

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5640258号
(P5640258)

(45) 発行日 平成26年12月17日(2014.12.17)

(24) 登録日 平成26年11月7日(2014.11.7)

| | | | | | |
|-------------------|------------------|-----|------------|------|--|
| (51) Int.Cl. | | F I | | | |
| G06F 19/00 | (2011.01) | | G06F 19/00 | 130 | |
| G06F 17/30 | (2006.01) | | G06F 17/30 | 180C | |
| G06F 17/27 | (2006.01) | | G06F 17/27 | Z | |

請求項の数 8 (全 15 頁)

| | |
|--|---|
| <p>(21) 出願番号 特願2010-83666 (P2010-83666)</p> <p>(22) 出願日 平成22年3月31日 (2010.3.31)</p> <p>(65) 公開番号 特開2011-215898 (P2011-215898A)</p> <p>(43) 公開日 平成23年10月27日 (2011.10.27)</p> <p>審査請求日 平成24年11月12日 (2012.11.12)</p> <p>特許法第30条第1項適用 平成22年3月19日 日本知能情報ファジィ学会発行の「第35回ファジィ・ワークショップ講演論文集」において発表</p> | <p>(73) 特許権者 801000027 学校法人明治大学 東京都千代田区神田駿河台1-1</p> <p>(74) 代理人 100092820 弁理士 伊丹 勝</p> <p>(72) 発明者 高木 友博 神奈川県川崎市多摩区東三田1-1-1 明治大学生田校舎内</p> <p>審査官 本郷 彰</p> |
|--|---|

最終頁に続く

(54) 【発明の名称】 類推方法、類推システム及び類推プログラム

(57) 【特許請求の範囲】

【請求項1】

複数の基底語 (base 1, base 2) の間の関係 R から、目標語 (target) との間で関係 R にある解 X を類推する構造写像理論に基づく類推方法であって、

類推に用いられる知識情報として形態素解析された複数の文が蓄積されたコーパスから、前記複数の基底語に基づき前記関係 R を抽出する関係抽出モジュールによって、前記複数の基底語が同時に出現する文を抽出する第1ステップと、

前記関係抽出モジュールによって前記抽出された文に含まれる単語から前記複数の基底語の間の関係 R を表す単語 r_i を抽出する第2ステップと、

前記関係抽出モジュールによって前記抽出された単語 r_i について、前記関係 R への所属度 $grade R (r_i)$ を算出する第3ステップと、

前記関係抽出モジュールにより抽出された関係 R を前記目標語に写像する関係写像モジュールによって、前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出する第4ステップと、

前記関係写像モジュールによって前記抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出する第5ステップと、

前記第4ステップ及び前記第5ステップを前記抽出された全ての単語 r_i について行い、これにより抽出された全ての単語 x_j に対して前記関係写像モジュールにより前記解 X への所属度 $grade X (x_j)$ を算出する第6ステップと、

前記算出された所属度 $grade X (x_j)$ の値が高い所定数の単語 x_j を前記解 X に

10

20

含まれる前記目標語に關係する候補語として前記關係写像モジュールにより抽出する第7ステップと

を備え、

前記第5ステップでは、前記単語 x_j は前記複数の基底語及び前記単語 r_i の前記文における記載順序に基づき抽出される

ことを特徴とする類推方法。

【請求項2】

前記第4ステップでは、前記所属度 $gradeR(r_i)$ の値が所定値よりも高い単語 r_i と前記目標語とが同時に出現する文のみを抽出する

ことを特徴とする請求項1記載の類推方法。

10

【請求項3】

前記關係 R は、この關係 R を構成する単語を r_i とし、メンバーシップ関数の値を $gradeR(r_i)$ とした場合、

$$R = \{gradeR(r_i)/r_i\} \quad (i=1,2,\dots,n)$$

として表される概念ファジィ集合であり、

前記解 X は、この解 X を構成する単語を x_j とし、メンバーシップ関数の値を $gradeX(x_j)$ とした場合、

20

$$X = \{gradeX(x_j)/x_j\} \quad (j=1,2,\dots,n)$$

として表される概念ファジィ集合である

ことを特徴とする請求項1又は2記載の類推方法。

【請求項4】

前記所属度 $gradeR(r_i)$ の値は、

$count(r_i)$ を關係 R として抽出された単語 r_i に対する頻度、 $N(base1 \cap base2)$ を複数の基底語が同時に出現した文の数、 $N(r_i)$ をコーパスに含まれる文中に単語 r_i が現れる文の数、 $Ndoc$ をコーパスに含まれる全ての文の数とした場合、

30

$$\begin{aligned} gradeR(r_i) &= \alpha \cdot \log \left(1 + \frac{\beta \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right) \\ &= \log \left(1 + \frac{count(r_i) \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right) \cdot \log_{10} (1 + count(r_i)) \end{aligned}$$

として表される(ただし、 α 、 β は調整可能なパラメータ)

ことを特徴とする請求項1～3のいずれか1項記載の類推方法。

40

【請求項5】

前記所属度 $gradeX(x_j)$ の値は、

$gradeX(x_j | r_i)$ を単語 r_i が目標語に写像されたときに単語 x_j が解 X として正しいかを示す指数、 Nr を目標語に代表される単語 r_i の数とした場合、

$$gradeX(x_j) = \sum_{i=1}^{Nr} gradeX(x_j | r_i)$$

として表される

ことを特徴とする請求項1～4のいずれか1項記載の類推方法。

50

【請求項6】

前記 $gradeX(x_j | r_i)$ の値は、

$count(x_j | r_i)$ を単語 r_i が目標語に写像されたときの解 X の候補として抽出される頻度、 $N(r_i \cap target)$ をコーパスに含まれる単語 r_i と目標語とが同時に出現した文の数、 $N(x_j)$ をコーパスに含まれる文中に単語 x_j が現れる文の数、 $Ndoc$ をコーパスに含まれる全ての文の数とした場合、

$$gradeX(x_j | r_i) = \gamma \cdot \log \left(1 + \frac{\delta \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot gradeR(r_i)$$

$$= \log \left(1 + \frac{count(x_j | r_i) \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot \log(1 + count(x_j | r_i)) \cdot gradeR(r_i)$$

10

として表される（ただし、 δ は調整可能なパラメータ）

ことを特徴とする請求項5記載の類推方法。

【請求項7】

複数の基底語 ($base1, base2$) の間の関係 R から、目標語 ($target$) との間で関係 R にある解 X を類推する構造写像理論に基づく類推方式を用いた類推システムであって、

類推に用いられる知識情報として形態素解析された複数の文をコーパスとして蓄積する蓄積手段と、

20

前記コーパスから前記複数の基底語が同時に出現する文を抽出すると共に、抽出された文に含まれる単語から前記複数の基底語の間の関係 R を表す単語 r_i を抽出する第1抽出手段と、

前記抽出された単語 r_i について、前記関係 R への所属度 $gradeR(r_i)$ を算出する第1算出手段と、

前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出すると共に、抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出することを、前記抽出された全ての単語 r_i について行う第2抽出手段と、

前記抽出された全ての単語 x_j に対して前記解 X への所属度 $gradeX(x_j)$ を算出する第2算出手段と、

30

前記算出された所属度 $gradeX(x_j)$ の値が高い所定数の単語 x_j を前記解 X に含まれる前記目標語に関係する候補語として抽出する第3抽出手段と

を備え、

前記第2抽出手段は、前記単語 x_j を前記複数の基底語及び前記単語 r_i の前記文における記載順序に基づき抽出する

ことを特徴とする類推システム。

【請求項8】

複数の基底語 ($base1, base2$) の間の関係 R から、目標語 ($target$) との間で関係 R にある解 X を類推する構造写像理論に基づく類推方式をコンピュータに実行させる類推プログラムであって、

40

コンピュータに、

類推に用いられる知識情報として形態素解析された複数の文が蓄積されたコーパスから、前記複数の基底語が同時に出現する文を抽出させる第1処理と、

前記抽出された文に含まれる単語から前記複数の基底語の間の関係 R を表す単語 r_i を抽出させる第2処理と、

前記抽出された単語 r_i について、前記関係 R への所属度 $gradeR(r_i)$ を算出させる第3処理と、

前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出させる第4処理と、

50

前記抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出させる第 5 処理と、

前記第 4 処理及び前記第 5 処理を前記抽出された全ての単語 r_i について行い、これにより抽出された全ての単語 x_j に対して前記解 X への所属度 $grade X(x_j)$ を算出させる第 6 処理と、

前記算出された所属度 $grade X(x_j)$ の値が高い所定数の単語 x_j を前記解 X に含まれる前記目標語に関係する候補語として抽出させる第 7 処理と

を実行させ、

前記第 5 処理では、前記単語 x_j は前記複数の基底語及び前記単語 r_i の前記文における記載順序に基づき抽出される

10

ことを特徴とする類推プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

この発明は、構造写像理論に基づく類推方式を用いた類推方法、類推システム及び類推プログラムに関する。

【背景技術】

【0002】

従来より、類推の最も基本的な形として、 $A : B = C : X$ という 4 項類推が知られている。4 項類推は、例えば A と B 間の関係を推理し、基底領域（ベース）から目標領域（ターゲット）へ関係を写像し、推理された関係を C と X に適用するというようなアプローチで行われ、その処理過程はベースからターゲットへと関係を写像することが中心となる（非特許文献 1 及び 2）。なお、ここでベースとは、類推する際に用いる既存の知識のことであり、ターゲットとは、解決しなければならない未知の問題を指す。

20

【0003】

また、関係の写像においては、構造写像理論が知られている（非特許文献 3）。この構造写像理論では、写像の際に膨大に生じてしまう無意味な知識を、構造を利用することで排除するアプローチが取られている。この理論によれば、妥当な類推写像は、属性の非写像、構造の一貫性、システム性原理の 3 つの基準を満たすとされている。

【先行技術文献】

30

【非特許文献】

【0004】

【非特許文献 1】R.J.Sternberg, "Intelligence, Information Processing and Analogical Reasoning", Lawrence Erlbaum Associates, 1977

【非特許文献 2】R.J.Sternberg, "Component Processes in Analogical Reasoning", Psychological Review, 1977

【非特許文献 3】D.Gentner, "Structure-Mapping: A Theoretical Framework for Analogy", Cognitive Science, 1983

【発明の概要】

【発明が解決しようとする課題】

40

【0005】

この発明は、構造写像理論に基づく類推方式によってベースとターゲットが異なる概念に属している場合でも精度良くある程度正しい解を得ることができる類推方法、類推システム及び類推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】

【0006】

本発明に係る類推方法は、複数の基底語（base 1, base 2）の間の関係 R から、目標語（target）との間で関係 R にある解 X を類推する構造写像理論に基づく類推方法であって、類推に用いられる知識情報として形態素解析された複数の文が蓄積されたコーパスから、前記複数の基底語が同時に出現する文を抽出する第 1 ステップと、前記

50

抽出された文に含まれる単語から前記複数の基底語の関係 R を表す単語 r_i を抽出する第2ステップと、前記抽出された単語 r_i について、前記関係 R への所属度 $gradeR(r_i)$ を算出する第3ステップと、前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出する第4ステップと、前記抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出する第5ステップと、前記第4ステップ及び前記第5ステップを前記抽出された全ての単語 r_i について行い、これにより抽出された全ての単語 x_j に対して前記解 X への所属度 $gradeX(x_j)$ を算出する第6ステップと、前記算出された所属度 $gradeX(x_j)$ の値が高い所定数の単語 x_j を前記解 X に含まれる前記目標語に関係する候補語として抽出する第7ステップとを備えたことを特徴とする。

10

【0007】

好ましい実施形態においては、例えば前記第4ステップでは、前記所属度 $gradeR(r_i)$ の値が所定値よりも高い単語 r_i と前記目標語とが同時に出現する文のみを抽出し、例えば前記第5ステップでは、前記単語 x_j は前記複数の基底語及び前記単語 r_i の前記文における記載順序に基づき抽出される。

【0008】

また、前記関係 R は、例えばこの関係 R を構成する単語を r_i とし、メンバーシップ関数の値を $gradeR(r_i)$ とした場合、

$$R = \{gradeR(r_i) / r_i\} \quad (i=1,2,\dots,n)$$

20

として表される概念ファジィ集合であり、前記解 X は、例えばこの解 X を構成する単語を x_j とし、メンバーシップ関数の値を $gradeX(x_j)$ とした場合、

$$X = \{gradeX(x_j) / x_j\} \quad (j=1,2,\dots,n)$$

として表される概念ファジィ集合である。

【0009】

また、前記所属度 $gradeR(r_i)$ の値は、例えば $count(r_i)$ を関係 R として抽出された単語 r_i に対する頻度、 $N(base1 \cap base2)$ を複数の基底語が同時に出現した文の数、 $N(r_i)$ をコーパスに含まれる文中に単語 r_i が現れる文の数、 $Ndoc$ をコーパスに含まれる全ての文の数とした場合、

$$\begin{aligned} gradeR(r_i) &= \alpha \cdot \log \left(1 + \frac{\beta \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right) \\ &= \log \left(1 + \frac{count(r_i) \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right) \cdot \log_{10} (1 + count(r_i)) \end{aligned}$$

40

として表される。ただし、 α 、 β は調整可能なパラメータである。

【0010】

更に、前記所属度 $gradeX(x_j)$ の値は、例えば $gradeX(x_j | r_i)$ を単語 r_i が目標語に写像されたときに単語 x_j が解 X として正しいかを示す指数、 Nr を目標語に代表される単語 r_i の数とした場合、

$$gradeX(x_j) = \sum_{i=1}^{Nr} gradeX(x_j | r_i)$$

として表される。

50

【0011】

なお、前記 $gradeX(x_j | r_i)$ の値は、例えば $count(x_j | r_i)$ を単語 r_i が目標語に写像されたときの解 X の候補として抽出される頻度、 $N(r_i \cap target)$ をコーパスに含まれる単語 r_i と目標語とが同時に出現した文の数、 $N(x_j)$ をコーパスに含まれる文中に単語 x_j が現れる文の数、 $Ndoc$ をコーパスに含まれる全ての文の数とした場合、

$$gradeX(x_j | r_i) = \gamma \cdot \log \left(1 + \frac{\delta \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot gradeR(r_i)$$

$$= \log \left(1 + \frac{count(x_j | r_i) \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot \log(1 + count(x_j | r_i)) \cdot gradeR(r_i)$$

10

として表される。ただし、 δ は調整可能なパラメータである。

【0012】

本発明に係る類推システムは、複数の基底語 ($base1$, $base2$) の間の関係 R から、目標語 ($target$) との関係 R にある解 X を類推する構造写像理論に基づく類推方式を用いた類推システムであって、類推に用いられる知識情報として形態素解析された複数の文をコーパスとして蓄積する蓄積手段と、前記コーパスから前記複数の基底語が同時に出現する文を抽出すると共に、抽出された文に含まれる単語から前記複数の基底語の間の関係 R を表す単語 r_i を抽出する第1抽出手段と、前記抽出された単語 r_i について、前記関係 R への所属度 $gradeR(r_i)$ を算出する第1算出手段と、前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出すると共に、抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出することを、前記抽出された全ての単語 r_i について行う第2抽出手段と、前記抽出された全ての単語 x_j に対して前記解 X への所属度 $gradeX(x_j)$ を算出する第2算出手段と、前記算出された所属度 $gradeX(x_j)$ の値が高い所定数の単語 x_j を前記解 X に含まれる前記目標語に関係する候補語として抽出する第3抽出手段とを備えたことを特徴とする。

20

【0013】

本発明に係る類推プログラムは、複数の基底語 ($base1$, $base2$) の間の関係 R から、目標語 ($target$) との間で関係 R にある解 X を類推する構造写像理論に基づく類推方式をコンピュータに実行させる類推プログラムであって、コンピュータに、類推に用いられる知識情報として形態素解析された複数の文が蓄積されたコーパスから、前記複数の基底語が同時に出現する文を抽出させる第1処理と、前記抽出された文に含まれる単語から前記複数の基底語の間の関係 R を表す単語 r_i を抽出させる第2処理と、前記抽出された単語 r_i について、前記関係 R への所属度 $gradeR(r_i)$ を算出させる第3処理と、前記目標語と前記単語 r_i とが同時に出現する文を前記コーパスから抽出させる第4処理と、前記抽出された文に含まれる単語の中から、前記目標語との間で関係 R にある単語 x_j を抽出させる第5処理と、前記第4処理及び前記第5処理を前記抽出された全ての単語 r_i について行い、これにより抽出された全ての単語 x_j に対して前記解 X への所属度 $gradeX(x_j)$ を算出させる第6処理と、前記算出された所属度 $gradeX(x_j)$ の値が高い所定数の単語 x_j を前記解 X に含まれる前記目標語に関係する候補語として抽出させる第7処理とを実行させることを特徴とする。

30

40

【発明の効果】

【0014】

本発明によれば、構造写像理論に基づく類推方式によってベースとターゲットが異なる概念に属している場合でも精度良くある程度正しい解を得ることができる類推方法、類推システム及び類推プログラムを提供することができる。

【図面の簡単な説明】

50

【0015】

【図1】本発明の一実施形態に係る類推システムの全体概要を説明するための図である。

【図2】同類推システムにおける類推方法による類推処理手順を示すフローチャートである。

【図3】同類推システムにおける類推方法による類推処理手順の一例を説明するための図である。

【発明を実施するための形態】

【0016】

以下に、添付の図面を参照して、この発明に係る類推方法、類推システム及び類推プログラムの実施の形態を詳細に説明する。図1は、本発明の一実施形態に係る類推システムの全体概要を説明するための図である。本実施形態に係る類推システムでは、構造写像理論に基づく類推方式として類推のベースとターゲットとが明確で、最も簡潔なモデルである4項類推を例に挙げて説明する。

10

【0017】

ここで、4項類推とは、 $A : B = C : D$ という形式の問題を指す。この4項類推の左辺第一項Aを複数の基底語のうちの一つであるベース1 (base 1) とし、左辺第二項Bを複数の基底語のうち他の一つであるベース2 (base 2) とする。また、右辺第一項Cを目標語であるターゲット (target) とし、右辺第二項Dを類推により求めるべき解Xとする。そして、写像対象となるベース1とベース2との関係をRとする。

【0018】

20

図1に示すように、コーパス10は、類推に用いられる知識情報として形態素解析された複数の文が蓄積されたデータベース(DB)である。関係抽出モジュール(Extract Relation Module:ERM)20は、ベース1及びベース2に基づいて、関係Rを抽出するモジュールである。関係写像モジュール(Mapping Relation Module:MRM)30は、ERM20により抽出された関係Rをターゲットに写像するモジュールである。

【0019】

なお、このように構成された類推システムは、例えばパーソナルコンピュータやワークステーション等のハードウェア上で本発明に係る類推プログラムを実行することにより実現され、類推システムに対して入力されたベース1, 2及びターゲットから解Xを類推して出力するように機能する。パーソナルコンピュータやワークステーション等のハードウェア構成については公知であるため、ここでは説明を省略する。

30

【0020】

具体的には、図2に示すように、まず、コーパス10から入力されたベース1, 2が同時に出現する文11(図1参照)を全て抽出する(ステップS100)。例えば、図3に示すように、“fish:scale=bird:X”という問題が与えられた場合、図3中矢印(1)で示すように、ベース1, 2である「fish」、「scale」が同時に出現する文「Fish is covered with scale.」や「Fish has scales.」等をコーパス10から全て抽出する。

【0021】

40

次に、抽出された文に含まれる単語からベース1, 2により写像される概念ファジィ集合の要素となる関係Rを表す単語 r_i を抽出する(ステップS102)。例えば、図3中矢印(2)で示すように、「fish」、「scale」の関係Rを表す単語「is」、「has(have)」、「with」、「cover」等の単語 r_i を抽出された文から抽出する。

【0022】

関係Rは、この関係Rを構成する単語を r_i とし、メンバーシップ関数の値を $grade_R(r_i)$ とした場合、次式(1)として表される。

【0023】

【数1】

$$R = \{gradeR(r_i)/r_i\} \quad (i=1,2,\dots,n)$$

・・・(1)

【0024】

そして、抽出された全ての単語 r_i について、関係 R への所属度 $gradeR(r_i)$ を算出する(ステップS104)。この所属度 $gradeR(r_i)$ の値は、 $count(r_i)$ を関係 R として抽出された単語 r_i に対する頻度、 $N(base1 \text{ base} 2)$ を複数のベース1, 2が同時に出現した文の数、 $N(r_i)$ をコーパス10に含まれる文中に単語 r_i が現れる文の数、 $Ndoc$ をコーパス10に含まれる全ての文の数とした場合、次式(2)として表される。

【0025】

【数2】

$$gradeR(r_i) = \alpha \cdot \log \left(1 + \frac{\beta \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right)$$

・・・(2)

ここで、 α 、 β について調整し、書き換えると次のように表すことができる。

$$gradeR(r_i) = \log \left(1 + \frac{count(r_i) \cdot Ndoc}{N(base1 \cap base2) \cdot N(r_i)} \right) \cdot \log_{10}(1 + count(r_i))$$

【0026】

次に、入力されたターゲットに関係 R を写像して、ターゲットと単語 r_i とが同時に出現する文12(図1参照)をコーパス10から全て抽出する(ステップS106)。例えば、図3中矢印(3)で示すように、ターゲットである「bird」と単語 r_i である「is」、「has(have)」、「with」、「cover」とが同時に出現する文「Bird is covered with feather.」や「Bird has wing.」等をコーパス10から全て抽出する。

【0027】

そして、抽出された文に含まれる単語の中から、関係 R に基づく概念ファジィ集合を構成する解 X の候補となる単語 x_j を抽出する(ステップS108)。例えば、図3中矢印(4)で示すように、「bird」と関係 R のような概念ファジィ集合を構成する解 X の候補となるような単語「wing」、「feather」等の単語 x_j を抽出する。

【0028】

解 X は、この解 X を構成する単語を x_j とし、メンバーシップ関数の値を $gradeX(x_j)$ とした場合、次式(3)として表される。

【0029】

【数3】

$$X = \{gradeX(x_j)/x_j\} \quad (j=1,2,\dots,n)$$

・・・(3)

【0030】

その後、抽出された全ての単語 x_j についてステップS106及びステップS108の処理が行われたか否かを判断し(ステップS110)、行われていない場合(ステップS110のN)は上記ステップS106に移行して処理を繰り返すと共に、行われた場合(ステップS110のY)は、抽出された全ての単語 x_j に対して解 X への所属度 $gradeX(x_j)$ を算出する(ステップS112)。

10

20

30

40

50

【 0 0 3 1 】

この所属度 $gradeX(x_j)$ の値は、 $gradeX(x_j | r_i)$ を単語 r_i が目標語に写像されたときに単語 x_j が解 X として正しいかを示す指数、 N_r を目標語に代表される単語 r_i の数とした場合、次式 (4) として表される。

【 0 0 3 2 】

【 数 4 】

$$gradeX(x_j) = \sum_{i=1}^{N_r} gradeX(x_j | r_i)$$

10

・・・ (4)

【 0 0 3 3 】

最後に、例えば算出された所属度 $gradeX(x_j)$ の値が高い所定数の単語 x_j を解 X に含まれるターゲットに関係する候補語として抽出し (ステップ S 1 1 4)、本フローチャートによる処理が終了される。なお、 $gradeX(x_j | r_i)$ の値は、 $count(x_j | r_i)$ を単語 r_i がターゲットに写像されたときの解 X の候補として抽出される頻度、 $N(r_i \cap target)$ をコーパス 10 に含まれる単語 r_i とターゲットとが同時に出現した文の数、 $N(x_j)$ をコーパス 10 に含まれる文中に単語 x_j が現れる文の数、 $Ndoc$ をコーパス 10 に含まれる全ての文の数とした場合、次式 (5) として表される。

20

【 0 0 3 4 】

【 数 5 】

$$gradeX(x_j | r_i) = \gamma \cdot \log \left(1 + \frac{\delta \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot gradeR(r_i)$$

・・・ (5)

ここで、 δ について調整し、書き換えると次のように表すことができる。

$$gradeX(x_j | r_i) = \log \left(1 + \frac{count(x_j | r_i) \cdot Ndoc}{N(r_i \cap target) \cdot N(x_j)} \right) \cdot \log(1 + count(x_j | r_i)) \cdot gradeR(r_i)$$

30

【 0 0 3 5 】

ここで、上記ステップ S 1 0 6 では、所属度 $gradeR(r_i)$ の値が高い単語 r_i についての関係 R のみをターゲットに写像するようにしても良い。このようにすれば、最終的に解 X の候補語として不適当な単語が抽出されることを避けることが可能となる。

【 0 0 3 6 】

また、解 X の正しい答えの候補語を得るために、 $A : B = C : X$ と $B : A = C : X$ の違いは考慮される必要があるので、上記ステップ S 1 0 8 では、例えば単語 x_j はベース 1、2 及び単語 r_i の抽出された文における記載順序に基づき抽出するようにしても良い。この場合、単語 x_j は、次の規則の下に抽出される。

【 0 0 3 7 】

すなわち、(1) 単語 r_i が記載順序「ベース 1 単語 r_i ベース 2」の関係下で頻繁に抽出された場合は、単語 r_i の後に記載された単語が単語 x_j として抽出される。また、(2) 単語 r_i が記載順序「ベース 2 単語 r_i ベース 1」の関係下で頻繁に抽出された場合は、単語 r_i の前に記載された単語が単語 x_j として抽出される。更に、(3) 単語 r_i が上記 (1) 及び (2) の関係下でそれぞれ等しく抽出された場合は、単語 x_j は単語 r_i の前後いずれかに記載された単語として抽出される。

40

【 0 0 3 8 】

本実施形態に係る類推システムでは、このような処理により、ベース 1、2 とターゲットとが異なる概念に属している場合でも、ある程度正しい解 X の候補を精度良く得ることが可能となる。また、得られた解 X を類推による候補として類推システムに備えられた図示しない表示手段に表示したり、印刷手段で印刷出力したり、音声出力手段で報知したり

50

して利用者に提示するようにすれば、類推による推薦を行うことも可能となる。次に、上述した類推システムを用いた本出願人による試験について説明する。

【0039】

この試験においては、フリー百科辞典の英語版 Wikipedia に書かれたある時点の全ての文（約 3,691,000 項目、約 43,670,000 文）を蓄積したコーパス 10 を用いた。この Wikipedia に書かれた文をコーパス 10 に記憶する際には、フリーソフトウェアの Tree Tagger を用いて形態素解析を行った。また、コーパス 10 の作成と検索にはフリーソフトウェアの Lucene を使用した。

【0040】

そして、試験では、単語で表記され、且つベース 1、ベース 2、ターゲット、解 X が全て 1 単語の名詞である問題のみを対象にした。更に、2 つの名詞の関係としては動詞が適切である場合が多いため、関係 R は動詞のみとした。なお、4 項類推では絶対的な正解を規定することが困難である。これは 4 項類推が、回答者の持つ知識や主観に少なからず依存するためである。そこで、次のように類推システムの評価を行うこととした。

【0041】

まず、参加した評価者はそれぞれ 4 項類推の解 X を求め、次に、同じ 4 項類推について類推システムが解 X を求めてその候補に当たる上位 10 個を評価者に提示した。評価者は提示された解 X を基に、類推システムが行った類推がどの程度正しいかを 2, 1, 0 の 3 段階で評価した。このような評価を 10 人の評価者が、合計 20 題の問題に対して行うことで、類推システムの評価を行った。

【0042】

以下の表 1 に「earth : sun = moon : X」という 4 項類推の問題に対して類推システムが求めた関係 R と解 X の候補上位 10 個を示す。

【0043】

【表 1】

“earth : sun = moon : X” の類推結果

| 関係 | gradeR(r_i) | 解 | gradeX(x_j) |
|------------|-----------------|-------------|-----------------|
| revolve | 1.0000 | earth | 1.0000 |
| orbit | 0.8511 | sun | 0.8339 |
| rotate | 0.4370 | planet | 0.6362 |
| eclipse | 0.2800 | orbit | 0.5570 |
| illuminate | 0.2620 | jupiter | 0.4738 |
| shine | 0.2441 | saturn | 0.3829 |
| absorb | 0.2376 | ecliptic | 0.3145 |
| envelop | 0.2339 | spacecraft | 0.2929 |
| radiate | 0.2249 | inclination | 0.2860 |
| obscure | 0.2062 | eclipse | 0.2464 |

【0044】

問題「earth : sun = moon : X」は、評価者全員が、類推システムが求めた解 X を正しいと評価した問題である。上記表 1 を見ると、「earth」と「sun」の関係として「revolve」や「orbit」等が挙げられている。また、解 X の候補

1位に「earth」という「earth:sun=moon:X」の解として納得できる単語が挙げられている。

【0045】

この結果から、類推システムは、「earth:sun=moon:X」という問題に関して、「earth」と「sun」の関係Rを正しく理解し、その関係Rを基に正しい解Xを導くことに成功したと言える。

【0046】

また、以下の表2は、「fish:scale=bird:X」という問題と、「fish:fin=bird:X」という問題とについて類推システムが求めた解を比較したものである。この2つの問題は、与えられた3項のうち、ベース2のみが異なる。

【0047】

【表2】

base2 の変化による類推結果の相違

| fish : scale = bird : X | |
|-------------------------|-------------------------|
| Xの要素 | gradeX(x _j) |
| insect | 1.0000 |
| feather | 0.8708 |
| prey | 0.8032 |
| nectar | 0.7349 |
| flycatcher | 0.6799 |
| nest | 0.6704 |
| buttonquails | 0.6445 |
| ground-nesting | 0.6165 |
| specie | 0.6116 |
| wing | 0.6039 |

| fish : fin = bird : X | |
|-----------------------|-------------------------|
| Xの要素 | gradeX(x _j) |
| feather | 1.0000 |
| wing | 0.7189 |
| beak | 0.6494 |
| tail | 0.5421 |
| plumage | 0.4391 |
| warbler | 0.4323 |
| specie | 0.4182 |
| insect | 0.4151 |
| nest | 0.4143 |
| bill | 0.4140 |

【0048】

上記表2から、類推システムは、2つの問題に対して適切に異なる解Xを求めたことが分かる。実際に、「fish:fin=bird:X」の解Xでは、「fish:scale=bird:X」の解Xよりも、「wing」や「tail」の順位が上がっている。これは、類推システムが、それぞれの問題に適した関係Rを求め、より解Xとして適した単語を導き出した結果であると言える。

【0049】

なお、表2に示した「fish:scale=bird:X」という問題で、類推システムが求めた解Xの候補1位に「insect」という単語が選ばれているが、これは、ERM20によって、「fish」と「scale」の関係として「feed」という単語が抽出されたためである。これは、正しい関係Rを得られないために、解Xに不適切な単語が含まれた例として挙げられる。

【0050】

最後に、10人の評価者による20問の問題に対する類推システムの評価を表3に示す。

【0051】

10

20

30

40

【表3】

20問の問題に対する類推結果とシステム評価

| 番号 | 問題 | 解の1位 | 解の2位 | 解の3位 | 評価の合計 |
|----|-----------------------------------|------------|-------------|-------------|-------|
| 1 | fish : scale = bird : ? | insect | feather | Prey | 10 |
| 2 | bird : feather = fish : ? | fin | water | fishing | 6 |
| 3 | fish : fin = bird : ? | feather | wing | beak | 17 |
| 4 | bird : sky = fish : ? | fishing | water | trout | 7 |
| 5 | earth : sun = moon : ? | earth | sun | planet | 20 |
| 6 | earth : sun = electron : ? | nucleus | atom | energy | 19 |
| 7 | river : water = circuit : ? | voltage | capacitor | device | 10 |
| 8 | teacher : student = doctor : ? | patient | tardis | hospital | 19 |
| 9 | sardine : fish = sparrow : ? | bird | seed | insect | 15 |
| 10 | frog : snake = mouse : ? | cat | lizard | snake | 20 |
| 11 | beer : barley = wine : ? | grape | beverage | winery | 20 |
| 12 | france : paris = spain : ? | madrid | france | toledo | 13 |
| 13 | spain : madrid = france : ? | french | paris | switzerland | 10 |
| 14 | america : washington = france : ? | french | switzerland | finland | 4 |
| 15 | animal : food = automobile : ? | vehicle | car | motor | 6 |
| 16 | automobile : gasoline = yacht : ? | boat | sailing | vessel | 4 |
| 17 | automobile : gasoline = train : ? | locomotive | passenger | rail | 2 |
| 18 | winter : snow = summer : ? | winter | olympic | snow | 1 |
| 19 | surplus : deficit = black : ? | hawk | hole | color | 2 |
| 20 | up : down = left : ? | hand | right | arm | 7 |

【0052】

上記表3では、20問の問題と、類推結果として類推システムが提示した解Xのうち所属度 $gradeR(x_j)$ の値が高い上位3つの要素、類推システムが求めた解Xについて10人の評価者が付けた評価点の合計を示している。また、この試験での評価の満点は20点となっている。

【0053】

表3に挙げられた結果を見ると、10点以上の評価を得た問題が11問あった。そのうち3問が満点の20点の評価を得ていた。また、5点以下の評価を得た問題が5問あった

10

20

30

40

50

。表3に示された評価の平均は10.6となった。

【0054】

そして、評価者全員が類推システムが求めた解Xが正しいとした3問の問題は、「earth:sun=moon:X」、「snake:frog=cat:X」、「beer:barley=wine:X」であり、それぞれ解Xの1位に「earth」、「mouse」、「grape」を挙げていた。

【0055】

逆に結果が悪かった問題は、「winter:snow=summer:X」という問題で、解Xの候補1位には「winter」が挙げられていた。

【0056】

このように、結果が良かった問題では、関係Rとして正しいと考えられるものが抽出されていた。逆に、結果が悪かった問題では、関係Rとして正しくないと考えられるものが抽出されていた。従って、類推システムでは、関係Rの抽出が類推の可否に大きく影響することが分かったが、本実施形態に係る類推システムによれば、ベースとターゲットが異なる概念に属している場合でも、概ね精度良くある程度正しい解Xを得ることが可能なが判明した。

【符号の説明】

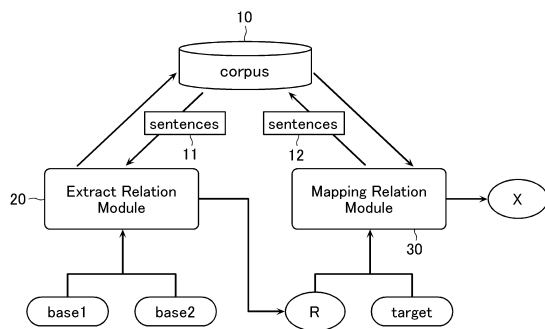
【0057】

- 10 コーパス
- 20 関係抽出モジュール (ERM)
- 30 関係写像モジュール (MRM)

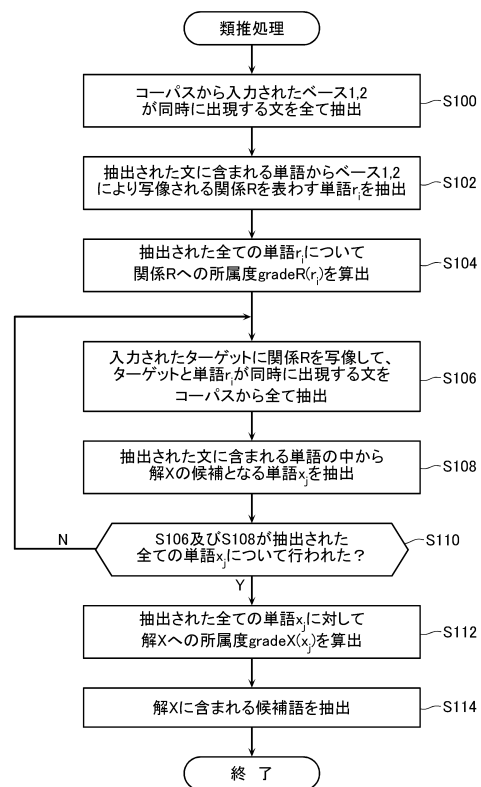
10

20

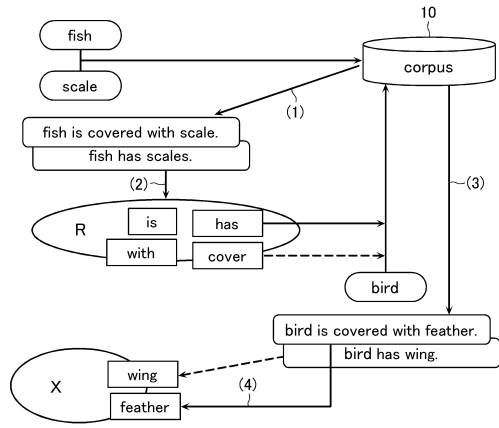
【図1】



【図2】



【 3 】



フロントページの続き

(56)参考文献 特開2009-059323(JP,A)

特開2011-118689(JP,A)

溝口 健二,萩原 将文,領域表現に基づいた四項類推を行なうニューラルネットワーク,電気学会論文誌C Vol.121-C No.7,日本,(社)電気学会,2001年7月1日,第121-C巻第7号【ISSN】0385-4221,1261-1267

加藤 誠,大島 裕明,小山 聡,田中 克己,語の共起を用いたWebからの類似関係検索,電子情報通信学会技術研究報告 Vol.108 No.329,日本,社団法人電子情報通信学会,2008年11月24日,第108巻第329号【ISSN】0913-5685,p.9-p.10

(58)調査した分野(Int.Cl.,DB名)

G06F 19/00

G06F 17/27-17/28

G06F 17/30