 以上記憶される類似情報記憶部で記憶されている類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張關係候補情報を生成し、当該拡張關係候補情報を、拡張關係候補情報が記憶される拡張關係候補情報記憶部に蓄積する拡張關係候補情報生成部、

前記拡張關係候補情報記憶部で記憶されている拡張關係候補情報が意味的關係を有する確からしさを示すスコアであり、前記拡張關係候補情報に含まれる 2 以上の言語表現の共起

10

20

30

40

50

に関するスコアである共起スコアと、前記拡張関係候補情報が意味的關係を有する確からしさを示すスコアであり、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとを取得するスコア取得部、前記スコア取得部が取得した共起スコアと経由スコアとを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該共起スコアが高く、かつ、当該経由スコアが高い拡張関係候補情報である拡張関係情報を選択する選択部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、意味的關係を有する2以上の言語表現を有する関係情報を拡張する関係情報拡張装置等に関する。

【背景技術】

【0002】

従来、意味的關係を有する2以上の言語表現を有する関係情報を取得することが行われていた。ここで、意味的關係とは、例えば、原因や、上位下位等の関係である。したがって、関係情報は、例えば、死亡の原因が心筋梗塞であることを示す「原因<心筋梗塞、死亡>」や、頭痛薬の上位語が薬であることを示す「上位<薬、頭痛薬>」等である。その関係情報を取得する方法として、例えば、非特許文献1で開示されている方法を用いることができる。その非特許文献1では、例えば、「XがYの原因である」という表現のパターンを手がかりとして、文書からX、Yの部分を取得することにより、関係情報を取得することができた。しかしながら、そのようなパターンを用いた方法では、その手がかりとなるパターンで書かれた関係しか取得できないという問題があった。

【0003】

その問題を解決するための方法として、類推を用いて関係情報を取得する方法が、例えば、非特許文献2で開示されている。

【先行技術文献】

【非特許文献】

【0004】

【非特許文献1】P. Pantel, M. Pennacchiotti, 「Espresso: Leveraging generic patterns for automatically harvesting semantic relations」, In Proc. of the 21st COLING and 44th ACL (COLING - ACL - 06), p. 113 - 120, 2006年

【非特許文献2】石川大介、石塚英弘、藤原譲、「特許文献における因果関係を用いた類推による仮説の生成と検証 - ライフサイエンス分野を対象として - 」情報知識学会誌、Vol. 17、No. 3、p. 164 - 181、2007年

【発明の概要】

【発明が解決しようとする課題】

【0005】

しかしながら、非特許文献2では、関係情報に含まれる言語表現を置換する際に、語基が共通する表現にしか置換できないという問題があった。例えば、「ペプチド」を「抗菌ペプチド」に置き換えることはできるが、それ語基を共通にしないものには置換できなかった。また、置換後の関係情報には、不適切なものも含まれてしまうという問題があった。

【0006】

本発明は、上記問題を解決するためになされたものであり、既存の関係情報を拡張する場合に、関係情報に含まれる言語表現を、語基の共通しないものにも拡張することができ、また、適切な意味的關係を有する関係情報に拡張することができる関係情報拡張装置等を提供することを目的とする。

10

20

30

40

50

【課題を解決するための手段】

【0007】

上記目的を達成するため、本発明による関係情報拡張装置は、意味的关系を有する2以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する2以上の言語表現を有する類似情報が2以上記憶される類似情報記憶部と、少なくとも1個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を、類似情報を用いて、言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報を拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部と、拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的关系を有する確からしさを示すスコアを取得するスコア取得部と、スコア取得部が取得したスコアを用いて、拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備えたものである。

10

【0008】

このような構成により、言語表現を類似する言語表現に置換するため、語基の共通しないものに言語表現を置換することができるようになる。また、拡張関係候補情報のスコアを取得し、そのスコアの高いものを選択するため、その選択結果である拡張関係情報は、意味的关系を適切に有するものとなりうる。

【0009】

また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報に含まれる2以上の言語表現の共起に関するスコアである共起スコアを取得してもよい。

20

意味的关系を有する言語表現は、共起する可能性が高いと考えられるため、このような構成により、共起スコアの高いものを選択することによって、適切な選択を行うことができる。

【0010】

また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアを取得してもよい。

【0011】

適切な拡張関係情報は、多くの関係情報から得られるものであると考えられるため、このような構成により、経由スコアの高いものを選択することによって、適切な選択を行うことができる。

30

【0012】

また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報に含まれる2以上の言語表現の共起に関するスコアである共起スコアと、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとを取得し、選択部は、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択してもよい。

このような構成により、共有スコアと、経由スコアとの高いものを選択することによって、より適切な選択を行うことができる。

40

【0013】

また、本発明による関係情報拡張装置では、前記スコア取得部は、前記拡張関係候補情報に含まれる2以上の言語表現と、当該拡張関係候補情報の生成時に用いられた関係情報の意味的关系と同じ種類の意味的关系を有する各関係情報に含まれる2以上の言語表現に対して共起の高い言語表現である共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。

このような構成により、その共起スコアによって、拡張関係候補情報の生成で用いられた関係情報と同様の関係性を有するものほどより高い値となる共起スコアを取得できる。

【0014】

50

また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する2以上の言語表現の意味的關係の種類を識別する情報である種類識別情報をも有するものであり、前記拡張関係候補情報生成部は、拡張関係候補情報の生成に用いる関係情報が有する種類識別情報を有する拡張関係候補情報を生成し、種類識別情報と、当該種類識別情報に対応する、当該種類識別情報で識別される意味的關係の種類に対応する1以上の共起言語表現とを有する対応情報が1以上記憶される対応情報記憶部をさらに備え、前記スコア取得部は、前記拡張関係候補情報に含まれる2以上の言語表現と、当該拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。

10

このような構成により、対応情報を用いて、前述のような共起スコアを取得することができる。

【0015】

また、本発明による関係情報拡張装置では、前記スコア取得部は、2以上の言語表現の組に含まれる当該2以上の言語表現と共起する言語表現を少なくとも素性として用い、当該素性の値及び2以上の言語表現の組に対する意味的關係の有無を教師データとする機械学習を行い、前記拡張関係候補情報に含まれる2以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得してもよい。

このような構成により、機械学習を用いて、前述のような共起スコアを取得することができる。

20

【0016】

また、本発明による関係情報拡張装置では、経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアであってもよい。

【0017】

適切な拡張関係情報は、多くの関係情報から高い類似度で得られるものであると考えられるため、このような構成により、類似度も考慮して経由スコアを取得することができ、より適切な選択を行うことができる。

【0018】

また、本発明による関係情報拡張装置では、関係情報及び拡張関係候補情報は、第1の言語表現と第2の言語表現とである2個の言語表現を有するものであり、スコア取得部は、ある拡張関係候補情報について、拡張関係候補情報と第2の言語表現が一致する各関係情報の第1の言語表現と、拡張関係候補情報の第1の言語表現との類似度の和である第1の計算値と、拡張関係候補情報と第1の言語表現が一致する各関係情報の第2の言語表現と、拡張関係候補情報の第2の言語表現との類似度の和である第2の計算値と、関係情報記憶部で記憶されている各関係情報と拡張関係候補情報との第1の言語表現同士の類似度及び第2の言語表現同士の類似度の積の和である第3の計算値とのうち、任意の1以上の計算値を引数とする増加関数の値である経由スコアを取得してもよい。

30

このような構成により、関係情報が2個の言語表現を含む場合に、第1から第3の計算値の任意の1以上の計算値を用いて、経由スコアを取得することができる。

40

【0019】

また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する2以上の言語表現の意味的關係の種類を識別する情報である種類識別情報をも有するものであり、前記類似情報記憶部では、種類識別情報と、当該種類識別情報に対応する類似情報とが記憶されており、前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも1個の言語表現を置換する際に、当該関係情報が有する種類識別情報に対応する類似情報を用いて置換を行ってもよい。

このような構成により、関係情報にふさわしい類似情報を用いて、拡張関係候補情報を生成することができる。したがって、拡張関係候補情報の精度がより高いものとなり、

50

その結果、拡張関係情報の精度もより高くなりうる。

【0020】

また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する2以上の言語表現の意味的關係の種類を識別する情報である種類識別情報をも有するものであり、前記類似情報記憶部では、種類識別情報と、置換対象でない1以上の言語表現と、当該種類識別情報及び当該置換対象でない1以上の言語表現に対応する類似情報とが記憶されており、前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる1個の言語表現を置換する際に、当該関係情報が有する種類識別情報及び置換対象でない言語表現に対応する類似情報を用いて置換を行ってもよい。

このような構成により、関係情報や、置換対象でない言語表現にふさわしい類似情報を用いて、拡張関係候補情報を生成することができうる。したがって、拡張関係候補情報の精度がより高いものとなり、その結果、拡張関係情報の精度もより高くなりうる。

10

【発明の効果】

【0021】

本発明による関係情報拡張装置等によれば、既存の関係情報を用いて、新たな適切な意味的關係を取得することができる。

【図面の簡単な説明】

【0022】

【図1】本発明の実施の形態1による関係情報拡張装置の構成を示すブロック図

【図2】同実施の形態による関係情報拡張装置の動作を示すフローチャート

20

【図3】同実施の形態による関係情報拡張装置の動作を示すフローチャート

【図4】同実施の形態による関係情報拡張装置の動作を示すフローチャート

【図5】同実施の形態における類似情報の一例を示す図

【図6】同実施の形態における拡張関係候補情報等の一例を示す図

【図7】同実施の形態における拡張関係候補情報等の一例を示す図

【図8】同実施の形態における実験結果を示す図

【図9】同実施の形態による関係情報拡張装置の構成の他の一例を示すブロック図

【図10】同実施の形態における対応情報の一例を示す図

【図11】同実施の形態における類似情報記憶部で記憶されている情報の一例を示す図

【図12】同実施の形態における類似情報記憶部で記憶されている情報の一例を示す図

30

【図13】同実施の形態におけるサポートベクトルマシンについて説明するための図

【図14】同実施の形態におけるコンピュータシステムの外観一例を示す模式図

【図15】同実施の形態におけるコンピュータシステムの構成の一例を示す図

【発明を実施するための形態】

【0023】

以下、本発明による関係情報拡張装置について、実施の形態を用いて説明する。なお、以下の実施の形態において、同じ符号を付した構成要素及びステップは同一または相当するものであり、再度の説明を省略することがある。

【0024】

(実施の形態1)

40

本発明の実施の形態1による関係情報拡張装置について、図面を参照しながら説明する。本実施の形態による関係情報拡張装置は、既存の関係情報に含まれる少なくとも1個の言語表現を類似する言語表現に置換し、その置換後のものについてスコアを取得し、そのスコアの高いものを選択することによって、関係情報の拡張を行うものである。

【0025】

図1は、本実施の形態による関係情報拡張装置1の構成を示すブロック図である。本実施の形態による関係情報拡張装置1は、関係情報記憶部11と、類似情報記憶部12と、拡張関係候補情報生成部13と、拡張関係候補情報記憶部14と、スコア取得部15と、選択部16と、拡張関係情報記憶部17とを備える。

【0026】

50

関係情報記憶部 11 では、意味的關係を有する 2 以上の言語表現を有する関係情報が記憶される。関係情報記憶部 11 で記憶されている関係情報の個数は問わないが、後述するスコア取得部 15 によって、経由スコアを取得する場合には、複数の関係情報が記憶されていることが好適である。一方、共起スコアのみ取得しか行わない場合には、関係情報記憶部 11 において、一個の関係情報が記憶されていてもよく、あるいは、複数の関係情報が記憶されていてもよい。なお、言語表現は、通常、単語（例えば、「健康」等）であるが、連続した単語の並び（例えば、「早朝散歩」等の単語列）であってもよい。また、その単語は、例えば、自立語の単語であってもよく、特に名詞の単語であってもよい。また、単語が名詞である場合には、その名詞を助詞や前置詞でつないだものが言語表現であつてもよい（例えば、「私の友人」や「friend of mine」等）。また、関係情報は、通常、2 個の言語表現を有するものであるが、3 個以上の言語表現を有してもよい。なお、本実施の形態では、関係情報が主に 2 個の言語表現を有する場合について説明する。また、本実施の形態では、関係情報を < 第 1 の言語表現、第 2 の言語表現 > のように表記する。

【0027】

また、意味的關係は、何らかの意味的關係であれば、その種類を問わない。例えば、上位と下位の関係（例えば、< 飲み物、コーヒー > 等）であってもよく、原因結果の関係（例えば、< 脳梗塞、死亡 > 等）であってもよく、ライバルや対義語の関係（例えば、< 上、下 > や < 高速、低速 > 等）であってもよく、製品とメーカーの関係（例えば、< 掃除機、A 社 > 等）であってもよく、事象と方法の関係（例えば、< 爆発、爆弾 > 等）であってもよく、事象とツールの関係（例えば、< 授業、教科書 > 等）であってもよく、事象と防ぐものの関係（例えば、< 病気、薬 > ）であってもよく、物と材料の関係（例えば、< 缶、アルミニウム > 等）であってもよく、全体と部分の関係（例えば、< 自転車、サドル > 等）であってもよく、事象や物とトラブルの関係（例えば、< 自動車、パンク > 等）であってもよく、事象と対策との関係（例えば、< 雨、傘 > 等）であってもよく、事象と必須のものとの関係（例えば、< 記念撮影、写真機 > 等）であってもよく、近距離の地点の関係（例えば、< 大阪、神戸 > 等）であってもよく、地点と名物や名所の関係（例えば、< 東京、東京タワー > 等）であってもよく、料理と含まれる食材との関係（例えば、< シチュー、じゃがいも > 等）であってもよく、その他の種類の関係であってもよい。また、意味的關係は、ある言語表現と、他の言語表現とが、例えば、関連あり、ゆかりの人物、ゆかりの寺・神社、食材効能、効く食材、意外な食材、料理効能、効く料理、意外な料理、栄養効能、効く栄養素、成分、旬、旬の食材、代用食品、類似語、仏像・神様、所蔵する寺、祀る神社等の関係を有することであってもよい。なお、意味的關係は、3 個以上の言語表現に関する関係であってもよい。例えば、食生活と検査結果と病気の関係（例えば、< 高カロリー、高血糖、糖尿病 >、< 塩分過多、高血圧、脳卒中 > 等）であってもよい。

【0028】

また、関係情報には、その関係情報に応じた意味的關係を識別する情報が含まれてもよく、あるいは、含まれなくてもよい。前者の場合には、例えば、関係情報は、「原因 < 脳梗塞、死亡 >」であってもよい。この場合には、「原因」が意味的關係を識別する情報であり、死亡の原因が脳梗塞であることを示している。

【0029】

また、関係情報記憶部 11 で記憶される関係情報は、手作業で作成されたものであつてもよく、自動的に取得されたものであつてもよい。後者の場合には、例えば、前述の非特許文献 1 の手法を用いてもよい。

【0030】

なお、関係情報に含まれる言語表現は、言語表現そのものであつてもよく、あるいは、その情報を特定可能な情報であってもよい。後者の場合には、関係情報に含まれる言語表現は、例えば、言語表現が格納されている領域を示すポインタやアドレスであってもよい。このことは、他の情報についても同様であるとする。

10

20

30

40

50

【 0 0 3 1 】

類似情報記憶部 1 2 では、2 以上の類似情報が記憶される。類似情報は、類似する 2 以上の言語表現を有する情報である。類似情報は、2 個の言語表現を有するものであってもよく、3 個以上の言語表現を有するものであってもよい。なお、同一の類似情報に含まれる言語表現は、互いに類似するものであるとする。また、類似情報は、言語表現の類似の程度を示す類似度を含んでいてもよく、あるいは、含んでいなくてもよい。また、類似情報は、手作業で作成されたものであってもよく、自動的に取得されたものであってもよい。

【 0 0 3 2 】

拡張関係候補情報生成部 1 3 は、関係情報記憶部 1 1 で記憶されている関係情報に含まれる少なくとも 1 個の言語表現を、類似情報を用いて、その言語表現に類似する言語表現に置換した拡張関係候補情報を生成する。そして、拡張関係候補情報生成部 1 3 は、その生成した拡張関係候補情報を拡張関係候補情報記憶部 1 4 に蓄積する。なお、拡張関係候補情報生成部 1 3 は、関係情報に含まれる少なくとも 1 個の言語表現を類似するものに置換するものである。したがって、拡張関係候補情報生成部 1 3 は、関係情報に含まれる 1 個の言語表現を類似するものに置換してもよく、あるいは、関係情報に含まれる 2 個以上の言語表現をそれぞれ類似するものに置換してもよい。また、拡張関係候補情報生成部 1 3 は、関係情報に含まれるすべての言語表現を類似するものに置換してもよい。また、関係情報に含まれる言語表現に類似する言語表現が 2 以上存在する場合には、拡張関係候補情報生成部 1 3 は、関係情報に含まれる言語表現を、その 2 以上の各言語表現に置換した拡張関係候補情報をそれぞれ生成してもよい。

【 0 0 3 3 】

なお、拡張関係候補情報生成部 1 3 は、異なる関係情報から、同じ拡張関係候補情報を生成することがありうる。例えば、拡張関係候補情報生成部 1 3 が、異なる関係情報<脳梗塞、死亡>、<心筋梗塞、死亡>から、同じ拡張関係候補情報<脳卒中、死亡>を生成する場合などである。その場合には、例えば、生成後の拡張関係候補情報に対するユニーク処理を行い、拡張関係候補情報の重複を解消してもよく、あるいは、拡張関係候補情報の蓄積時に、すでに蓄積されている拡張関係候補情報を蓄積しないようにしてもよい。

【 0 0 3 4 】

また、拡張関係候補情報生成部 1 3 は、関係情報と同じ情報である拡張関係候補情報を拡張関係候補情報記憶部 1 4 に蓄積してもよく、あるいは、蓄積しなくてもよい。前者の場合には、拡張関係候補情報生成部 1 3 は、さらに積極的に、関係情報記憶部 1 1 で記憶されている各関係情報をそのまま拡張関係候補情報記憶部 1 4 に蓄積するようにしてもよい。

【 0 0 3 5 】

拡張関係候補情報記憶部 1 4 では、拡張関係候補情報が記憶される。拡張関係候補情報は、前述のように、少なくとも 1 個の言語表現が類似する言語表現に置換された関係情報である。拡張関係候補情報記憶部 1 4 で記憶されている拡張関係候補情報は、前述のように、拡張関係候補情報生成部 1 3 によって生成されたものである。

【 0 0 3 6 】

スコア取得部 1 5 は、拡張関係候補情報記憶部 1 4 で記憶されている拡張関係候補情報が意味的關係を有する確からしさを示すスコアを取得する。このスコアは、拡張関係候補情報に含まれる 2 以上の言語表現の共起に関するスコアであってもよい。すなわち、スコア取得部 1 5 は、拡張関係候補情報に含まれる 2 以上の言語表現の共起に関するスコアである共起スコアを取得してもよい。また、このスコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアであってもよい。すなわち、スコア取得部 1 5 は、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経路スコアを取得してもよい。その経路スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、その拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言

10

20

30

40

50

語表現とが類似しているほど、より高い値となるスコアであってもよい。なお、「置換前の言語表現」とは、置換の対象となる言語表現のことである。また、スコア取得部15は、共起スコアと、経路スコアとの一方のみを取得してもよく、あるいは、両方を取得してもよい。

【0037】

ここで、共起スコアの取得方法と、経路スコアの取得方法とについて説明する。

(1) 共起スコアの取得方法

まず、共起について説明する。言語表現Aと、言語表現Bとが共起するとは、決められた長さの範囲内(例えば、一文の範囲内、一段落の範囲内、一ページの範囲内、一の文書の範囲内、一のウェブページの範囲内等)において、言語表現Aと言語表現Bとが同時に出現することである。共起スコアは、その共起を示す尺度のことである。共起を示す尺度としては、例えば、共起頻度や共起率、Simpson係数、コサイン距離、ダイス係数、相互情報量等が存在する。言語表現Aと言語表現Bとの共起頻度とは、決められた長さの範囲内において、言語表現A, Bが同時に出現する数である。共起率とは、共起頻度を、言語表現Aの出現数(Xとする)と言語表現Bの出現数(Yとする)との和から共起頻度(Zとする)を引いたもの(すなわち、 $X + Y - Z$)で割った数である。また、Simpson係数は、共起率の分母を、X, Yの最小値にしたものである。また、コサイン距離は、共起率の分母を、XとYの積の絶対値の自乗根にしたものである。なお、これらの共起を示す尺度についてはすでに公知である。例えば、次の文献を参照されたい。また、上記以外の共起の尺度を用いて共起スコアを取得してもよいことは言うまでもない。

【0038】

文献：相澤彰子、「共起に基づく類似性尺度」、オペレーションズ・リサーチ、経営の科学52(11)、p.706-712、2007年11月

【0039】

なお、その共起スコアを算出する際には、多くの文書を有するデータベースが必要になる。そのデータベースは、例えば、関係情報拡張装置1が保持していてもよく、あるいは、装置外に存在してもよい。後者の場合には、例えば、そのデータベースは、ウェブサイトであってもよく、あるいは、所定のサーバが有するデータベースであってもよい。また、その共起スコアを算出する際の検索等の処理は、スコア取得部15がそのデータベースにアクセスすることによって行ってもよく、あるいは、スコア取得部15は、外部の装置やサーバに対して、2以上の言語表現を渡し、その外部の装置やサーバにおいて生成された共起スコアを受け取るだけであってもよい。

【0040】

また、あらゆる言語表現の組合せについてあらかじめ共起スコアを算出して保持しておくことによって、ある言語表現A, Bの共起スコアを、その保持している情報から取得するようにしてもよい。具体的には、言語表現Aと言語表現Bと、両言語表現の共起スコアとが対応付けられて保持されており、スコア取得部15は、言語表現A, Bの共起スコアの取得方法を取得する際には、その言語表現A, Bに対応付けられている共起スコアを読み出してもよい。このように、共起スコアの取得は、共起スコアを算出することであってもよく、共起スコアを読み出すことであってもよい。

【0041】

また、拡張関係候補情報が3個以上の言語表現を有する場合には、スコア取得部15は、その3個以上の言語表現の共起スコアを取得する。すなわち、スコア取得部15は、拡張関係候補情報に含まれるすべての言語表現に対する共起スコアを取得する。その共起スコアが、共起頻度や共起率等のように、3個以上の言語表現についても取得できるものであれば、スコア取得部15は、その3個以上の言語表現に対応する共起スコアを取得する。一方、共起スコアがダイス係数や相互情報量等のように、2個の言語表現に対してのみ定義されている場合には、スコア取得部15は、3個以上の言語表現から2個の言語表現のすべてのペアを作り、そのすべてのペアについてダイス係数等の共起の尺度を取得し、そのすべてのペアの共起の尺度を引数とする関数の値を共起スコアとしてもよい。なお

、その関数は、各引数の増加関数であるとする。例えば、その関数は、すべてのペアの共起の尺度の平均や和、積等であってもよい。

【 0 0 4 2 】

(2) 経路スコアの取得方法

次に、経路スコアについて説明する。ここでは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、その拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアである経路スコアについて説明する。また、関係情報と、拡張関係候補情報において、2個の言語表現が含まれている場合について説明する。その関係情報と、拡張関係候補情報とにおいて、第1項の言語表現のことを第1の言語表現と呼び、第2項の言語表現のことを第2の言語表現と呼ぶことにする。すなわち、関係情報や、拡張関係候補情報が $\langle X, Y \rangle$ である場合には、第1の言語表現が X となり、第2の言語表現が Y となる。

10

【 0 0 4 3 】

ある拡張関係候補情報を $\langle f_h, s_h \rangle$ とする。ここで、 f_h, s_h は、言語表現である。意味的關係を有する2個の言語表現を有する与えられた関係情報の集合を、 R_{given} とする。 $R_{given} = \{ r_1 = \langle f_1, s_1 \rangle, \dots, r_n = \langle f_n, s_n \rangle \}$ とする。そして、第1の計算値 $S_{FA}(f_h, s_h)$ と、第2の計算値 $S_{SA}(f_h, s_h)$ と、第3の計算値 $S_{FULL}(f_h, s_h)$ とを、次式のようにして算出する。

【 0 0 4 4 】

なお、第1の計算値 $S_{FA}(f_h, s_h)$ は、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第2の言語表現 s_h が一致する各関係情報の第1の言語表現と、拡張関係候補情報 $\langle f_h, s_h \rangle$ の第1の言語表現 f_h との類似度の和である。その和は、次式で示されるように、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第2の言語表現 s_h が一致する各関係情報の第1の言語表現の集合に関する和である。したがって、その第1の計算値 $S_{FA}(f_h, s_h)$ は、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第2の言語表現 s_h が一致する関係情報に含まれる第1の言語表現の集合を特定し、その集合に含まれる各第1の言語表現と、拡張関係候補情報 $\langle f_h, s_h \rangle$ の第1の言語表現 f_h との類似度の、集合の各要素に関する和であるということもできる。

20

【 0 0 4 5 】

また、第2の計算値 $S_{SA}(f_h, s_h)$ は、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第1の言語表現 f_h が一致する各関係情報の第2の言語表現と、拡張関係候補情報 $\langle f_h, s_h \rangle$ の第2の言語表現 s_h との類似度の和である。その和は、次式で示されるように、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第1の言語表現 f_h が一致する各関係情報の第2の言語表現の集合に関する和である。したがって、その第2の計算値 $S_{SA}(f_h, s_h)$ は、拡張関係候補情報 $\langle f_h, s_h \rangle$ と第1の言語表現 f_h が一致する関係情報に含まれる第2の言語表現の集合を特定し、その集合に含まれる各第2の言語表現と、拡張関係候補情報 $\langle f_h, s_h \rangle$ の第2の言語表現 s_h との類似度の、集合の各要素に関する和であるということもできる。

30

【 0 0 4 6 】

また、第3の計算値 $S_{FULL}(f_h, s_h)$ は、関係情報記憶部 11 で記憶されている各関係情報と拡張関係候補情報 $\langle f_h, s_h \rangle$ との第1の言語表現同士の類似度及び第2の言語表現同士の類似度の積の和である。その和は、次式で示されるように、 R_{given} における第1の言語表現と第2の言語表現とのペアの集合に関する和である。したがって、その第3の計算値 $S_{FULL}(f_h, s_h)$ は、 R_{given} に含まれる各関係情報の言語表現のペアの集合に含まれる一の関係情報について、第1の言語表現と拡張関係候補情報 $\langle f_h, s_h \rangle$ の第1の言語表現 f_h との類似度と、第2の言語表現と拡張関係候補情報 $\langle f_h, s_h \rangle$ の第2の言語表現 s_h との類似度との積を算出し、その類似度の積の、集合の各要素に関する和であるということもできる。

40

【数 1】

$$S_{FA}(f_h, s_h) = \sum_{f_i \in FA(s_h)} \text{sim}(f_h, f_i)$$

$$S_{SA}(f_h, s_h) = \sum_{s_i \in SA(f_h)} \text{sim}(s_h, s_i)$$

$$S_{FULL}(f_h, s_h) = \sum_{\langle f_i, s_i \rangle \in R_{given}} \text{sim}(f_h, f_i) \text{sim}(s_h, s_i)$$

【0047】

ここで、 $FA(s)$ は、 R_{given} で第 2 の言語表現が s である関係情報の第 1 の言語表現の集合である。また、 $SA(f)$ は、 R_{given} で第 1 の言語表現が f である関係情報の第 2 の言語表現の集合である。 sim は、類似度である。この類似度は、例えば、次の文献の類似度のように自動的に算出されるものを用いてもよく、手作業で設定された類似度を用いてもよく、類義語、同義語など、意味が類似している表現として登録されている 2 つの言語表現を高い類似度としてもよい。また、言語表現を、意味的な階層構造、木構造で整理している辞書を用いる場合には、2 つの言語表現から上の構造を辿って行き、最初の共通の場所までに辿る階層の数が少ないほど類似度が高く、逆に大きいほど類似度が低くなるように設定してもよい。そのような辞書としては、日本語のものとしては、例えば、「分類語彙表」(国立国語研究所)、「日本語語彙大系」(岩波書店)、「角川類語国語辞典」(角川書店)、「日本語大シソーラス」(大修館書店)、「EDR 概念体系辞書」(EDR プロジェクト)、「デジタル類語辞典」(ジャングル)、「JST 科学技術用語シソーラス」(JST 科学技術振興機構)等が存在する。また、英語のものとしては、例えば、「ロジェ類語辞典」、「WordNet」、「MeSH (Medical Subject Headings)」等が存在する。

【0048】

文献：風間淳一、Stijn De Saeger、鳥澤健太郎、村田真樹、「係り受けの確率的クラスタリングを用いた大規模類似語リストの作成」、言語処理学会第 15 回年次大会、p. 84 - 87、2009 年 3 月

【0049】

なお、上記各計算値を取得する際に、類似情報記憶部 12 で記憶されている、類似度を含む類似情報を用いる場合には、言語表現 A 、 B が類似するものではなく、類似情報記憶部 12 で記憶されている類似情報に含まれないため、 $\text{sim}(A, B)$ が分からない場合もありうる。その場合には、 $\text{sim}(A, B) = 0$ にするものとする。

【0050】

スコア取得部 15 は、拡張関係候補情報 $\langle f_h, s_h \rangle$ の経由スコアを、前述の第 1 の計算値 $S_{FA}(f_h, s_h)$ と、第 2 の計算値 $S_{SA}(f_h, s_h)$ と、第 3 の計算値 $S_{FULL}(f_h, s_h)$ とのうち、任意の 1 以上の計算値を引数とする増加関数の値を算出することによって取得してもよい。なお、その増加関数は、各引数の増加関数であり、例えば、第 1 から第 3 の計算値の和である $S^{sum}(f_h, s_h)$ であってもよく、第 1 から第 3 の計算値の積である $S^{prod}(f_h, s_h)$ であってもよい。 S^{sum} は、第 1 から第 3 の計算値のいずれかが高い場合に高くなる。一方、 S^{prod} は、第 1 から第 3 の計算値のすべてが高い場合に高くなる。つまり、 S^{prod} は、バランスよく両方の言語表現に基づき生成される拡張関係候補情報がよい値となる、と考えられる点で、 S^{sum} と異なる。なお、 S^{prod} の計算では、0 になることを回避するため、各計算値に十分に小さい値を足すものとする。

【0051】

なお、上記の式において、 $\text{sim}(A, B)$ を、言語表現 A 、 B が類似する場合に「1」となり、言語表現 A 、 B が類似しない場合に「0」となる関数であるとすると、経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとなる。すなわち、類似の程度を考慮しないスコアとな

10

20

30

40

50

る。したがって、そのようにして前述の計算値を算出するようにしてもよい。例えば、スコア取得部15は、 $sim(A, B)$ を取得する際に、類似情報記憶部12で記憶されている類似情報によって、言語表現A, Bが類似することが示される場合には、 $sim(A, B) = 1$ として、言語表現A, Bが類似しないことが示される場合には、 $sim(A, B) = 0$ としてもよい。

【0052】

また、関係情報や拡張関係候補情報がm個の言語表現を含む場合の経路スコアの算出について簡単に説明する。なお、mは2以上の整数である。ここで、ある拡張関係候補情報を $\langle f_h^1, f_h^2, \dots, f_h^m \rangle$ とする。 f_h^1, f_h^2 等は、拡張関係候補情報に含まれる言語表現である。また、関係情報の集合 R_{given} は次のようであるとする。なお、 f_h^j や、 f_i^j を第jの言語表現や、j番目の言語表現と呼ぶことにする。

$R_{given} = \{ r_1 = \langle f_1^1, f_1^2, \dots, f_1^m \rangle, r_2 = \langle f_2^1, f_2^2, \dots, f_2^m \rangle, \dots, r_n = \langle f_n^1, f_n^2, \dots, f_n^m \rangle \}$

【0053】

その場合に、計算値 $S_{k_1 k_2 \dots k_j}(f_h^1, \dots, f_h^m)$ は、次のようになる。

【数2】

$$S_{k_1 k_2 \dots k_j}(f_h^1, \dots, f_h^m) = \sum_{(f_i^{k_1}, f_i^{k_2}, \dots, f_i^{k_j}) \in M} sim(f_h^{k_1}, f_i^{k_1}) sim(f_h^{k_2}, f_i^{k_2}) \dots sim(f_h^{k_j}, f_i^{k_j})$$

【0054】

ここで、計算値 $S_{k_1 k_2 \dots k_j}(f_h^1, \dots, f_h^m)$ は、拡張関係候補情報が、k1番目の言語表現と、k2番目の言語表現と、...、kj番目の言語表現とを置換して生成されたものである場合における経路スコアの計算値である。なお、上記式における集合Mについて説明する。まず、集合 $M_1 = \{1, 2, 3, \dots, m\}$ として、集合 $M_2 = \{k_1, k_2, k_3, \dots, k_j\}$ とする。その集合M2の各要素k1, k2, k3, ..., kjは、集合M1から重複しないように選択した1個以上、m個以下の要素である。また、 $k_1 < k_2 < \dots < k_j$ であるとする。また、集合M3を、集合M1から、集合M2に含まれる各要素を除去したものであるとする。その集合M3を、 $M_3 = \{p_1, p_2, \dots, p_{m-j}\}$ とする。すると、Mは、 R_{given} において、第p1項が $f_h^{p_1}$ であり、第p2項が $f_h^{p_2}$ であり、...、第pm-j項が $f_h^{p_{m-j}}$ である、第k1項、第k2項、...、第kj項の組の集合である。ただし、j = mの場合には、M3は空集合となり、Mは R_{given} となる。そして、スコア取得部15は、j = 1からj = mまでの各値におけるすべてのM2のうち、任意の1以上のM2に対する計算値を引数とする増加関数（この増加関数は、各引数の増加関数であるとする）の値を算出することによって、m個の言語表現を含む拡張関係候補情報に対する経路スコアを取得することができる。その増加関数は、例えば、すべてのM2について算出した計算値の和であってもよく、すべてのM2について算出した計算値の積であってもよい。

【0055】

また、ここでは、共起スコアと、経路スコアとの2個のスコアについて説明したが、スコア取得部15は、それら以外の拡張関係候補情報が意味的關係を有する確からしさを示すスコアを取得してもよい。

【0056】

また、スコア取得部15は、共起スコアと、経路スコアとの両方を取得する場合に、拡張関係候補情報記憶部14で記憶されているすべての拡張関係候補情報について、共起スコアと、経路スコアとを取得してもよく、あるいは、一方のスコアを取得し、その取得したスコアの高いものについてのみ、他方のスコアを取得するようにしてもよい。本実施の形態では、後者の場合について主に説明する。

【0057】

10

20

30

40

50

また、スコア取得部 15 が取得したスコアは、拡張関係候補情報に対応付けられて蓄積されてもよい。具体的には、スコア取得部 15 は、取得したスコアを、拡張関係候補情報記憶部 14 で記憶されている、そのスコアに対応する拡張関係候補情報に対応付けて蓄積してもよく、あるいは、取得したスコアを、そのスコアに対応する拡張関係候補情報を識別する情報に対応付けて図示しない記録媒体に蓄積してもよい。

【0058】

選択部 16 は、スコア取得部 15 が取得したスコアを用いて、拡張関係候補情報記憶部 14 で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報を選択する。この選択部 16 が選択した拡張関係候補情報が拡張関係情報となる。スコア取得部 15 が共起スコアと、経路スコアとの両方を取得した場合には、選択部 16 は、共起スコアが高く、かつ、経路スコアが高い拡張関係候補情報を選択してもよい。なお、スコアが高い拡張関係候補情報とは、例えば、しきい値以上のスコアである拡張関係候補情報であってもよく、あるいは、スコアの高いものからあらかじめ決められた個数の拡張関係候補情報であってもよい。なお、その 2 個の条件を合わせて用いてもよい。例えば、スコアの高い拡張関係候補情報は、しきい値以上のスコアである拡張関係候補情報であって、かつ、スコアの高いものからあらかじめ決められた個数内の拡張関係候補情報であってもよい。そのしきい値は、例えば、スコアの最大値にあらかじめ決められた 1 未満の数（例えば、0.9 や 0.8 など）を掛けた値であってもよく、あらかじめ決められた値であってもよい。

【0059】

選択部 16 は、選択した拡張関係候補情報を後述する拡張関係情報記憶部 17 に蓄積してもよく、あるいは、拡張関係候補情報記憶部 14 において、選択した拡張関係候補情報に対して、拡張関係情報であることを示すフラグ等の識別情報を設定してもよい。すなわち、拡張関係候補情報を、選択されたものと、選択されていないものとに区別できるのであれば、その選択結果を示す方法は問わない。なお、本実施の形態では、前者の場合、すなわち、選択結果である拡張関係情報が拡張関係情報記憶部 17 に蓄積される場合について説明する。

【0060】

また、拡張関係候補情報記憶部 14 に関係情報と同じ情報である拡張関係候補情報が記憶されている場合に、選択部 16 は、関係情報と同じ拡張関係情報を選択してもよく、あるいは、しなくてもよい。後者の場合には、選択部 16 は、関係情報記憶部 11 で記憶されている関係情報を参照し、その関係情報と一致する拡張関係候補情報を選択しないようにしてもよい。

【0061】

拡張関係情報記憶部 17 では、選択部 16 によって選択された拡張関係候補情報である拡張関係情報が記憶される。この拡張関係情報は、前述のように、選択部 16 によって蓄積されたものである。

【0062】

なお、関係情報記憶部 11 や、類似情報記憶部 12 に関係情報や 2 以上の類似情報が記憶される過程は問わない。例えば、記録媒体を介して関係情報等が関係情報記憶部 11 等で記憶されるようになってよく、通信回線等を介して送信された関係情報等が関係情報記憶部 11 等で記憶されるようになってよく、あるいは、入力デバイスを介して入力された関係情報等が関係情報記憶部 11 等で記憶されるようになってよく、また、関係情報記憶部 11 や類似情報記憶部 12、拡張関係候補情報記憶部 14、拡張関係情報記憶部 17 での記憶は、RAM 等における一時的な記憶でもよく、あるいは、長期的な記憶でもよい。また、また、関係情報記憶部 11 や類似情報記憶部 12、拡張関係候補情報記憶部 14、拡張関係情報記憶部 17 は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。

【0063】

また、関係情報記憶部 11 と、類似情報記憶部 12 と、拡張関係候補情報記憶部 14 と

、拡張関係情報記憶部 17 とのうち、任意の 2 以上の記憶部は、同一の記録媒体によって実現されてもよく、あるいは、別々の記録媒体によって実現されてもよい。前者の場合には、例えば、関係情報を記憶している領域が関係情報記憶部 11 となり、類似情報を記憶している領域が類似情報記憶部 12 となる。

【0064】

次に、本実施の形態による関係情報拡張装置 1 の動作について、図 2 のフローチャートを用いて説明する。

(ステップ S101) 拡張関係候補情報生成部 13 は、関係情報記憶部 11 で記憶されている関係情報の少なくとも 1 個の言語表現を、類似情報記憶部 12 で記憶されている類似情報を用いて類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報記憶部 14 に蓄積する。なお、この処理の詳細については、図 3 のフローチャートを用いて説明する。

10

【0065】

(ステップ S102) スコア取得部 15 は、拡張関係候補情報記憶部 14 で記憶されている各拡張関係候補情報について、スコアを取得する。なお、この処理の詳細については、図 4 のフローチャートを用いて後述する。

【0066】

(ステップ S103) 選択部 16 は、スコア取得部 15 が取得したスコアの高い拡張関係候補情報である拡張関係情報を選択する。この選択は、例えば、前述のように、しきい値よりも大きいスコアの拡張関係候補情報を選択することによって行われてもよく、スコアの高いものからあらかじめ決められた個数の拡張関係候補情報を選択することによって行われてもよい。そして、選択部 16 は、その選択結果の拡張関係情報を、拡張関係情報記憶部 17 に蓄積する。このようにして、関係情報を拡張する一連の処理が終了する。

20

【0067】

図 3 は、図 2 のフローチャートにおける拡張関係候補情報の生成の処理 (ステップ S101) の詳細を示すフローチャートである。

(ステップ S201) 拡張関係候補情報生成部 13 は、カウンタ i を 1 に設定する。

【0068】

(ステップ S202) 拡張関係候補情報生成部 13 は、カウンタ j を 1 に設定する。

【0069】

30

(ステップ S203) 拡張関係候補情報生成部 13 は、関係情報記憶部 11 で記憶されている i 番目の関係情報に含まれる j 番目の言語表現に類似する言語表現が、類似情報記憶部 12 で記憶されているかどうか判断する。そして、 i 番目の関係情報に含まれる j 番目の言語表現に類似する言語表現が類似情報記憶部 12 で記憶されている場合には、ステップ S204 に進み、そうでない場合には、ステップ S206 に進む。なお、この判断は、例えば、その j 番目の言語表現を検索キーとして、類似情報記憶部 12 を検索することによって行ってもよい。そして、その検索でヒットした場合には、 j 番目の言語表現に類似する言語表現が記憶されていることになり、ヒットしなかった場合には、 j 番目の言語表現に類似する言語表現が記憶されていないことになる。

【0070】

40

(ステップ S204) 拡張関係候補情報生成部 13 は、 i 番目の関係情報に含まれる j 番目の言語表現に類似する言語表現を類似情報記憶部 12 から読み出す。具体的には、ステップ S203 の検索でヒットした類似情報から、 i 番目の関係情報に含まれる j 番目の言語表現以外の言語表現を読み出すことによって行われてもよい。その読み出した言語表現は、図示しない記録媒体で一時的に記憶されてもよい。そして、拡張関係候補情報生成部 13 は、 i 番目の関係情報に含まれる j 番目の言語表現を、類似情報記憶部 12 から読み出した言語表現に置換した拡張関係候補情報を生成する。例えば、類似情報記憶部 12 から 2 以上の言語表現を読み出した場合には、拡張関係候補情報生成部 13 は、 i 番目の関係情報に含まれる j 番目の言語表現を、その 2 以上の言語表現のそれぞれに置換した 2 以上の拡張関係候補情報を生成することになる。

50

【 0 0 7 1 】

なお、このステップ S 2 0 4 において、拡張関係候補情報生成部 1 3 は、i 番目の関係情報に関して、(j - 1) 番目までの少なくともいずれかの言語表現を置換した拡張関係候補情報についても、j 番目の言語表現の置換を行ってもよい。例えば、関係情報が 3 個の言語表現を含んでおり、j = 3 である場合に、i 番目の関係情報に関して、1 番目の言語表現のみが置換された拡張関係候補情報と、2 番目の言語表現のみが置換された拡張関係候補情報と、1 番目の 2 番目の言語表現の両方が置換された拡張関係候補情報とが存在する場合には、拡張関係候補情報生成部 1 3 は、その 3 個の拡張関係候補情報について、j 番目 (= 3 番目) の言語表現を類似する言語表現に置換する処理を行ってもよい。

【 0 0 7 2 】

(ステップ S 2 0 5) 拡張関係候補情報生成部 1 3 は、置換後の拡張関係候補情報を拡張関係候補情報記憶部 1 4 に蓄積する。なお、ステップ S 2 0 4 において、i 番目の関係情報に関して、それまでに置換の行われた拡張関係候補情報についても置換を行う場合には、このステップ S 2 0 5 の蓄積時に、蓄積対象の拡張関係候補情報が、i 番目の関係情報に関するものであることが分かるように蓄積することが好適である。拡張関係候補情報生成部 1 3 は、例えば、カウンタ i の値に対応付けて拡張関係候補情報を蓄積してもよい。

【 0 0 7 3 】

また、拡張関係候補情報生成部 1 3 は、ステップ S 2 0 4 と、ステップ S 2 0 5 との処理を、1 個の拡張関係候補情報を生成することに繰り返して実行してもよい。例えば、i 番目の関係情報に含まれる j 番目の言語表現に類似する言語表現が 2 以上あった場合には、拡張関係候補情報生成部 1 3 は、j 番目の言語表現を各言語表現に置換することに、置換後の拡張関係候補情報を拡張関係候補情報記憶部 1 4 に蓄積してもよい。

【 0 0 7 4 】

(ステップ S 2 0 6) 拡張関係候補情報生成部 1 3 は、カウンタ j を 1 だけインクリメントする。

【 0 0 7 5 】

(ステップ S 2 0 7) 拡張関係候補情報生成部 1 3 は、i 番目の関係情報に j 番目の言語表現が存在するかどうか判断する。そして、存在する場合には、ステップ S 2 0 3 に戻り、そうでない場合には、ステップ S 2 0 8 に進む。なお、関係情報に含まれる言語表現の個数はあらかじめ決まっているため、例えば、その個数を図示しない記録媒体で記憶しておき、拡張関係候補情報生成部 1 3 は、その記録媒体で記憶されている言語表現の個数と、カウンタ j の値とを比較することによって、このステップ S 2 0 7 の処理を行ってもよい。その場合には、j > (記憶されている言語表現の個数) である場合には、ステップ S 2 0 8 に進むことになる。

【 0 0 7 6 】

(ステップ S 2 0 8) 拡張関係候補情報生成部 1 3 は、カウンタ i を 1 だけインクリメントする。

【 0 0 7 7 】

(ステップ S 2 0 9) 拡張関係候補情報生成部 1 3 は、関係情報記憶部 1 1 に i 番目の関係情報が存在するかどうか判断する。そして、存在する場合には、ステップ S 2 0 2 に戻り、そうでない場合には、図 2 のフローチャートに戻る。

【 0 0 7 8 】

図 4 は、図 2 のフローチャートにおけるスコアの取得の処理 (ステップ S 1 0 2) の詳細を示すフローチャートである。なお、図 4 のフローチャートでは、共起スコアを取得し、その後共起スコアの高いものを暫定的に選択して、その選択された拡張関係候補情報についてのみ、経路スコアを取得する場合について説明する。

【 0 0 7 9 】

(ステップ S 3 0 1) スコア取得部 1 5 は、拡張関係候補情報記憶部 1 4 で記憶されているすべての拡張関係候補情報について、共起スコアを取得する。

10

20

30

40

50

【 0 0 8 0 】

(ステップS302) 選択部16は、ステップS301で取得された共起スコアの高い拡張関係候補情報を暫定的に選択する。選択部16は、例えば、その選択後の拡張関係候補情報を拡張関係候補情報記憶部14や、図示しない記録媒体に蓄積してもよく、あるいは、暫定的に選択した、拡張関係候補情報記憶部14で記憶されている拡張関係候補情報に対して、暫定的に選択されたことを示すフラグ等の識別情報を設定してもよい。

【 0 0 8 1 】

(ステップS303) スコア取得部15は、ステップS302で暫定的に選択されたすべての拡張関係候補情報について、経由スコアを取得する。

【 0 0 8 2 】

このように、共起スコアを取得し、共起スコアの高いものを暫定的に選択し、その暫定的に選択された拡張関係候補情報について経由スコアを取得することによって、経由スコアを取得する拡張関係候補情報の個数を減らすことができる。通常、共起スコアの取得よりも、経由スコアの取得の方が負荷の高い処理であるため、この順序でスコアの取得を行うことによって、スコアの取得の処理を軽減することができる。この場合には、ステップS103の選択の処理において、経由スコアの高い拡張関係候補情報を選択すればよいことになる。

【 0 0 8 3 】

なお、図4のフローチャートにおけるスコアの取得の方法は一例であり、拡張関係候補情報記憶部14で記憶されているすべての拡張関係候補情報について、共起スコアと経由スコアとの両方を取得してもよい。この場合には、ステップS103の選択の処理において、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択すればよいことになる。また、共起スコアと、経由スコアとの一方のみを用いた選択を行う場合には、図4のフローチャートにおいて、拡張関係候補情報記憶部14で記憶されているすべての拡張関係候補情報について、その選択で用いるスコアのみを取得を行ってもよい。この場合には、ステップS103の選択の処理において、選択で用いるスコア(共起スコア、または、経由スコア)の高い拡張関係候補情報を選択すればよいことになる。また、図4のフローチャートとは逆に、まず経由スコアを取得し、その後には経由スコアの高いものを暫定的に選択して、その選択された拡張関係候補情報についてのみ、共起スコアを取得してもよい。

【 0 0 8 4 】

次に、本実施の形態による関係情報拡張装置1の動作について、簡単な具体例を用いて説明する。この具体例において、共起スコアのみを用いて選択を行うものとする。また、この具体例では、関係情報記憶部11において、死亡の原因が心筋梗塞であることを示す関係情報<心筋梗塞、死亡>のみが記憶されているものとする。また、類似情報記憶部12では、図5で示される類似情報が記憶されているものとする。図5において、一つのレコードが、一つの類似情報である。また、一つの類似情報に含まれる各言語表現は、互いに類似するものである。例えば、心筋梗塞、脳梗塞、脳卒中、うつ病は、互いに類似する言語表現である。

【 0 0 8 5 】

関係情報を拡張する処理が開始されると、まず、拡張関係候補情報生成部13が、関係情報記憶部11で記憶されている関係情報<心筋梗塞、死亡>から拡張関係候補情報を生成する処理を行う(ステップS101)。具体的には、拡張関係候補情報生成部13は、その関係情報の1番目の言語表現「心筋梗塞」を検索キーとして類似情報記憶部12を検索する。その結果、1番目のレコードに含まれる「心筋梗塞」がヒットするため、拡張関係候補情報生成部13は、その1番目のレコードから、検索キー以外の言語表現「脳梗塞」「脳卒中」「うつ病」を読み出して図示しない記録媒体に蓄積すると共に、1番目の関係情報の1番目の言語表現「心筋梗塞」に類似する言語表現が存在すると判断する(ステップS201~S203)。そして、拡張関係候補情報生成部13は、関係情報<心筋梗塞、死亡>の1番目の言語表現「心筋梗塞」を、それに類似する言語表現「脳梗塞」「脳

10

20

30

40

50

卒中」「うつ病」に置換した拡張関係候補情報をそれぞれ生成し、それらの拡張関係候補情報を、その時点のカウンタ*i*の値に対応付けて拡張関係候補情報記憶部14に蓄積する(ステップS204, S205)。図6の1番目から3番目のレコードは、そのようにして蓄積された拡張関係候補情報を含んでいる。なお、図6において、関係情報IDは、カウンタ*i*の値である。共起スコアは、後にスコア取得部15によって取得されるものであるため、現段階では空欄である。その後、拡張関係候補情報生成部13は、関係情報<心筋梗塞、死亡>の2番目の言語表現「死亡」に類似する言語表現「病死」「急死」を類似情報記憶部12から読み出して蓄積すると共に、1番目の関係情報の2番目の言語表現「死亡」に類似する言語表現が存在すると判断する(ステップS206, S207, S203)。そして、拡張関係候補情報生成部13は、関係情報<心筋梗塞、死亡>と、それまでに蓄積された関係情報ID「1」に対応する拡張関係候補情報<脳梗塞、死亡>、<脳卒中、死亡>、<うつ病、死亡>との2番目の言語表現「死亡」を、それに類似する言語表現「病死」「急死」に置換した拡張関係候補情報をそれぞれ生成し、それらの拡張関係候補情報を、その時点のカウンタ*i*の値に対応付けて拡張関係候補情報記憶部14に蓄積する(ステップS204, S205)。その結果、拡張関係候補情報記憶部14で記憶されている情報は、図6で示されるようになる。

【0086】

次に、スコア取得部15は、図6で示される各拡張関係候補情報に含まれる2個の言語表現の共起スコアをそれぞれ取得し、その拡張関係候補情報に対応付けて拡張関係候補情報記憶部14に蓄積する(ステップS102, S301)。その結果、拡張関係候補情報記憶部14で記憶されている情報は、図7で示されるようになったとする。なお、この共起スコアは共起頻度であるとする。その後、選択部16は、共起スコアを用いた選択を行う。この選択では、しきい値が50に設定されており、そのしきい値以上の共起スコアの拡張関係候補情報が、拡張関係情報として選択されるものとする。すると、選択部16は、図7でしめされる拡張関係情報のうち、<うつ病、病死>と<うつ病、急死>以外の拡張関係情報を選択して拡張関係情報記憶部17に蓄積する(ステップS103)。その結果、拡張関係情報記憶部17では、図7の1番目から9番目までの拡張関係候補情報である拡張関係情報が記憶されることになる。このようにして、関係情報<心筋梗塞、死亡>を、拡張関係情報<脳梗塞、死亡>等に拡張することができる。なお、この具体例で示した共起スコア等は、本実施の形態による関係情報拡張装置1の動作の詳細を説明するために示したものであり、実際の文書等を用いて取得したデータではない。

【0087】

なお、この具体例では、互いに類似する2以上の言語表現が一の類似情報に含まれる場合について説明したが、そうでなくてもよい。類似情報は、例えば、類似する2個の言語表現を有する情報であってもよい。その場合には、例えば、図5の1番目のレコードは、心筋梗塞と脳梗塞のペア、心筋梗塞と脳卒中のペア、心筋梗塞とうつ病のペア、脳梗塞と脳卒中のペア、脳梗塞とうつ病のペア、脳卒中とうつ病のペアというように、6個の類似情報に分かれることになる。

【0088】

次に、本実施の形態による関係情報拡張装置1の実験例について説明する。この実験例では、「XはYの原因となる」の関係性を有する関係情報<X, Y>から得られた拡張関係情報の精度の評価と、従来のパターンベースの方法で取得することが困難であった関係を取得できているかどうかの評価とを行う。なお、この実験例において、言語表現は、名詞または連続する名詞である。

【0089】

まず、評価方法について説明する。評価は、3人の評価者によって行った。そして、(1)常識的に正解と判断された場合、あるいは、(2)常識的に正解と判断されなくても、ウェブに正しいと支持するエビデンスが1つ以上見つかった場合を正解とした。なお、正解であるとは、「XはYの原因となる」という関係が成り立つことである。また、3人の評価者のうち、2名以上一致(lenient)、3名一致(strict)で精度を

10

20

30

40

50

測定した。

【0090】

なお、(2)では、1つの関係に関して、YahooAPIを用いて「X、Y、原因」のAND検索で10ページを獲得し、各ページから「X、Y、原因」が200文字以内に存在するテキストセグメントを最大3つ抽出して、最大30個(=10×3)のセグメントを評価者に提示することによって行った。その最大30個のセグメントのうち、少なくとも1つが評価者によって妥当であると判断された場合には、正解であることになる。

【0091】

本実験では、各評価者で合計400個の評価を行ったが、評価者間のkappa値は、平均で、0.629であった。一般的に、kappa値が0.6以上ならば「かなりよい一致率」と言われていることから、評価者間の判定の一致率は概ねよいといえる。

10

【0092】

次に、この実験例で用いた類似情報について説明する。類似語の獲得には、前述の風間らの文献の方法で作成された約50万名詞に対する類似度付き類似語リスト(ALAGINフォーラムで公開されている文脈類似語データベースVersion1のold.500k-2k.data.)を用いた。

【0093】

風間らの方法では、大量コーパスから各名詞nの(助詞、動詞)、(の、名詞)の大きく2種類の係り受け関係depを収集し、Torisawaの文献の手法(次の文献参照)、

20

【数3】

$$P(n, dep) = \sum_{c_i \in C} P(c_i)P(n | c_i)P(dep | c_i)$$

に基づき、EMアルゴリズムでP(c)、P(n|c)、P(dep|c)を推定する(確率モデルとしてはPLSIと等価である)。これによって、depをそのまま素性とする場合と比べてスムージング効果が期待できる。次に、上記パラメータからP(c|n)を計算し、名詞n1、n2の類似度をP(c|n1)、P(c|n2)のJensen-Shannon(JS)ダイバージェンスとして求める。JSダイバージェンスは確率分布間の距離の一種で、以下の式で計算する。

30

【数4】

$$JS(P1 || P2) = \frac{1}{2}(KL(P1 || P_{mean}) + KL(P2 || P_{mean}))$$

【0094】

ここで、P1、P2は確率分布、KL(P1||P2)はKLダイバージェンス、P_{mean}はP1、P2をベクトルとしてみた場合の平均である。JSダイバージェンスは0から1を取り、小さいほど類似していることになる。そのため、単語n1、n2の類似度は次のようにする。

【数5】

40

$$sim(n1, n2) = 1 - JS(P(c | n1) || P(c | n2))$$

【0095】

最終的に、可能な単語集合の中の全ペアについて、(A)sim(n1, n2)がしきい値T_{sim}以上である、(B)互いの類似度のトップM単語に含まれる、の2つの条件を満たす単語ペアを類似情報として獲得した。この実験例では、しきい値T_{sim}=0.7、M=20として類似情報を生成した。なお、実験例で用いた各類似情報には、類似している単語のペアと、その単語のペアの類似度とが含まれている。

【0096】

文献：K.Torisawa, 「An Unsupervised Method

50

for Canonicalization of Japanese Postpositions」, In Proc. of the 6th NLP RS, p. 211 - 218, 2001年

【0097】

また、この実験例において、単語共起頻度である共起スコアを用いた。その単語共起頻度は、約1億文書で上記と同じ約50万名詞の全ペアに対して、近接4文内で共起する文書頻度を計算したデータ (ALAGINフォーラムで公開されている単語共起頻度データベース Version 1 の 500k - 500k . 100m - docs . w4 . data) を用いた。その共起頻度である共起スコアのしきい値は、 $T_{c.o.c} = 20$ とした。これらのしきい値等のパラメータは、共起スコアを用いた選択を行った場合に、経験的に、関係情報の約10倍の関係が生成されることを目安に設定した。

10

【0098】

次に、この実験例で用いた関係情報について説明する。関係情報は、De Saegerらの文献の方法 (次の文献参照) で獲得した関係から、明らかに不適切な関係をクリーニングした上で、トップ1万個を用いた。そのDe Saegerらの文献の方法による関係獲得でパターン学習に用いたデータは約5千万文書で、対象の単語集合は、上記と同じ約50万名詞である。その方法の詳細は文献に譲るが、シードパターンを入力し、それらシードパターンと同じ2語を抽出できる全パターンを用いて関係を再獲得してランキングするため、パターンベースの方法では、最高レベルの網羅性と考えられる。前述の評価法と同様の基準で関係情報の精度を測定したところ、lenientで0.80、strictで0.70であった。つまり、ノイズが含まれる関係情報からの類推となる。ただし、本評価は、De Saegerらの文献と方法が異なり、De Saegerらの文献と比較するとやや低めの値となる傾向にある点に注意されたい。

20

【0099】

文献: S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, M. Murata, 「Large Scale Relation Acquisition Using Class Dependent Patterns」, In Proc. of the 9th ICDM, p. 764 - 769, 2009年

【0100】

これらの類似情報、関係情報、単語共起頻度を用いて、拡張関係候補情報を生成し、共起スコアを用いた選択を行ったところ、1万の関係情報から102290個の新しい関係 (拡張関係候補情報) が生成された。

30

【0101】

次に、この実験例で用いた経由スコアについて説明する。この実験例では、経由スコアとして、前述のsumとprodとの2種類を用い、それぞれでランキングした結果を評価した。精度は、関係情報の関係を除いた上で、各経由スコアのトップ1万から100個、1万から3万の100個の200個を評価した。結果を図8に示す。図8の15000位以降の精度は、トップ1万までの精度と、トップ1万から3万までの精度を用いて補間したものである。lenientを正解とすると、prod (prod) は、トップ1万の精度が0.63、sum (sum) は、0.53であった。関係情報の精度 (lenientで0.80) と比較するとやや精度が低下しているが、文中での書かれ方を用いずに、これだけの精度を達成できた。なお、トップ1万は、1万個の関係情報から1万個の拡張関係情報 (これには関係情報は含まれない) を取得したことになる。したがって、少しの精度の低下によって、関係の個数を倍に拡張できることが分かる。また、図8の通り、sumとprodを比較すると、prodの方が上位の精度が高いため、よいスコアとなっていると考えられる。

40

【0102】

次に、パターンベースで獲得困難である関係を取得できたかどうかを調べた。具体的には、トップ1万のlenientを正解と考え、De Saegerらの文献の方法で、順位が100万位以下である関係の数の割合を100個のサンプリングを用いて調査し、

50

その割合を用いてトップ1万に含まれる正解の係数を推定した。ただし、De Saegerらの文献の方法では、5千万文書を用いていることに対して、この実験例では1億文書での共起頻度を用いているので、フェアな比較とはいえない。厳密には、文書集合を揃えた上で比較すべきで、De Saegerらとの比較は、参考的なものである。結果は、次のようになった。

【0103】

	De Saegerら方法で100万位以下の係数
sum	約3100
prod	約3300

【0104】

上記の結果より、この実験では、パターンベースの従来法では獲得困難であった関係が実際に獲得されていることが確認できた。例えば、関係情報に含まれなかった関係として、<ミネラル不足、花粉症>や、<食習慣、ニキビ>を取得することができた。なお、前者は、関係情報<カルシウム不足、アトピー>等から両方の単語を置換することによって生成されたものであり、後者は、関係情報<生活習慣、ニキビ>等から一方の単語を置換することによって生成されたものである。このように、本実験例では、関係情報に含まれる両方の単語を置換した場合であっても、スコアを用いた選択を行うことによって、精度を維持できていると考えられる。

【0105】

以上のように、本実施の形態による関係情報拡張装置1によれば、類似情報を用いて関係情報を拡張するため、従来のパターンベースの手法では獲得することのできなかつた関係を生成することができる。また、言語表現を類似する言語表現に置換することによって新たな関係を生成するため、語基の共通しないものに言語表現を置換することができ、前述の非特許文献2の場合よりも、より広範囲な拡張が可能となる。また、拡張した関係に対して、スコアを取得し、そのスコアを用いた選択を行うことによって、不適切な関係を除去することができる。したがって、選択された拡張関係情報は、意味的關係を適切に有するものになりうる。

【0106】

ここで、このようにして生成された拡張関係情報によって示される関係の使い方について簡単に説明する。例えば、本実施の形態による関係情報拡張装置1によって、拡張関係情報<心筋梗塞、急死>、<脳梗塞、急死>、<脳卒中、急死>を得ることができていた場合には、情報検索システムのキーワード推薦として、ユーザが「急死」を入力した場合に、「"急死"の"原因"には、「心筋梗塞」、「脳梗塞」、「脳卒中」などがあります」等のように、意味的關係で整理された推薦を行うことができるようになる。また、拡張関係情報に、上位<薬、抗ウイルス薬剤>や、効果<抗ウイルス薬剤、インフルエンザ>が存在する場合には、「インフルエンザに効く薬は？」という質問に対して、適切な推論によって「抗ウイルス薬剤」と答えることができる。また、拡張関係情報の示す関係を、他の種々の用途で用いることもできる。なお、その際に、拡張関係情報のみを用いてもよく、拡張関係情報と関係情報とをマージしたものをを用いてもよい。

【0107】

なお、本実施の形態による関係情報拡張装置1において、スコアを取得する方法は、前述のものに問われないことは言うまでもない。例えば、スコア取得部15が取得するスコアは、共起スコアと、経路スコアとを2個の引数とする各引数に関する増加関数の値であってもよい。そして、増加関数の値であるスコアを用いて、選択が行われてもよい。その増加関数は、例えば、 $C1 \times \text{共起スコア} + C2 \times \text{経路スコア}$ であってもよい。ここで、 $C1$ 、 $C2$ は、正の係数である。

【0108】

また、本実施の形態による関係情報拡張装置1において、共起スコアを取得する方法は、前述のものに問われないことは言うまでもない。例えば、スコア取得部15は、拡張関係候補情報に含まれる2以上の言語表現と、その拡張関係候補情報に対応する共起言語表

10

20

30

40

50

現とが共起する方が、その拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。ここで、拡張関係候補情報に対応する共起言語表現とは、その拡張関係候補情報の生成時に用いられた関係情報の意味的關係と同じ種類の意味的關係を有する各関係情報(この関係情報は、関係情報記憶部11で記憶されている関係情報であってもよく、あるいは、そうでなくてもよい。)に含まれる2以上の言語表現に対して共起の高い言語表現である。なお、共起が高い言語表現とは、前述のスコアが高い場合と同様に、例えば、しきい値以上の共起の頻度である言語表現であってもよく、あるいは、共起の頻度の高いものからあらかじめ決められた個数の言語表現であってもよい。具体例を用いて説明すると次のようになる。例えば、拡張関係候補情報<心筋梗塞、急死>が、関係情報<心筋梗塞、死亡>を用いて生成されたとする。そして、その関係情報<心筋梗塞、死亡>の意味的關係の種類が「原因」であったとする。また、意味的關係の種類が「原因」である2以上(多数であることが好適である)の各関係情報に含まれる2以上の言語表現(すべての言語表現)と共起の高い言語表現として、「原因」「理由」「要因」...があったとする。すると、「原因」「理由」「要因」...が、共起言語表現になる。また、拡張関係候補情報<心筋梗塞、急死>に含まれる2個の言語表現が、共起言語表現「原因」「理由」「要因」...のいずれかと共起する場合には、拡張関係候補情報<心筋梗塞、急死>に含まれる2個の言語表現が、共起言語表現「原因」「理由」「要因」...のいずれとも共起しない場合に比べて、共起スコアは高くなる。その場合に、(1)保持している共起言語表現を用いて共起スコアを取得する方法と、(2)機械学習を用いて共起スコアを取得する方法とがある。以下、その各方法について説明する。なお、(1)(2)以外の方法によって、上述のように共起スコアを取得してもよいことは言うまでもない。

【0109】

(1)保持している共起言語表現を用いて共起スコアを取得する方法

この方法では、関係情報拡張装置は、図9で示されるように、1以上の対応情報が記憶される対応情報記憶部21をさらに備えている。ここで、対応情報は、種類識別情報と、その種類識別情報に対応する、その種類識別情報で識別される意味的關係の種類に対応する1以上の共起言語表現とを有する情報である。ここで、種類識別情報は、関係情報の意味的關係の種類を識別する情報である。本方法の場合には、関係情報記憶部11で記憶されている各関係情報は、その関係情報が有する2以上の言語表現の意味的關係の種類を識別する情報である種類識別情報をも有するものであるとする。ここでは、関係情報が、「種類識別情報<第1の言語表現、第2の言語表現>」の形式で示されるものとする。例えば、「原因<心筋梗塞、死亡>」となる。また、拡張関係候補情報生成部13は、拡張関係候補情報の生成に用いる関係情報が有する種類識別情報を有する拡張関係候補情報を生成するものとする。したがって、関係情報「原因<心筋梗塞、死亡>」を用いて生成された拡張関係候補情報は、例えば、「原因<脳梗塞、死亡>」となる。そして、スコア取得部15は、拡張関係候補情報に含まれる2以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、その拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコアを取得するものとする。ここで、「拡張関係候補情報に含まれる2以上の言語表現のみが共起する」とは、その拡張関係候補情報に含まれる2以上の言語表現が、共起言語表現のいずれとも共起しないことである。

【0110】

なお、共起スコアが、拡張関係候補情報に含まれる2以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、その拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となることは、結果としてそのようになればよいのであって、その方法は問わない。例えば、スコア取得部15は、拡張関係候補情報に含まれる2以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応するいずれかの共起言語表現とが共起する場合には、前述のように、拡張関係候補情報に含まれる2以上の言語表現に対する共起頻度などを用いて取得した共

10

20

30

40

50

起の尺度に対して、1を超える数（例えば、1.2や1.5、2など）を掛けた値を共起スコアとしてもよい。また、共起する共起言語表現の数が多いほど、共起スコアが高くなるようにしてもよい。例えば、ある拡張関係候補情報について、その拡張関係候補情報が有する種類識別情報に対応する共起言語表現の数がAであるとする。また、そのA個の共起言語表現のうち、その拡張関係候補情報に含まれる2以上の言語表現と共起する共起言語表現の数がBであったとする。また、 $R = B / A$ とする。そして、スコア取得部15は、Rを引数とする増加関数の値である共起スコアを取得してもよい。具体的には、前述のように、共起頻度などを用いて取得した共起の尺度に対して、 $(1 + C \times R)$ を掛けた値を共起スコアにしてもよい。なお、Cは、正の係数である。さらに、共起言語表現との共起の程度が高いほど、共起スコアがより高くなるようにしてもよい。例えば、前述のBの値を、拡張関係候補情報に含まれる2以上の言語表現と共起言語表現との共起頻度の和などにしてもよい。

10

【0111】

ここで、具体例を用いて説明する。対応情報記憶部21において、図10で示される対応情報が記憶されていたとする。図10の対応情報において、種類識別情報と、共起言語表現とが対応付けられている。例えば、種類識別情報「食材」に、共起言語表現「材料」「レシピ」...が対応している。したがって、種類識別情報「食材」を有する関係情報（例えば、食材<シチュー、じゃがいも>のように、シチューの食材がじゃがいもであることを示す関係情報等）に含まれるすべての言語表現と共起の高い言語表現が、「材料」や「レシピ」等であることが示されていることになる。また、拡張関係候補情報記憶部14において、拡張関係候補情報「食材<シチュー、サツマイモ>」が記憶されていたとする。すると、スコア取得部15は、前述のように、2個の言語表現「シチュー」「サツマイモ」の共起の尺度を算出する。また、スコア取得部15は、その拡張関係候補情報「食材<シチュー、サツマイモ>」に含まれる種類識別情報「食材」に対応する共起言語表現「材料」「レシピ」...を、図10の対応情報を用いて取得する。そして、共起言語表現を順番に変えながら、3個の言語表現「シチュー」「サツマイモ」「共起言語表現」が共起するかどうか判断する。ここで、例えば、種類識別情報「食材」に対応する共起言語表現の総数が20個であり、そのうち、「シチュー」「サツマイモ」と共起した共起言語表現の個数が5個であったとする。すると、スコア取得部15は、前述のように、共起スコア = 共起の尺度 $\times (1 + C \times 5 / 20)$ を取得してもよい。一方、例えば、拡張関係候補情報「食材<シチュー、デンプン>」に対しては、2個の言語表現「シチュー」「サツマイモ」と共起する共起言語表現が存在しなかったとする。すると、その拡張関係候補情報「食材<シチュー、デンプン>」に対しては、スコア取得部15は、2個の言語表現「シチュー」「サツマイモ」の共起の尺度そのものを共起スコアにする。

20

30

【0112】

(2) 機械学習を用いて共起スコアを取得する方法

この方法では、スコア取得部15は、機械学習を用いて、共起スコアを取得する。すなわち、スコア取得部15は、2以上の言語表現の組に含まれるその2以上の言語表現と共起する言語表現を少なくとも素性として用い、その素性の値及び2以上の言語表現の組に対する意味的関係の有無（なお、この意味的関係の有無は、その言語表現の組に含まれる2以上の言語表現の意味的関係の有無である）を教師データとする機械学習を行い、拡張関係候補情報に含まれる2以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得する。その機械学習について、以下、説明する。

40

【0113】

この機械学習の問題（入力）は、共起スコアを取得する対象となる、言語表現の組（その言語表現の組は、2以上の言語表現を有している。また、その言語表現の組に含まれる言語表現の数は、関係情報に含まれる言語表現の数と同じであるとする。）である。具体的には、拡張関係候補情報である。また、その機械学習の解（出力）は、問題（入力）である2以上の言語表現の組（拡張関係候補情報）に含まれる2以上の言語表現が、その拡張関係候補情報に対応する意味的関係の種類と同じ意味的関係の種類を有する関係情報と

50

同様の共起であるかどうかの確信度である。拡張関係候補情報に対応する意味的關係の種類とは、その拡張関係候補情報の生成で用いられた関係情報の意味的關係の種類である。なお、その解（出力）には、問題（入力）の拡張関係候補情報に含まれる2以上の言語表現が、その拡張関係候補情報に応じた意味的關係の種類と同じ意味的關係の種類を有する関係情報と同様の意味的關係を有するかどうかの情報が含まれてもよい。また、その機械学習の素性には、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現（すべての言語表現）と共起する言語表現のリストが含まれるものとする。そのリストは、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現と共起するすべての言語表現のリストであってもよく、あるいは、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現と共起の高い言語表現のリストであってもよい。このリストを作成するためには、スコア取得部15は、例えば、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現と、あらゆる言語表現とが共起するかどうかを判断してもよい。ここで、その処理で用いられるあらゆる言語表現は、例えば、あらかじめ図示しない記録媒体で記憶されている言語表現群であってもよい。そして、共起すると判断された言語表現を、そのリストに含めるようにしてもよい。また、共起の高いもののみをリストに含める場合には、スコア取得部15は、共起の尺度（例えば、共起頻度や共起率等である。なお、ここで用いられる共起の尺度は、例えば、ダイス係数や相互情報量などのように、2個の言語表現に対してのみ定義されている共起の尺度を用いたものではなく、3個以上の言語表現に対しても定義されているものであることが好適である）も算出し、それに応じて共起の高いものを選択してもよい。なお、共起が高いものとは、前述のスコアが高い場合と同様に、例えば、しきい値以上の共起の頻度である言語表現であってもよく、あるいは、共起の頻度の高いものからあらかじめ決められた個数の言語表現であってもよい。また、その機械学習の素性には、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現の共起の尺度、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現の属性（例えば、言語表現の品詞や、言語表現の上位語等）、問題（入力）である2以上の言語表現の組としての拡張関係候補情報に対応する意味的關係の種類、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現と共起する言語表現のリストに含まれる各言語表現に関する、問題（入力）である2以上の言語表現の組に含まれる2以上の言語表現との共起の尺度（例えば、共起頻度や共起率等であり、前述のように、3個以上の言語表現に対しても定義されている共起の尺度であることが好適である）のうち、任意の1以上のものが素性に含まれてもよい。なお、意味的關係の種類を素性に用いる場合には、例えば、問題（入力）である2以上の言語表現の組に種類識別情報が含まれており、その種類識別情報を素性に用いてもよい。また、意味的關係の種類を素性に用いずに、意味的關係の種類ごとに学習を行い、その意味的關係の種類ごとの学習結果を用いて、共起スコアを取得してもよい。例えば、種類識別情報「原因」に対応する拡張関係候補情報に対する共起スコアを取得する際には、種類識別情報「原因」に対応して学習された学習結果を用いて、共起スコアを取得してもよい。

【0114】

また、その機械学習で用いられる教師データ（訓練データ）は、2以上の言語表現の組に対する意味的關係の有無と、その2以上の言語表現の組に対応する、前述の素性の各値とである。例えば、教師データの正例（すなわち、意味的關係のあるもの）である、種類識別情報「原因」に対応する2以上の言語表現の組としては、種類識別情報「原因」で識別される意味的關係の種類である関係情報を用いてもよい。その関係情報は、関係情報記憶部11で記憶されているものであってもよく、あるいは、そうでないものであってもよい。また、教師データの負例（すなわち、意味的關係のないもの）である、種類識別情報「原因」に対応する2以上の言語表現の組としては、任意のコーパスからランダムに取得した2以上の言語表現の組を用いてもよい。

【0115】

教師データを用いた学習の後に、判断の対象となる、拡張関係候補情報を入力すると、

10

20

30

40

50

その拡張関係候補情報に関する素性の各値が取得され、その拡張関係候補情報に含まれる2以上の言語表現の意味的關係に関する確信度が出力される。例えば、その確信度は、-1から1までの範囲の値であってもよい。また、先述のように、意味的關係を有するかどうかの結果も出力されてもよい。例えば、確信度が-1から0までであれば意味的關係を有しないという結果になり、確信度が0を超えて1までであれば意味的關係を有するという結果になる。スコア取得部15は、その確信度に応じた共起スコアを取得する。具体的には、確信度が-1から1までの値である場合には、共起スコアは、その確信度を引数とする増加関数の値であってもよい。具体的には、共起スコア = $C \times (\text{確信度} + 1)$ であってもよい。なお、Cは、任意の係数である。また、共起スコアが正の値になるように、確信度に1を足している。なお、拡張関係候補情報に含まれる2以上の言語表現が意味的關係を有する場合にも、有しない場合にも、確信度が0から1までの範囲の値であるのであれば、意味的關係を有する場合には、共起スコア = $C \times (1 + \text{確信度})$ として、意味的關係を有しない場合には、共起スコア = $C \times (1 - \text{確信度})$ としてもよい。また、共起スコアは、拡張関係候補情報に含まれる2以上の言語表現が意味的關係を有するという結果の場合の確信度のみを用いてもよい。その場合には、例えば、共起スコア = $C \times \text{確信度}$ であってもよい。このように、機械学習を用いて共起スコアを取得することによって、結果として、拡張関係候補情報に含まれる2以上の言語表現と、その拡張関係候補情報に対応する共起言語表現とが共起する方が、その拡張関係候補情報に含まれる2以上の言語表現のみが共起するよりも高い値となる共起スコアを取得できることになる。

【0116】

また、本実施の形態では、類似情報記憶部12で記憶されているすべての類似情報を用いて拡張関係候補情報を生成する場合について説明したが、そうでなくてもよい。すなわち、拡張関係候補情報生成部13は、類似情報記憶部12で記憶されている一部の類似情報を用いて、拡張関係候補情報の生成を行ってもよい。そのため、例えば、関係情報記憶部11で記憶されている関係情報は、その関係情報が有する2以上の言語表現の意味的關係の種類を識別する情報である種類識別情報をも有するものであってもよい。また、類似情報記憶部12では、種類識別情報と、その種類識別情報に対応する類似情報とが記憶されていてもよい。そして、拡張関係候補情報生成部13は、関係情報記憶部11で記憶されている関係情報に含まれる少なくとも1個の言語表現を置換する際に、その関係情報が有する種類識別情報に対応する類似情報を用いて置換を行ってもよい。

【0117】

具体的には、類似情報記憶部12で、図11で示される類似情報が記憶されていたとする。図11において、類似情報と、種類識別情報とが対応付けられている。なお、図11の類似情報は、図5の類似情報とは異なり、互いに類似する2個の言語表現のみを対応付ける情報である。そして、関係情報記憶部11において、関係情報「名産<愛媛、みかん>」が記憶されていたとする。この関係情報は、愛媛の名産がみかんであることを示すものである。この関係情報を用いて拡張関係候補情報を生成する場合には、拡張関係候補情報生成部13は、その関係情報から種類識別情報「名産」を取得し、その種類識別情報「名産」に対応付けられている類似情報を特定する。そして、拡張関係候補情報生成部13は、その特定した類似情報を用いて、関係情報「名産<愛媛、みかん>」に含まれる各言語表現「愛媛」「みかん」の少なくとも1個を置換した拡張関係候補情報を生成する。具体的には、愛媛が香川に置換された拡張関係候補情報「名産<香川、みかん>」等が生成され、拡張関係候補情報記憶部14に蓄積されることになる。なお、拡張関係候補情報には、種類識別情報が含まれていてもよく、あるいは、含まれていなくてもよい。なお、図11で示される種類識別情報と、類似情報との対応は、手作業で生成されたものであってもよく、あるいは、その他の方法によって生成されたものであってもよい。

【0118】

なお、ここでは、種類識別情報に対応する類似情報を用いて拡張関係候補情報を生成する場合について説明したが、さらに、種類識別情報と、置換対象でない言語表現とに対応する類似情報を用いて、拡張関係候補情報を生成してもよい。その場合にも、関係情報は

種類識別情報を有するものであるとする。また、類似情報記憶部 12 では、種類識別情報と、置換対象でない言語表現と、それらに対応する類似情報とが記憶されているものとする。そして、拡張関係候補情報生成部 13 は、関係情報記憶部 11 で記憶されている関係情報に含まれる 1 個の言語表現を置換する際に、その関係情報が有する種類識別情報と、その関係情報に含まれる置換対象ではない言語表現とに対応する類似情報を用いて置換を行うものとする。ここで、関係情報に N 個 (N は 2 以上の整数) の言語表現が含まれている場合には、置換対象でない言語表現は、(N - 1) 個となる。したがって、その場合には、類似情報記憶部 12 において、類似情報は、種類識別情報と、(N - 1) 個の置換対象でない言語表現とに対応付けられていることになる。例えば、N = 2 の場合に、類似情報記憶部 12 において、図 12 で示される情報が記憶されていたとする。図 12 において、種類識別情報と、1 個の置換対象でない言語表現と、類似情報とが対応付けられている。また、図 12 の類似情報は、図 11 の場合と同様に、2 個の言語表現を対応付ける類似情報である。そして、例えば、拡張関係候補情報生成部 13 が、関係情報「原因<心筋梗塞、死亡>」を用いて拡張関係候補情報を生成する処理について説明する。その処理において、関係情報の第 1 の言語表現「心筋梗塞」の置換を行う場合には、拡張関係候補情報生成部 13 は、その関係情報の種類識別情報「原因」と、置換対象でない言語表現「死亡」とを取得する。そして、図 12 の情報を参照し、それらに対応する類似情報を特定する。そして、その特定した類似情報を用いて、第 1 の言語表現「心筋梗塞」を、「脳梗塞」等に置換した拡張関係候補情報「原因<脳梗塞、死亡>」「原因<脳卒中、死亡>」等を生成して、拡張関係候補情報記憶部 14 に蓄積する。なお、図 12 で示される種類識別情報と、類似情報との対応は、手作業で生成されたものであってもよく、あるいは、その他の方法によって生成されてもよい。後者の場合には、例えば、置換対象でない言語表現を含む文書のみから、類似情報を生成してもよい。または、例えば、図 11 の各レコードにおいて、置換対象でない言語表現と、類似情報に含まれるすべての言語表現との共起が高い場合に、そのレコードに、その置換対象でない言語表現を含めたレコードを作成し、図 12 のレコードとしてもよい。例えば、図 11 の 1 番目のレコードにおいて、置換対象でない言語表現「みかん」と、類似情報に含まれるすべての言語表現「愛媛」「香川」との共起が高いとすると、その図 11 の 1 番目のレコードに置換対象でない言語表現「みかん」を追加したレコードを生成し、図 12 の情報に追加してもよい。なお、図 12 の情報は、拡張関係候補情報の生成時に、一時的に生成されて類似情報記憶部 12 で記憶されるものであってもよい。例えば、種類識別情報が「原因」である場合に、置換対象でない言語表現が「死亡」であれば、それらに対応する類似情報を生成して類似情報記憶部 12 に蓄積し、次に、置換対象でない言語表現が「急死」になれば、種類識別情報「原因」、置換対象でない言語表現「急死」に対応する類似情報を生成して類似情報記憶部 12 に蓄積するようにしてもよい。その類似情報等の生成は、例えば、拡張関係候補情報生成部 13 が行ってもよく、他の構成要素が行ってもよい。

【0119】

また、本実施の形態による関係情報拡張装置 1 は、拡張関係情報記憶部 17 で記憶された拡張関係情報や、あるいは、拡張関係候補情報記憶部 14 においてフラグ等によって拡張関係情報であることが示された拡張関係候補情報を出力する出力部を備えてもよい。その出力部による出力は、例えば、表示デバイス (例えば、CRT や液晶ディスプレイなど) への表示でもよく、所定の機器への通信回線を介した送信でもよく、プリンタによる印刷でもよく、記録媒体への蓄積でもよい。なお、その出力部は、出力を行うデバイス (例えば、表示デバイスやプリンタなど) を含んでもよく、あるいは含まなくてもよい。また、その出力部は、ハードウェアによって実現されてもよく、あるいは、それらのデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

【0120】

また、本実施の形態による関係情報拡張装置 1 が処理を行う関係情報や拡張関係候補情報等に含まれる言語表現の言語は問わない。言語表現は、例えば、日本語や英語、ドイツ語、フランス語、ロシア語、中国語、スペイン語等で記述されたものであってもよい。た

10

20

30

40

50

だし、関係情報記憶部 1 1、類似情報記憶部 1 2、拡張関係候補情報記憶部 1 4、拡張関係情報記憶部 1 7 で記憶される関係情報等の言語は、すべて共通しているものとする。

【 0 1 2 1 】

[機械学習に関する説明]

ここで、上記実施の形態で用いられる機械学習について説明する。機械学習の手法は、問題 - 解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どのような問題のときにどのような解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるようにする方法である。例えば、次の文献を参照されたい。

【 0 1 2 2 】

文献：村田真樹、「機械学習に基づく言語処理」, 龍谷大学理工学部・招待講演、2004年 (<http://www2.nict.go.jp/jt/a132/members/murata/ps/rk1-siryou.pdf>)

文献：村田真樹, 馬青, 内元清貴, 井佐原均、「サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳」, 電子情報通信学会言語理解とコミュニケーション研究会 NLC2000-78, 2001年

文献：村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均、「NSEVAL2」辞書タスクでのCRLの取り組み」, 電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-40, 2001年

【 0 1 2 3 】

機械学習アルゴリズムを動作させるために、問題の状況を機械に伝える際に、素性（解析に用いる情報で問題を構成する各要素）というものが必要になる。問題を素性によって表現するのである。例えば、日本語文末表現の時制の推定の問題において、問題：「彼が話す。」 - - 解「現在」が与えられた場合に、素性の一例は、「彼が話す。」「が話す。」「話す。」「す」「。」となる。

【 0 1 2 4 】

すなわち、機械学習の手法は、素性の集合 - 解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どのような素性の集合のときにどのような解になるかを学習し、その学習結果を利用して、新しい問題のときも、その問題から素性の集合を取り出して、その素性に対応する解を推測する方法である。なお、ここで、「解」とは、例えば、前述の回答情報であるかどうかや、分類情報などである。

【 0 1 2 5 】

機械学習の手法として、例えば、k近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いることができる。なお、以下の説明では、文書を分類する場合（問題 - 解のセットが、文 - 分類である場合）の機械学習について主に説明するが、それ以外の機械学習についても、同様に適用可能であることは言うまでもない。

【 0 1 2 6 】

k近傍法は、最も類似する一つの事例のかわりに、最も類似するk個の事例を用いて、このk個の事例での多数決によって解（分類）を求める手法である。kは、あらかじめ定める整数の数字であって、一般的に、1から9の間の奇数を用いる。

【 0 1 2 7 】

シンプルベイズ法は、ベイズの定理にもとづいて各解（分類）の確率を推定し、その確率値が最も大きい解を、求める解とする方法である。

【 0 1 2 8 】

シンプルベイズ法において、文脈bで分類aを出力する確率は、次式で与えられる。

10

20

30

40

【数6】

$$p(a|b) = \frac{p(a)}{p(b)} p(b|a)$$

$$\cong \frac{\tilde{p}(a)}{p(b)} \prod_i \tilde{p}(f_i|a)$$

【0129】

ただし、ここで文脈 b は、あらかじめ設定しておいた素性 f_j ($F, 1 \leq j \leq k$) の集合である。 $p(b)$ は、文脈 b の出現確率である。ここで、分類 a に非依存であって定数のために計算しない。 $P(a)$ (ここで P は p の上部にチルダ) と $P(f_i|a)$ は、それぞれ教師データから推定された確率であって、分類 a の出現確率、分類 a のときに素性 f_i を持つ確率を意味する。 $P(f_i|a)$ として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、上記の2行目の式の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スムージングを行う。ここでは、次式を用いてスムージングを行ったものを用いる。

10

【数7】

$$p(f_i|a) = \frac{\text{freq}(f_i, a) + 0.01 * \text{freq}(a)}{\text{freq}(a) + 0.01 * \text{freq}(a)}$$

20

【0130】

ただし、 $\text{freq}(f_i, a)$ は、素性 f_i を持ち、かつ分類が a である事例の個数、 $\text{freq}(a)$ は、分類が a である事例の個数を意味する。

なお、スムージングは、上記式を用いた方法に限られるものではなく、その他の方法を用いてもよいことは言うまでもない。

【0131】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力を与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

30

【0132】

決定リスト方法では、あらかじめ設定しておいた素性 f_j ($F, 1 \leq j \leq k$) のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈 b で分類 a を出力する確率は、次式によって与えられる。

【数8】

$$p(a|b) = p(a|f_{\max})$$

【0133】

ただし、 f_{\max} は、次式によって与えられる。

【数9】

$$f_{\max} = \arg \max_{f_j \in F} \max_{a_i \in A} \tilde{p}(a_i|f_j)$$

40

【0134】

また、 $P(a_i|f_j)$ (ここで P は p の上部にチルダ) は、素性 f_j を文脈に持つ場合の分類 a_i の出現の割合である。

【0135】

最大エントロピー法は、あらかじめ設定しておいた素性 f_j ($1 \leq j \leq k$) の集合を F とするとき、以下の所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布 $p(a, b)$ を求め、その確率分布にしたがって求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。

50

【 0 1 3 6 】

所定の条件式は、次式で与えられる。

【 数 1 0 】

$$\sum_{a \in A, b \in B} p(a, b) g_j(a, b) = \sum_{a \in A, b \in B} \tilde{p}(a, b) g_j(a, b)$$

for $\forall f_j (1 \leq j \leq k)$

【 0 1 3 7 】

また、エントロピーを意味する式は、次式で与えられる。

【 数 1 1 】

$$H(p) = - \sum_{a \in A, b \in B} p(a, b) \log(p(a, b))$$

10

【 0 1 3 8 】

ただし、A、Bは分類と文脈の集合を意味する。また、 $g_j(a, b)$ は文脈bに素性 f_j があって、なおかつ分類がaの場合1となり、それ以外で0となる関数を意味する。また、 $P(a_i | f_j)$ (ここでPはpの上部にチルダ)は、既知データでの(a, b)の出現の割合を意味する。

【 0 1 3 9 】

上記の条件式は、確率pと出力と素性の組の出現を意味する関数gをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化(確率分布の平滑化)を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の文献を参照されたい。

20

【 0 1 4 0 】

文献: Eric Sven Ristad, 「Maximum Entropy Modeling for Natural Language」, (ACL/EACL Tutorial Program, Madrid, 1997年)

文献: Eric Sven Ristad, 「Maximum Entropy Modeling Toolkit, Release 1.6 beta」, (<http://www.mnemonic.com/software/memt>), 1998年

30

【 0 1 4 1 】

サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。

【 0 1 4 2 】

図13にサポートベクトルマシン法のマージン最大化の概念を示す。図13において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図13(A)は、正例と負例の間隔が狭い場合(スモールマージン)の概念図、図13(B)は、正例と負例の間隔が広い場合(ラージマージン)の概念図である。

40

【 0 1 4 3 】

このとき、二つの分類が正例と負例からなるものとする、学習データにおける正例と負例の間隔(マージン)が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図13(B)に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【 0 1 4 4 】

基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張(カーネル関数の導入)がなされたものが用いられる。

【 0 1 4 5 】

50

この拡張された方法は、以下の識別関数（ $f(x)$ ）を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【数 1 2】

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b\right) \quad (\text{M1})$$

$$b = -\frac{\max_{i, y_i=-1} b_i + \min_{i, y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^l \alpha_j y_j K(x_j, x_i)$$

10

【0 1 4 6】

ただし、 x は識別したい事例の文脈（素性の集合）を、 x_i と y_j （ $i = 1, \dots, l$, $y_j \in \{1, -1\}$ ）は学習データの文脈と分類先を意味し、関数 sgn は、

$$\operatorname{sgn}(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (\text{otherwise}) \end{cases}$$

であり、また、各 α_i は、式 (M3) と式 (M4) の制約のもと、式 (M2) を最大にする場合のものである。

【数 1 3】

20

$$L(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (\text{M2})$$

$$0 \leq \alpha_i \leq C \quad (i = 1, \dots, l) \quad (\text{M3})$$

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad (\text{M4})$$

【0 1 4 7】

また、関数 K はカーネル関数と呼ばれ、様々なものが用いられるが、本形態では、例えば、以下の多項式のものを用いる。

30

$$K(x, y) = (x \cdot y + 1)^d$$

【0 1 4 8】

ここで、 C 、 d は実験的に設定される定数である。例えば、 C はすべての処理を通して 1 に固定した。また、 d は、1 と 2 の二種類を試している。ここで、 $\alpha_i > 0$ となる x_i は、サポートベクトルと呼ばれ、通常、式 (M1) の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【0 1 4 9】

なお、拡張されたサポートベクトルマシン法の詳細については、次の文献を参照されたい。

40

文献：Nello Cristianini, John Shawe-Taylor, 「An Introduction to Support Vector Machines and other kernel-based learning methods」, Cambridge University Press, 2000年

文献：Taku Kudoh, 「Tinysvm: Support Vector machines」, (<http://cl.aistnara.ac.jp/taku-ku/software/TinySVM/index.html>), 2000年

【0 1 5 0】

サポートベクトルマシン法は、分類の数が 2 個のデータを扱うものである。したがって

50

、分類の数が3個以上の事例を扱う場合には、通常、これにペアワイズ法またはワンVSレスト法などの手法を組み合わせる用いることになる。

【0151】

ペアワイズ法は、 n 個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア($n(n-1)/2$ 個)を生成し、ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、 $n(n-1)/2$ 個の二値分類による分類先の多数決によって、分類先を求める方法である。

【0152】

ワンVSレスト法は、例えば、 a 、 b 、 c という三つの分類先があるときは、分類先 a とその他、分類先 b とその他、分類先 c とその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定すべき問題が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある解くべき問題が、「分類先 a とその他」の組の学習処理で作成したサポートベクトルマシンにおいて分離平面から最も離れた場合には、その解くべき問題の分類先は、 a と推定する。

10

【0153】

図示しない解推定手段が推定する、解くべき問題についての、どのような解(分類先)になりやすいかの度合いの求め方は、図示しない機械学習手段が機械学習の手法として用いる様々な方法によって異なる。

20

【0154】

例えば、機械学習手段が、機械学習の手法として k 近傍法を用いる場合、機械学習手段は、教師データの事例同士で、その事例から抽出された素性の集合のうち重複する素性の割合(同じ素性をいくつ持っているかの割合)にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として学習結果記憶手段に記憶しておく。

【0155】

そして、解推定手段は、解くべき問題の素性(文章群属性情報)が抽出されたときに、学習結果記憶手段において定義された類似度と事例を参照して、抽出された解くべき問題の素性について、その解くべき問題の素性の類似度が高い順に k 個の事例を学習結果記憶手段の事例から選択し、選択した k 個の事例での多数決によって決まった分類先を、解くべき問題の分類先(解)として推定する。すなわち、解推定手段では、解くべき問題についての、どのような解(分類先)になりやすいかの度合いを、選択した k 個の事例での多数決の票数とする。

30

【0156】

また、機械学習手法として、シンプルベイズ法を用いる場合には、図示しない機械学習手段は、教師データの事例について、前記事例の解と素性の集合との組を学習結果情報として学習結果記憶手段に記憶する。そして、解推定手段は、解くべき問題の素性が抽出されたときに、学習結果記憶手段の学習結果情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて、解くべき問題の素性の集合の場合の各分類になる確率を算出して、その確率の値が最も大きい分類を、その解くべき問題の素性の分類(解)と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

40

【0157】

また、機械学習手法として決定リスト法を用いる場合には、図示しない機械学習手段は、教師データの事例について、素性と分類先との規則を所定の優先順序で並べたリストを、予め、何らかの手段により、学習結果記憶手段に記憶させる。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段のリストの優先順序の高い順に、抽出された解くべき問題の素性と規則の素性とを比較し、素性が一致した規則の分類

50

先をその解くべき問題の分類先（解）として推定する。

【0158】

また、機械学習手法として最大エントロピー法を使用する場合には、図示しない機械学習手段は、教師データの事例から解となりうる分類を特定し、所定の条件式を満足し、かつエントロピーを示す式を最大にするときの素性の集合と解となりうる分類の二項からなる確率分布を求めて、学習結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の確率分布を利用して、抽出された解くべき問題の素性の集合についてその解となりうる分類の確率を求めて、最も大きい確率値を持つ解となりうる分類を特定し、その特定した分類をその解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

10

【0159】

また、機械学習手法としてサポートベクトルマシン法を使用する場合には、図示しない機械学習手段は、教師データの事例から解となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数にしたがって事例の素性の集合を次元とする空間上で、その事例の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて学習結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の超平面を利用して、解くべき問題の素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果にもとづいて定まる分類を、その解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、分離平面からのその解くべき問題の事例への距離の大きさとする。

20

【0160】

また、上記各実施の形態において、ある構成要素が機械学習を用いて処理を行う場合に、その所望の処理が実行されるまでに学習が行われるのであれば、その学習のタイミングは問わない。

【0161】

また、上記実施の形態では、関係情報拡張装置1がスタンドアロンである場合について説明したが、関係情報拡張装置1は、スタンドアロンの装置であってもよく、サーバ・クライアントシステムにおけるサーバ装置であってもよい。後者の場合には、拡張関係情報等が、通信回線を介して出力されてもよい。

30

【0162】

また、上記実施の形態において、各処理または各機能は、単一の装置または単一のシステムによって集中処理されることによって実現されてもよく、あるいは、複数の装置または複数のシステムによって分散処理されることによって実現されてもよい。

【0163】

また、上記実施の形態において、各構成要素が実行する処理に関する情報、例えば、各構成要素が受け付けたり、取得したり、選択したり、生成したり、送信したり、受信したりした情報や、各構成要素が処理で用いるしきい値や数式、アドレス等の情報等は、上記説明で明記していない場合であっても、図示しない記録媒体において、一時的に、あるいは長期にわたって保持されていてもよい。また、その図示しない記録媒体への情報の蓄積を、各構成要素、あるいは、図示しない蓄積部が行ってもよい。また、その図示しない記録媒体からの情報の読み出しを、各構成要素、あるいは、図示しない読み出し部が行ってもよい。

40

【0164】

また、上記実施の形態において、各構成要素等で用いられる情報、例えば、各構成要素が処理で用いるしきい値やアドレス、各種の設定値等の情報がユーザによって変更されてもよい場合には、上記説明で明記していない場合であっても、ユーザが適宜、それらの情報を変更できるようにしてもよく、あるいは、そうでなくてもよい。それらの情報をユーザが変更可能な場合には、その変更は、例えば、ユーザからの変更指示を受け付ける図示

50

しない受付部と、その変更指示に応じて情報を変更する図示しない変更部とによって実現されてもよい。その図示しない受付部による変更指示の受け付けは、例えば、入力デバイスからの受け付けでもよく、通信回線を介して送信された情報の受信でもよく、所定の記録媒体から読み出された情報の受け付けでもよい。

【 0 1 6 5 】

また、上記実施の形態において、関係情報拡張装置 1 に含まれる 2 以上の構成要素が通信デバイスや入力デバイス等を有する場合に、2 以上の構成要素が物理的に単一のデバイスを有してもよく、あるいは、別々のデバイスを有してもよい。

【 0 1 6 6 】

また、上記実施の形態において、各構成要素は専用のハードウェアにより構成されてもよく、あるいは、ソフトウェアにより実現可能な構成要素については、プログラムを実行することによって実現されてもよい。例えば、ハードディスクや半導体メモリ等の記録媒体に記録されたソフトウェア・プログラムを CPU 等のプログラム実行部が読み出して実行することによって、各構成要素が実現され得る。なお、上記実施の形態における関係情報拡張装置 1 を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、意味的關係を有する 2 以上の言語表現を有する関係情報が記憶される関係情報記憶部で記憶されている関係情報に含まれる少なくとも 1 個の言語表現を、類似する 2 以上の言語表現を有する類似情報が 2 以上記憶される類似情報記憶部で記憶されている類似情報を用いて、言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報を、拡張関係候補情報が記憶される拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部、拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的關係を有する確からしさを示すスコアを取得するスコア取得部、スコア取得部が取得したスコアを用いて、拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部として機能させるためのプログラムである。

【 0 1 6 7 】

なお、上記プログラムにおいて、情報を送信する送信ステップや、情報を受信する受信ステップなどでは、ハードウェアでしか行われない処理、例えば、送信ステップにおけるモデムやインターフェースカードなどで行われる処理は少なくとも含まれない。

【 0 1 6 8 】

また、このプログラムは、サーバなどからダウンロードされることによって実行されてもよく、所定の記録媒体（例えば、CD-ROM などの光ディスクや磁気ディスク、半導体メモリなど）に記録されたプログラムが読み出されることによって実行されてもよい。また、このプログラムは、プログラムプロダクトを構成するプログラムとして用いられてもよい。

【 0 1 6 9 】

また、このプログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【 0 1 7 0 】

図 1 4 は、上記プログラムを実行して、上記実施の形態による関係情報拡張装置 1 を実現するコンピュータの外観の一例を示す模式図である。上記実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムによって実現されうる。

【 0 1 7 1 】

図 1 4 において、コンピュータシステム 9 0 0 は、CD-ROM (Compact Disk Read Only Memory) ドライブ 9 0 5、FD (Floppy (登録商標) Disk) ドライブ 9 0 6 を含むコンピュータ 9 0 1 と、キーボード 9 0 2 と、マウス 9 0 3 と、モニター 9 0 4 とを備える。

【 0 1 7 2 】

図 1 5 は、コンピュータシステム 9 0 0 の内部構成を示す図である。図 1 5 において、コンピュータ 9 0 1 は、CD-ROM ドライブ 9 0 5、FD ドライブ 9 0 6 に加えて、M

10

20

30

40

50

PU (Micro Processing Unit) 911と、ブートアッププログラム等のプログラムを記憶するためのROM 912と、MPU 911に接続され、アプリケーションプログラムの命令を一時的に記憶すると共に、一時記憶空間を提供するRAM (Random Access Memory) 913と、アプリケーションプログラム、システムプログラム、及びデータを記憶するハードディスク914と、MPU 911、ROM 912等を相互に接続するバス915とを備える。なお、コンピュータ901は、LANへの接続を提供する図示しないネットワークカードを含んでいてもよい。

【0173】

コンピュータシステム900に、上記実施の形態による関係情報拡張装置1の機能を実行させるプログラムは、CD-ROM 921、またはFD 922に記憶されて、CD-ROMドライブ905、またはFDドライブ906に挿入され、ハードディスク914に転送されてもよい。これに代えて、そのプログラムは、図示しないネットワークを介してコンピュータ901に送信され、ハードディスク914に記憶されてもよい。プログラムは実行の際にRAM 913にロードされる。なお、プログラムは、CD-ROM 921やFD 922、またはネットワークから直接、ロードされてもよい。

10

【0174】

プログラムは、コンピュータ901に、上記実施の形態による関係情報拡張装置1の機能を実行させるオペレーティングシステム(OS)、またはサードパーティプログラム等を必ずしも含んでいなくてもよい。プログラムは、制御された態様で適切な機能(モジュール)を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいてもよい。コンピュータシステム900がどのように動作するのかについては周知であり、詳細な説明は省略する。

20

【0175】

また、本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【0176】

以上より、本発明による関係情報拡張装置等によれば、関係情報を適切に拡張することができるという効果が得られ、新たな関係を取得する装置等として有用である。

30

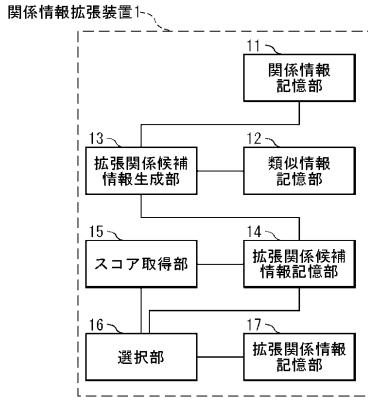
【符号の説明】

【0177】

- 1 関係情報拡張装置
- 11 関係情報記憶部
- 12 類似情報記憶部
- 13 拡張関係候補情報生成部
- 14 拡張関係候補情報記憶部
- 15 スコア取得部
- 16 選択部
- 17 拡張関係情報記憶部
- 21 対応情報記憶部

40

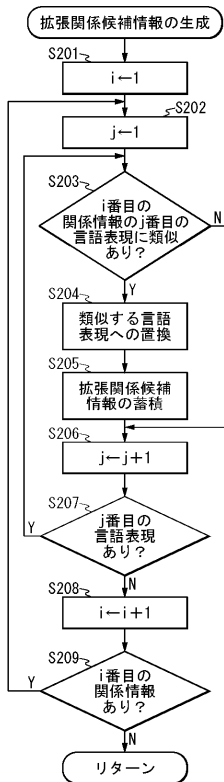
【図1】



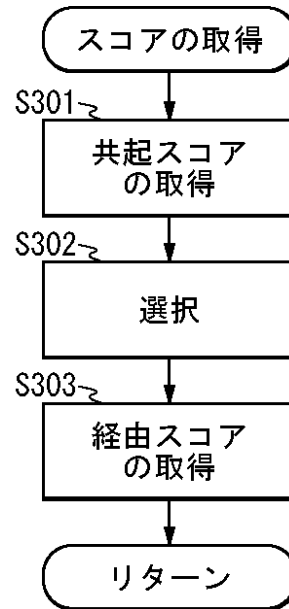
【図2】



【図3】



【図4】



【図5】

類似情報	
心筋梗塞, 脳梗塞, 脳卒中, うつ病	
死亡, 病死, 急死	
⋮	

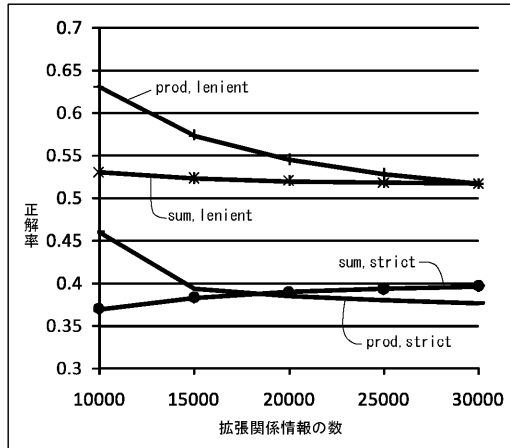
【図6】

関係情報ID	拡張関係候補情報	共起スコア
1	<脳梗塞, 死亡>	—
1	<脳卒中, 死亡>	—
1	<うつ病, 死亡>	—
1	<心筋梗塞, 病死>	—
1	<心筋梗塞, 急死>	—
1	<脳梗塞, 病死>	—
1	<脳梗塞, 急死>	—
1	<脳卒中, 病死>	—
1	<脳卒中, 急死>	—
1	<うつ病, 病死>	—
1	<うつ病, 急死>	—

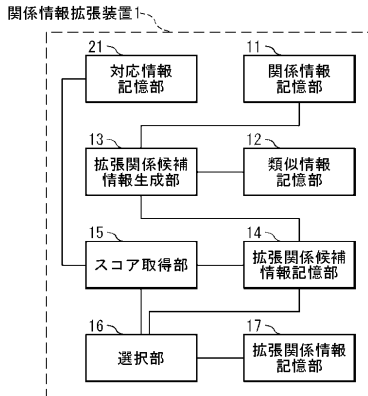
【図7】

関係情報ID	拡張関係候補情報	共起スコア
1	<脳梗塞, 死亡>	540
1	<脳卒中, 死亡>	190
1	<うつ病, 死亡>	280
1	<心筋梗塞, 病死>	220
1	<心筋梗塞, 急死>	360
1	<脳梗塞, 病死>	410
1	<脳梗塞, 急死>	470
1	<脳卒中, 病死>	120
1	<脳卒中, 急死>	150
1	<うつ病, 病死>	20
1	<うつ病, 急死>	16

【図8】



【図9】



【図11】

種類識別情報	類似情報
名産	愛媛, 香川
名産	香川, 徳島
名産	徳島, 高知
⋮	⋮
原因	死亡, 急死
⋮	⋮

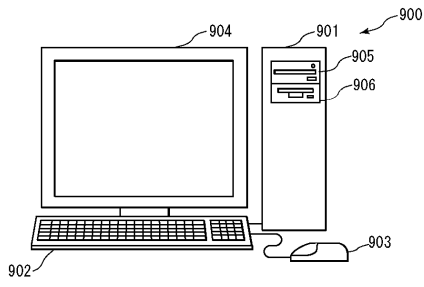
【図10】

対応情報	
種類識別情報	共起言語表現
原因	原因, 理由, 要因, . . .
食材	材料, レシピ, . . .
⋮	⋮

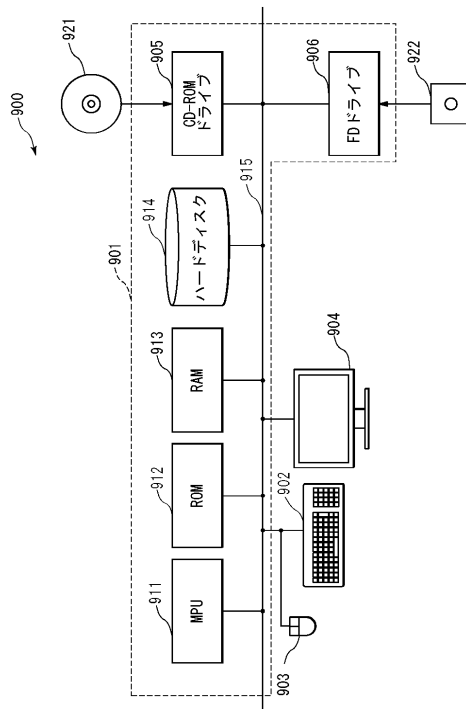
【図12】

種類識別情報	置換対象でない言語表現	類似情報
原因	死亡	心筋梗塞, 脳梗塞
原因	死亡	心筋梗塞, 脳卒中
⋮	⋮	⋮
名産	みかん	愛媛, 香川
⋮	⋮	⋮

【図14】

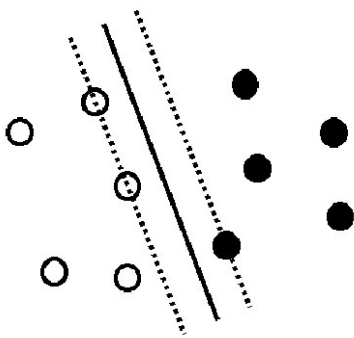


【図15】

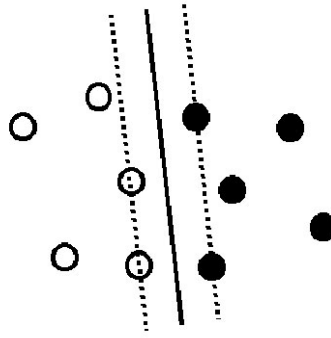


【図13】

(A) スモールマージン



(B) ラージマージン



フロントページの続き

- (72)発明者 鳥澤 健太郎
東京都小金井市貫井北町4 - 2 - 1 独立行政法人情報通信研究機構内
- (72)発明者 村田 真樹
東京都小金井市貫井北町4 - 2 - 1 独立行政法人情報通信研究機構内
- (72)発明者 風間 淳一
東京都小金井市貫井北町4 - 2 - 1 独立行政法人情報通信研究機構内
- (72)発明者 黒田 航
東京都小金井市貫井北町4 - 2 - 1 独立行政法人情報通信研究機構内

審査官 成瀬 博之

- (56)参考文献 特開2008 - 225565 (JP, A)
石川大介 他2名, 特許文献における因果関係を用いた類推による仮説の生成と検証, 情報知識
学会誌, 日本, 情報知識学会, 2007年 9月18日, 第17巻3号, 164 - 181頁

- (58)調査した分野(Int.Cl., DB名)
G06F 17/27 - 17/28