

日本語ワードネットを用いた意味の数学モデルの構築

斎藤 優太[†] 井上 尚^{††} 北川 高嗣^{†††}

[†] 筑波大学情報学群情報科学類 〒 305-8573 茨城県つくば市天王台 1-1-1

^{††} 筑波大学システム情報工学研究科コンピュータサイエンス専攻 〒 305-8573 茨城県つくば市天王台 1-1-1

^{†††} 筑波大学システム情報系 〒 305-8573 茨城県つくば市天王台 1-1-1

E-mail: [†]{yuta,inoue}@mma.cs.tsukuba.ac.jp, ^{††}takashi@cs.tsukuba.ac.jp

あらまし 我々は、計算機上で言葉と言葉の意味的相関の計量が可能なモデルとして「意味の数学モデル」を提案してきた。これまでは、モデル内の意味空間の作成に英英辞書を用いており英単語同士の意味的相関の計量のみ可能だった。そこで、本研究では日本語のシソーラスである日本語ワードネットを用いて意味空間を作成する手法を考案し実装した。さらに、日本語ワードネットで定義されている概念同士の上位下位、含意、などの関係を意味空間の作成に使用した。提案手法により日本語の単語同士の意味的相関の計量が可能になった。また、実験を行い提案手法の有効性を検証した。

キーワード 意味の数学モデル 日本語ワードネット 意味的相関計量

Yuta SAITO[†], Hisashi INOUE^{††}, and Takashi KITAGAWA^{†††}

[†] College of Information Science, University of Tsukuba. 305-8573 Japan

^{††} Department of Computer Science, Graduate school of SIE, University of Tsukuba. 305-8573 Japan

^{†††} Faculty of Engineering, Information and Systems, University of Tsukuba.

E-mail: [†]{yuta,inoue}@mma.cs.tsukuba.ac.jp, ^{††}takashi@cs.tsukuba.ac.jp

1. 序 論

我々は、人間は高度な知的活動を行うにあたり、様々な事象を言葉に変換し文脈によって言葉の意味を特定し思考しているという視点から、計算機による言葉と言葉の意味的相関の計量が可能になれば、計算機が人間の思考を理解し、人間の感性に基づいた高度な知的活動の支援ができるようになって考えている。

我々は計算機上で言葉と言葉の意味的相関の計量を可能にするモデルとして「意味の数学モデル」[1]を提案してきた。意味の数学モデルは、辞書をもとに言葉同士の意味的な関係を表す空間(以下、意味空間)を作成し、その空間を用いて言葉同士の意味的相関を計量するモデルである。各種メディアデータを言葉を用いて特徴づけ、意味の数学モデルに適応することで、メディアデータの形式によらず統一的に扱うことができる。

これまでは、モデル内の意味空間を Longman Dictionary of Contemporary English [2] という英英辞書を用いて作成していた。Longman 英英辞書は全ての見出し語を、The Longman Defining Vocabulary と呼ばれる約 2000 語の単語のみを用いて説明しているという特徴を持つ。以下、辞書の見出し語を説明している単語を説明語と呼ぶ。意味の数学モデル内の意味空間の次元

は、作成に使用した辞書の説明語の種類の数によって決まる。Longman 英英辞書を用いて従来の手法で作成した意味空間の次元はたかだか 2000 次元になる。

従来の意味の数学モデルは英単語同士の意味的相関のみ計量可能だった。他の言語を用いる場合はその都度英語に翻訳する必要があった。つまり、他の言語を英語に翻訳する際に、適切な単語や表現が英語に存在しない場合などでは、他の言語独特の単語の意味や表現を考慮できていなかった。また、説明語の種類数が限定されていない辞書を用いて従来の手法で意味空間を作成しようとした場合、計算コストが大きくなるという問題があった。

本研究で提案する手法では、日本語ワードネット [3][4] を辞書として用い意味空間を作成し、意味の数学モデルを構築する。また、類似する意味を持つ説明語を、日本語ワードネットを用いて一つのグループにまとめることで、各説明語の意味を保ちつつ、意味空間の作成に用いる説明語の種類数を減少する。これにより、意味空間の次元は、たかだか 6 万次元まで縮小され、計算コストを削減する。さらに、日本語ワードネットで定義されている概念同士の上位下位、含意などの関係を意味空間の作成に使用する。

そして、提案手法により構築した意味の数学モデルを用いて

意味的相関を計量し、検証を行った。

2. 先行研究

本章では、先行研究である意味の数学モデルと日本語ワードネットについて述べる。

2.1 意味の数学モデル

メディアデータを言葉を用いてメタ表現したデータ（以下、メタデータ）とする。本節では、辞書をもとに作られたメタデータ空間（以下、 MDS ）内に、メタデータを射影し、与えられた文脈を考慮した言葉と言葉の意味的相関を計量するモデルである意味の数学モデルについて述べる。

2.1.1 MDS の作成

(1) データ行列 M の作成

辞書などを用いて m 個の見出し語について、各々 n 個の特徴 $(f_1, f_2, \dots, f_n)$ で特徴づけられた特徴ベクトル $(d_1, d_2, \dots, d_m)$ を作成し、そのベクトルを行ベクトルとした $m \times n$ 行列をデータ行列 M とする。データ行列 M は列ごとに 2 ノルムで正規化する。図 1 に、データ行列 M の概形を示す。

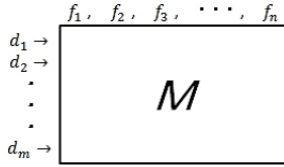


図 1 データ行列 M

(2) データ行列 M の相関行列 $M^T M$ の固有値分解

データ行列 M の相関行列 $M^T M$ を固有値分解する。

$$M^T M = Q \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_\nu & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} Q^T, \quad (1)$$

$$0 \leq \nu \leq n.$$

ここで行列 Q は、

$$Q = (q_1, q_2, \dots, q_n), \quad (2)$$

である。 $(q_1, q_2, \dots, q_n)$ は $M^T M$ の固有ベクトルである。相関行列の対称性より、固有値はすべて実数であり、固有ベクトルは互いに直交している。

(3) MDS の定義

非ゼロ固有値に対応する固有ベクトル $(q_1, q_2, \dots, q_\nu)$ によって形成される正規直交空間を MDS と定義する。

$$MDS := \text{span}(q_1, q_2, \dots, q_\nu). \quad (3)$$

$(q_1, q_2, \dots, q_\nu)$ は正規直交基底である。 MDS の次元はデータ行列 M のランクに一致する。この空間は ν 次元ユークリッド空間となる。また、 MDS の基底となる非ゼロ固有値に対応する固有ベクトル $(q_1, q_2, \dots, q_\nu)$ を意味素と呼ぶ。

2.1.2 メディアデータの MDS への写像

メディアデータのメタデータを 2.1.1 節で作成した MDS へ写像する。これにより、メタデータが同じ空間上に配置されることになり、メディアデータの種類の影響されことなくメディアデータ間の意味的相関を計算することが可能となる。

(1) メディアデータの特徴付け

メディアデータ P を t 個の特徴 $(o_1, o_2, \dots, o_t)$ を用いて (4) のように特徴づける。

$$P = (o_1, o_2, \dots, o_t). \quad (4)$$

ここで各特徴 $(o_1, o_2, \dots, o_t)$ はデータ行列 M と同じ特徴を用いて (5) のように表現されているベクトルである。

$$o_i = (o_{i1}, o_{i2}, \dots, o_{in}). \quad (5)$$

(2) メディアデータ P のベクトル表現

メディアデータ P を構成する t 個の特徴 $(o_1, o_2, \dots, o_t)$ がそれぞれ n 次元のベクトルで定義されている。 t 個の特徴 $(o_1, o_2, \dots, o_t)$ を (6) のように定義する結合演算子 $\oplus$ を用いて結合することにより、一つの n 次元ベクトルを生成する。このように生成した n 次元ベクトルをメタデータとする。

$$P = \oplus_{i=1}^t o_i := (\text{sign}(\max_{1 \leq i \leq t} |o_{i1}|) \max_{1 \leq i \leq t} o_{i1}, \text{sign}(\max_{1 \leq i \leq t} |o_{i2}|) \max_{1 \leq i \leq t} o_{i2}, \dots, \text{sign}(\max_{1 \leq i \leq t} |o_{in}|) \max_{1 \leq i \leq t} o_{in}). \quad (6)$$

結合演算子 $\oplus$ は t 個のベクトルから要素ごとに絶対値最大の成分を取り出す。ここで $\text{sign}(\alpha)$ は実数 α の符号を表す。以上のように作成したメタデータを MDS 内でフーリエ展開することにより、メタデータは MDS 内に写像される。

2.1.3 文脈により選択された部分空間上への射影

(1) 意味射影集合 Π_ν を構成

MDS から部分空間への射影（以下、意味射影）の集合を構成する。 P_{λ_i} を (7) のように定義する。

$$P_{\lambda_i} \xleftrightarrow{d} \lambda_i \text{ に対応する固有空間への射影,} \\ \text{i.e. } P_{\lambda_i} : MDS \rightarrow \text{span}(q_i). \quad (7)$$

意味射影の集合 Π_ν を (8) のように定義する。

$$\begin{aligned}
\Pi_\nu &:= \{0, P_{\lambda_1}, P_{\lambda_2}, \dots, P_{\lambda_\nu}, \\
&P_{\lambda_1} + P_{\lambda_2}, P_{\lambda_1} + P_{\lambda_3}, \dots, P_{\lambda_{\nu-1}} + P_{\lambda_\nu}, \\
&\vdots \\
&P_{\lambda_1} + P_{\lambda_2} + \dots + P_{\lambda_\nu}\}.
\end{aligned} \tag{8}$$

i 次元の部分空間を $\frac{\nu(\nu-1)\dots(\nu-i+1)}{i!}$ ($i = 1, 2, \dots, \nu$) 個作成できるので, Π_ν の要素の個数は 2^ν 個であり, 意味の数学モデルは 2^ν 通りの意味を表現できる.

(2) 意味解釈オペレータ S_p の構成

文脈を決定する長さ l の単語列 $s_l = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l)$ と, 閾値 ε ($0 < \varepsilon < 1$) が与えられた時, それらに応じて意味射影 $P_\varepsilon(s_l)$ を構成するオペレータが S_p である. T_l を長さ l の単語列の集合とすると S_p は以下のように定義される.

$$\begin{aligned}
S_p : T_l &\mapsto \Pi_\nu, \\
T_l \ni s_l, \Pi_\nu \ni P_\varepsilon(s_l).
\end{aligned} \tag{9}$$

ここで, $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l)$ はデータ行列 M と同一の特徴を用いて表されるベクトルである. 意味解釈オペレータ S_p は以下の計算を行う.

(a) $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l)$ のフーリエ展開

文脈を決定する単語列 s_l を構成する l 個の単語を各々 MDS に射影する. この射影では, l 個の単語を各々 MDS 内でフーリエ展開し, フーリエ係数を求める. これは各単語と各意味素の相関を求めることに相当する. $\mathbf{w}_i$ と $\mathbf{q}_j$ の内積 w_{ij} は (10) のように定める.

$$w_{ij} := (\mathbf{w}_i, \mathbf{q}_j), \quad j = 1, 2, \dots, \nu. \tag{10}$$

ベクトル $\hat{\mathbf{w}}_i \in MDS$ を (11) のように定める.

$$\hat{\mathbf{w}}_i := (w_{i1}, w_{i2}, \dots, w_{i\nu}). \tag{11}$$

$\hat{\mathbf{w}}_i$ は, 単語 $\mathbf{w}_i$ を MDS に射影したものである.

(b) 各単語列 s_l の意味重心 $\mathbf{G}^+(s_l)$ の計算

まず, 意味素ごとにフーリエ係数の総和を求める. これは, 単語列 s_l と各意味素との相関を求めることに相当する. このベクトルは, ν 個の意味素があるため ν 次元ベクトルとなる. このベクトルを無限大ノルムによって正規化したベクトルを, 単語列 s_l の意味重心 $\mathbf{G}^+(s_l)$ と呼び, (12) のように定める.

$$\mathbf{G}^+(s_l) := \frac{(\sum_{i=1}^l w_{i1}, \dots, \sum_{i=1}^l w_{i\nu})}{\|(\sum_{i=1}^l w_{i1}, \dots, \sum_{i=1}^l w_{i\nu})\|_\infty}. \tag{12}$$

(c) 意味射影 $P_\varepsilon(s_l)$ の決定

単語列 s_l の意味重心を構成する各要素において, 閾値 ε を越える要素に対応する意味素を, メタデータを射影する部分空間の構成に用いる $P_\varepsilon(s_l)$ を, (13) のように定める.

$$P_\varepsilon(s_l) := \sum_{i \in \Lambda_\varepsilon} P_{\lambda_i} \in \Pi_\nu. \tag{13}$$

ただし, $\Lambda_\varepsilon := \{i | |(G^+(s_l)_i)| > \varepsilon\}$ とする.

2.1.4 意味的相関計量

文脈により選択された部分空間上に射影されたメタデータのノルムを求めることにより, メディアデータの意味的相関の計量を行う. 部分空間におけるメタデータのノルムの大きさをその文脈とメディアデータとの意味的相関の大きさとする. メタデータ $\mathbf{x} \in MDS$ のノルム $norm(\mathbf{x})$ を (14) のように定める.

$$norm(\mathbf{x}) = \frac{\sqrt{\sum_{j \in \Lambda_\varepsilon} \{c_j x_j\}^2}}{\|\mathbf{x}\|_2}. \tag{14}$$

ここで, c_j は, 文脈を決定する単語列 s_l に依存して決まる重みであり, (15) のように定義する.

$$c_j := \frac{\sum_{i=1}^l w_{ij}}{\|(\sum_{i=1}^l w_{i1}, \dots, \sum_{i=1}^l w_{i\nu})\|_\infty}, \quad j \in \Lambda_\varepsilon. \tag{15}$$

ここで, メディアデータの特徴付けしている特徴の数が多いと, 文脈によりどのような部分空間が選ばれても, 部分空間におけるメディアデータのノルムが大きくなる傾向がある. そのため MDS でのメディアデータベクトルを 2 ノルムで正規化する.

また, メタデータ $\mathbf{x} \in MDS$ とメタデータ $\mathbf{y} \in MDS$ 間の文脈を考慮した動的距離 $\rho(\mathbf{x}, \mathbf{y})$ を (16) のように定める.

$$\rho(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j \in \Lambda_\varepsilon} \{c_j (x_j - y_j)\}^2}. \tag{16}$$

動的距離 $\rho(\mathbf{x}, \mathbf{y})$ の値が 0 に近いほど, メタデータ $\mathbf{x}$ とメタデータ $\mathbf{y}$ の与えられた文脈での意味的相関が大きいくことを表す.

2.2 日本語ワードネット

日本語ワードネットは, プリンストン大学の認知科学研究所によって運営されている英語の概念辞書である Princeton WordNet をベースに, 日本語を付与した日本語のシソーラスである. 日本語ワードネットは独立行政法人情報通信研究機構 (NICT) によって運営されている.

2.2.1 日本語ワードネットの構造について

日本語ワードネットでは, Princeton WordNet で定義されている synset に対して日本語が付加されている. synset とは, 語を類義関係のセットでグループ化したものであり, 一つの synset が一つの概念に対応する. 日本語ワードネット (ver.1.1) には synset として 57238 種類の概念が含まれている. また, synset 同士はリンクと呼ばれる関係で結ばれており, リンクには, 上位下位, 含意など 21 種類が定義されている.

Synset 02068974-n	1	1 synset番号(synset offset)
Jpn: 海豚, ドルフィン, イルカ	2	2 同義語(synonym)
Eng: dolphin		3 定義文
Jpn: ぐんばいのような鼻先を持つ様々な小型歯クジラ各種; マズミルよりも大きい		4 関連synsetとのリンク
Eng: any of various small toothed whales with a beaklike snout; larger than porpoises		5 他の言語資源とのリンク
Hype: toothed whale		
Hype: delphinus, delphis, white, whale, grampus, griseus, bottlenose, dolphin		
Hynd: whale, sea, orca, river, dolphin, porpoise		
Hynd: delphinidae	4	
SUMO: < AquaticMammal	5	

図2 日本語ワードネットの synset の内容

日本語ワードネットの synset の内容を, 図2を用いて説明する.

(1) synset 番号

synset 番号は各 synset に割り当てられた識別番号である。「8 ケタの数字+“(半角ハイフン)”+アルファベット」で表される。アルファベットは品詞を表しており、それぞれ n = 名詞, v = 動詞, r = 形容詞, a = 副詞 を表す。

(2) 同義語

同義語は、その synset の概念を言い表すのに使用される語である。英語の単語は Princeton WordNet で定義された語で、日本語の単語は日本語ワードネットが付加された語である。

(3) 定義文

定義文は、その synset が表す概念について説明している文である。英文は Princeton WordNet で定義された文で、日本語は日本語ワードネットが付加された文である。

(4) 関連 synset とのリンク

synset 同士の関係を表すリンクである。詳細は 2.2.2 節で説明する。

(5) 他の言語資源とのリンク

IEEE からリリースされている概念辞書である SUMO (Suggested Upper Merged Ontology) と synset のリンクである。

2.2.2 synset 同士の関係を表すリンク

各 synset は上位下位、含意などの 21 種類の関係を表すリンクで結ばれている。表 1 に日本語ワードネットで定義されているリンクの種類の一部を示す。また、表 2 にリンクの種類の具体例を示す。

表 1 リンクの定義

種類	定義
上位語	当該 synset が相手 synset に包含される
下位語	当該 synset が相手 synset を包含する
構成要素 (部分)	当該 synset が部分として相手 synset を構成する
被構成要素 (部分)	当該 synset が相手 synset という部分によって構成される
構成要素 (構成員)	当該 synset が相手 synset の構成員である
被構成要素 (構成員)	当該 synset が相手 synset という構成員によって構成される
構成要素 (物質・材料)	当該 synset が相手 synset を構成する物質、材料である
被構成要素 (物質・材料)	当該 synset が相手 synset という物質、材料によって構成される
被包含領域 (カテゴリ)	当該 synset が相手 synset のカテゴリに属する
例	当該 synset は相手 synset の例である
含意	当該 synset を行う場合、必ず相手 synset も行っている
属性	当該 synset が相手 synset という属性を表す際に使われる

表 2 リンクの例

種類	当該 synset	相手 synset
上位語	りんご	果物
下位語	果物	りんご
構成要素 (部分)	竜骨	船体
被構成要素 (部分)	船体	竜骨
構成要素 (構成員)	兄	兄弟
被構成要素 (構成員)	兄弟	兄
構成要素 (物質・材料)	酸素	水
被構成要素 (物質・材料)	水	酸素
被包含領域 (カテゴリ)	虚数	数学
例	ベートーヴェン	作曲家
含意	いびきをかく	寝る
属性	赤	色

3. 日本語ワードネットを用いた意味の数学モデルの構築

本章では、提案手法である、日本語ワードネットを用いた意味の数学モデルの構築手法について述べる。

3.1 データ行列 M の作成

本節では、日本語ワードネットを用いて MDS のもととなるデータ行列 M を作成する手法について述べる。以下にその手順を示す。

(1) 見出し語の選定

2.1.1 節で示したデータ行列 M における見出し語を選定する。日本語ワードネットの各 synset を synset 番号について昇順にソートしたものを、それぞれ順番に見出し語 ($d_1, d_2, \dots, d_n$) とする。

(2) 特徴の選定

2.1.1 節で示したデータ行列 M における特徴を選定する。日本語ワードネットの各 synset を synset 番号について昇順にソートしたものを、それぞれ順番に特徴 ($f_1, f_2, \dots, f_n$) とする。以降、この特徴を説明語と呼ぶ。

(3) 特徴ベクトルの作成

1. 2. より、見出し語と説明語は同じ synset 集合となり、 MDS の作成にはこの synset 集合に含まれる synset を用いる。つまり、 MDS の次元はこの synset 集合に含まれる synset の種類の数と一致する。また、見出し語 d_i と説明語 f_i は同じ synset を示す。

見出し語の定義文を形態素解析し、名詞、動詞、形容詞、副詞以外の単語と接頭語、接尾語を取り除く。また、定義文に含まれるが特定の意味を表さない単語を取り除く。特定の意味を表さない単語とは「いる、ある、もの、こと、…」など独自の意味を持たない単語のことである。例としては、「～である」「～ている」など辞書の定義文中で頻出する単語である。表 3 に取り除く単語の一覧を示す。

取り除かれなかった単語を同義語に含む synset を、見出し語の特徴づけに用いる。複数の synset が存在する場合、それらすべての synset を特徴づけに用いる。定義文中のすべての単語について処理を行った結果、特徴づけに用いる synset が一つもなかった場合、見出し語の synset は MDS 作成に用いる synset 集合から取り除かれる。なお、synset が synset 集合から取り除かれると、 MDS の次元は減少する。

見出し語 d_i について説明語 f_j が特徴づけに用いられていた場合、 d_i に f_j が採用されている、と表現する。 d_i に f_j が採用されていた場合、 d_i の j 番目の要素を 1 にする。 d_i に f_j が採用されていない場合、 d_i の j 番目の要素を 0 にする。また、 d_i の i 番目の要素は 1 にする。以上のような処理を各見出し語に行い、特徴ベクトルを作成する。

表 3 取り除く単語の一覧

いる	ある	もの	こと	てる	できる	いう	よう	とき
さま	ない	する	やる	事	行う	なる	ため	できる

(4) データ行列 M の作成

全ての見出し語について特徴ベクトルを作成し、見出し語の *synset* 番号について昇順にソートし、行ベクトルとして並べ、データ行列 M を作成する。見出し語と説明語は同じ *synset* 集合であることからデータ行列 M は正方行列となる。以下、データ行列 M の i 行 j 列の要素を m_{ij} と表すこととする。

3.2 孫引きや逆引きを考慮したデータ行列 M の作成

見出し語の定義文に含まれないが見出し語と意味的な関連を持つ説明語を採用するために、定義文を用いた説明語の孫引き、*synset* 同士の関係を表すリンクを用いた説明語の孫引き、逆引きの三種類の操作をデータ行列 M に対して行う。それぞれの操作を以下に示す。

3.2.1 操作の説明

(1) 定義文を用いた孫引き

人が辞書である見出し語 x について調べるとき、見出し語 x の定義文中に知らない単語 y が出てきた場合、それを見出し語 y としてもう一度辞書を引き直すことで、見出し語 x の単語の意味を理解しようとする。つまり、見出し語 y の定義文 Y は見出し語 x と関連性を持つといえる。このように再帰的に辞書を引いていく事を孫引きと定義する。*synset* の定義文について孫引きを行うと、直接定義文には記述されていない説明語を採用できる。しかし定義文中の単語全てを孫引きの対象とすると、説明語に採用される単語の数が多くなり見出し語の意味が曖昧になる。そのため、孫引きの対象とする単語を選出する必要がある。今回は、PageRank の値について閾値を設定し、対象とする単語の選出を行った。PageRank の値を用いた対象の選出については 3.2.2 節で説明する。

見出し語 d_i に説明語 f_j が採用されていたとき、見出し語 d_j に採用されている説明語 f_k を見出し語 d_i の説明語に採用する。すなわち、 $m_{ij} = 1$ かつ $m_{jk} = 1$ ならば $m_{ik} = 1$ とする。

(2) *synset* 同士の関係を表すリンクを用いた孫引き

2.2.2 節で示したように、日本語ワードネットには *synset* 同士の関係を表すリンクが定義されている。このリンクについても孫引きを行うことで、見出し語に意味的な関連のある説明語を採用できる。

見出し語 d_i と見出し語 d_j がリンクで結ばれていた時、 $m_{ij} = 1$ とする。また、この操作も今回は PageRank の値について閾値を設定し、対象を選出する。以下、この操作をリンクを用いた孫引きと呼ぶ

(3) 逆引き

「箸」の定義文中に「食べる」という単語が出てくるならば、見出し語「箸」の説明語に「食べる」は採用される。一方、「食べる」の定義文中に「箸」という単語が出てこないならば、見出し語「食べる」の説明語に「箸」は採用されない。しかし、「食べる」という行為は「箸」を使用して行うので、「食べる」と「箸」は互いに関連があると考えられる。そこで、提案手法では、見出し語「箸」に説明語「食べる」が採用されていた場合、見出し語「食べる」に説明語「箸」を採用する。この操作を逆引きと呼ぶ。

i 番目の見出し語に j 番目の説明語が採用されていた時、 j 番

目の見出し語に i 番目の説明語を採用する。すなわち、 $m_{ij} = 1$ ならば $m_{ji} = 1$ とする。

3.2.2 PageRank を用いた単語の選出

多くの見出し語に採用されている説明語は見出し語特有の意味に対する重要性は低く、少数の見出し語に採用されている説明語がその見出し語特有の意味をよく表しており、重要性が高いと考える。見出し語と採用されている説明語をリンクとして考え、 $m_{ij} = 1$ のとき見出し語 i から説明語 j へのリンクがあるとし、PageRank を計算する。定義文を用いた孫引きは、この見出し語と説明語間のリンクをたどる操作と考えられるので PageRank に着目した。PageRank の値が小さいということは、少数の単語に採用されておりその見出し語の意味をよく表す単語であると考えられる。従って、閾値を設定し、孫引きの対象候補の単語の PageRank の値が閾値よりも小さければ定義文を用いた孫引きの対象とする。

リンクを用いた孫引きについても同様に、*synset* 同士をつなぐリンクをリンクとして考え、PageRank を計算する。リンクを用いた孫引きは、このリンクをたどる操作と考えられるので PageRank に着目した。閾値を設定し、孫引きの対象候補の単語の PageRank の値が閾値よりも小さければリンクを用いた孫引きの対象とする。

3.2.3 意味の数学モデルの構築

以上のように作成したデータ行列 M について、相関行列 $M^T M$ を固有値分解し、非ゼロ固有値に対応した固有ベクトルを用いて MDS を作成する。作成した MDS を用いて意味の数学モデルを構築する。

4. 実験

本章では、3. 章で示した手法により日本語ワードネットを用いて構築した意味の数学モデルについて、いくつかの例について *synset* 間の意味的相関を計量し、提案方式の有効性を検証する。

4.1 実験の目的

実験 1 では、与えられた文脈にしたがって部分空間が選択され、メタデータが射影されていることを確認する。実験 2 では、*synset* 間の文脈を考慮した意味的相関の計量が実現されていることを確認する。以上の二つの実験を行い、提案手法の有効性を検証する。また、3.2 節で示した各種操作について検証するために、4 種類の MDS を作成し実験を行う。

4.2 実験に用いる MDS

それぞれの MDS は、3.1 節で作成したデータ行列 M は共通である。行った操作はそれぞれ以下の通りである。

- MDS1: 逆引きのみ行う
- MDS2: 逆引き、定義文を用いた孫引きを行う
- MDS3: 逆引き、リンクを用いた孫引きを行う
- MDS4: 逆引き、定義文を用いた孫引き、リンクを用いた孫引きを行う

それぞれの MDS を用いて実験を行い、各種操作について検証を行う。以降の MDS1 などの表記はここで示した MDS1 のことを示す。また、リンクを用いた孫引きを行う際に、表 1 に

示した 12 種類のリンクを対象とした。

4.3 実験環境

日本語ワードネットからデータ行列 M の作成を行うプログラムを java 言語を用いて作成した。また、固有値分解、射影の計算、意味的相関の計量は MATLAB 言語を用いて計算した。今回は、形態素解析エンジン Igo [5] を用いて形態素解析を行った。メディアデータとして日本語ワードネットの synset を入力し実験を行った。以下、 $\langle \rangle$ でくくられている単語は、その単語の概念を表す synset のことを表している。例えば、 $\langle \text{水} \rangle$ は水という概念を表す synset(synset 番号 14845743- n) のことを表す。

今回、定義文を用いた孫引きの対象を選定するために PageRank を計算した結果、最大 1.9609e-5、最少 6.4112e-10 となり、閾値を 8.0016e-10 とした。これより、56817 個中 44190 個の synset が定義文を用いた孫引きの対象となった。

また、リンクを用いた孫引きについても PageRank を計算し、最大 0.0042、最少 1.1383e-9 となり、閾値を 1.5749e-5 とした。これより、他の synset とリンクを持つ synset の 48905 個中 29304 個がリンクを用いた孫引きの対象となった。

4.4 実験 1

4.4.1 実験内容

本実験では、2.1.4 節で示したメタデータ $\mathbf{x}$ のノルム $norm(\mathbf{x})$ を意味的相関として計量する。この値が大きいくほど、与えられた文脈とメタデータ間の意味的相関が大きいことを示している。

文脈を与え、文脈と多義語の持つ複数の意味との意味的相関をそれぞれ計量する。そして、文脈と最も意味的相関が大きい意味を、その文脈における多義語の表す意味であると判断する。与えられた文脈の変化により、その文脈における多義語の表す意味も変化することを確認し、文脈にしたがって部分空間が選択され、メタデータが適切に射影されていることを確認する。4.2 節で示した 4 種類の MDS を用いて検証を行う。

実験に用いる計量対象語と文脈のセットを 3 種類用意し、以下に示す。

実験 1-1 計量対象語「夢」

- (1) 睡眠中に見る幻覚のこと (以下、夢 1 と表す)
- (2) 将来実現させたい望み (以下、夢 2 と表す)

$\langle \text{睡眠} \rangle$ と $\langle \text{欲求} \rangle$ という文脈を与え、 $\langle \text{夢 1} \rangle$ と $\langle \text{夢 2} \rangle$ の MDS でのノルムを計量する。 $\langle \text{睡眠} \rangle$ という文脈では $\langle \text{夢 1} \rangle$ の方が $\langle \text{夢 2} \rangle$ よりもノルムが大きく、 $\langle \text{欲求} \rangle$ という文脈では $\langle \text{夢 2} \rangle$ の方が $\langle \text{夢 1} \rangle$ よりもノルムが大きいと予想する。

実験 1-2 計量対象語「足」

- (1) 人間の体の一部 (以下、足 1 と表す)
 - (2) 移動の手段 (以下、足 2 と表す)
 - (3) 家具などの本体を支える突き出た部分 (以下、足 3 と表す)
- $\langle \text{人間} \rangle$ と $\langle \text{公共交通機関} \rangle$ と $\langle \text{家具} \rangle$ という文脈を与え、 $\langle \text{足 1} \rangle$ 、 $\langle \text{足 2} \rangle$ 、 $\langle \text{足 3} \rangle$ の MDS でのノルムを計量する。 $\langle \text{人間} \rangle$

という文脈では $\langle \text{足 1} \rangle$ の方が $\langle \text{足 2} \rangle$ 、 $\langle \text{足 3} \rangle$ よりもノルムが大きく、 $\langle \text{公共交通機関} \rangle$ という文脈では $\langle \text{足 2} \rangle$ の方が $\langle \text{足 1} \rangle$ 、 $\langle \text{足 3} \rangle$ よりもノルムが大きく、 $\langle \text{家具} \rangle$ という文脈では $\langle \text{足 3} \rangle$ の方が $\langle \text{足 1} \rangle$ 、 $\langle \text{足 2} \rangle$ よりもノルムが大きいと予想する。

実験 1-3 計量対象語「手」

- (1) 人間の手 (以下、手 1 と表す)
- (2) 何かを行う方法、手段 (以下、手 2 と表す)
- (3) トランプゲームにおいてプレイヤーが持っているカード (以下、手 3 と表す)

$\langle \text{四肢} \rangle$ と $\langle \text{解決法} \rangle$ と $\langle \text{トランプ} \rangle$ という文脈を与え、 $\langle \text{手 1} \rangle$ 、 $\langle \text{手 2} \rangle$ 、 $\langle \text{手 3} \rangle$ の MDS でのノルムを計量する。 $\langle \text{四肢} \rangle$ という文脈では $\langle \text{手 1} \rangle$ の方が $\langle \text{手 2} \rangle$ 、 $\langle \text{手 3} \rangle$ よりもノルムが大きく、 $\langle \text{解決法} \rangle$ という文脈では $\langle \text{手 2} \rangle$ の方が $\langle \text{手 1} \rangle$ 、 $\langle \text{手 3} \rangle$ よりもノルムが大きく、 $\langle \text{トランプ} \rangle$ という文脈では $\langle \text{手 3} \rangle$ の方が $\langle \text{手 1} \rangle$ 、 $\langle \text{手 2} \rangle$ よりもノルムが大きいと予想する。

4.4.2 実験結果

実験 1-1 の MDS1 での結果を表 4、MDS2 での結果を表 5、MDS3 での結果を表 6、MDS4 での結果を表 7 に示す。同様に、実験 1-2 の結果を表 8~11、実験 1-3 の結果を表 12~15 に示す。また、実験 1 の結果をまとめて表 16 に示す。表中の $\bigcirc$ は列の MDS を用いて行の文脈において実験を行った結果、予想通りの結果が示されたことを表し、 $\times$ は予想に反する結果が示されたことを表す。

表 4 実験 1-1 MDS1

計量対象	文脈	
	$\langle \text{睡眠} \rangle$	$\langle \text{欲求} \rangle$
$\langle \text{夢 1} \rangle$	0.0241	0.0244
$\langle \text{夢 2} \rangle$	0.0585	0.0279

表 5 実験 1-1 MDS2

計量対象	文脈	
	$\langle \text{睡眠} \rangle$	$\langle \text{欲求} \rangle$
$\langle \text{夢 1} \rangle$	0.0373	0.4717
$\langle \text{夢 2} \rangle$	0.0176	0.0083

表 6 実験 1-1 MDS3

計量対象	文脈	
	$\langle \text{睡眠} \rangle$	$\langle \text{欲求} \rangle$
$\langle \text{夢 1} \rangle$	0.0214	0.0131
$\langle \text{夢 2} \rangle$	0.0004	0.0265

表 7 実験 1-1 MDS4

計量対象	文脈	
	$\langle \text{睡眠} \rangle$	$\langle \text{欲求} \rangle$
$\langle \text{夢 1} \rangle$	0.0441	0.0246
$\langle \text{夢 2} \rangle$	0.0234	0.0511

表 8 実験 1-2 MDS1

計量対象	文脈		
	$\langle \text{人間} \rangle$	$\langle \text{公共交通機関} \rangle$	$\langle \text{家具} \rangle$
$\langle \text{足 1} \rangle$	0.0097	0.0839	0.147
$\langle \text{足 2} \rangle$	0.0149	0.8349	0.1821
$\langle \text{足 3} \rangle$	0.0088	0.0817	0.1021

表 9 実験 1-2 MDS2

計量対象	文脈		
	$\langle \text{人間} \rangle$	$\langle \text{公共交通機関} \rangle$	$\langle \text{家具} \rangle$
$\langle \text{足 1} \rangle$	0.0008	0.0583	0.0111
$\langle \text{足 2} \rangle$	0.0025	0.0345	0.0017
$\langle \text{足 3} \rangle$	0.00278	0.0062	0.0226

表 10 実験 1-2 MDS3

計量対象	文脈		
	〈人間〉	〈公共交通機関〉	〈家具〉
〈足 1〉	0.0056	0.1067	0.0199
〈足 2〉	0.0036	0.1532	0.0199
〈足 3〉	0.0008	0.1427	0.0258

表 11 実験 1-2 MDS4

計量対象	文脈		
	〈人間〉	〈公共交通機関〉	〈家具〉
〈足 1〉	0.0378	0.8149	1.2667
〈足 2〉	0.0207	0.9976	1.1413
〈足 3〉	0.0322	0.8084	1.3626

表 12 実験 1-3 MDS1

計量対象	文脈		
	〈四肢〉	〈解決法〉	〈トランプ〉
〈手 1〉	0.0083	0.0286	0.0021
〈手 2〉	0.0010	0.0279	0.0228
〈手 3〉	0.0076	0.0060	0.0101

表 13 実験 1-3 MDS2

計量対象	文脈		
	〈四肢〉	〈解決法〉	〈トランプ〉
〈手 1〉	0.0189	0.5785	0.0574
〈手 2〉	0.0098	0.0083	0.0494
〈手 3〉	0.0056	0.0667	0.0047

表 14 実験 1-3 MDS3

計量対象	文脈		
	〈四肢〉	〈解決法〉	〈トランプ〉
〈手 1〉	0.0113	0.0071	0.0155
〈手 2〉	0.0075	0.0046	0.0377
〈手 3〉	0.0042	0.0077	0.0742

表 15 実験 1-3 MDS4

計量対象	文脈		
	〈四肢〉	〈解決法〉	〈トランプ〉
〈手 1〉	0.5276	0.0094	0.0065
〈手 2〉	0.0093	0.0199	0.0105
〈手 3〉	0.0023	0.0032	0.0207

表 16 実験 1 まとめ

	文脈	MDS1	MDS2	MDS3	MDS4
実験 1-1 夢	睡眠	×	○	○	○
	欲求	○	×	○	○
実験 1-2 足	人間	×	×	○	○
	公共交通機関	○	×	○	○
	家具	×	○	○	○
実験 1-3 手	四肢	○	○	○	○
	解決法	×	×	×	○
	トランプ	×	×	○	○

4.4.3 実験考察

● 実験 1-1

MDS1 では、〈睡眠〉という文脈において〈夢 2〉の方が〈夢

1〉よりもノルムが大きいという予想に反する結果が示された。MDS2 では、〈欲求〉という文脈において、〈夢 1〉の方が〈夢 2〉よりもノルムが大きいという予想に反する結果が示された。MDS3, MDS4 では予想通りの結果が示された。

● 実験 1-2

MDS1, MDS2 では二種類の文脈において予想に反する結果が示され、MDS3, MDS4 では予想通りの結果が示された。

● 実験 1-3

MDS1, MDS2 では二種類の文脈において予想に反する結果が示され、MDS3 では〈解決法〉という文脈において予想に反する結果が示された。MDS4 では予想通りの結果が示された。

以上より、与えられた文脈が変化すると、その文脈における多義語の表す意味も変化することを確認した。すなわち、文脈にしたがって部分空間が選択され、メタデータが適切に射影されていることを確認した。また、MDS4 が 4 種類の MDS の中で予想通りの結果を示した数が一番多いということを確認した。

4.5 実験 2

4.5.1 実験内容

本実験では、synset 間の文脈を考慮した意味的相関の計量が実現されていることを確認する。2.1.4 節で示した動的距離 $\rho(x, y)$ を意味的相関として計量する。動的距離 $\rho(x, y)$ の値が 0 に近いほど〈x〉と〈y〉の与えられた文脈における意味的な距離が近い、つまり意味的相関が大きいといえる。

以下のように計量対象語と文脈のセットを二種類用意した。

実験 2-1 では、文脈として、〈車輪〉、〈背もたれ〉の二種類を与え、〈車いす〉と〈車〉、〈車いす〉と〈いす〉の間の意味的相関を計量する。〈車輪〉という文脈においては、〈車いす〉と〈車〉の間の意味的相関の方が、〈車いす〉と〈いす〉よりも大きいと予想する。

実験 2-2 では、文脈として、〈くちばし〉、〈爬虫類〉、〈絶滅する〉の三種類を与え、〈翼竜〉と〈鳥〉、〈翼竜〉と〈恐竜〉の間の意味的相関を計量する。〈くちばし〉という文脈においては、〈翼竜〉と〈鳥〉の間の意味的相関のほうが、〈翼竜〉と〈恐竜〉よりも大きいと予想し、〈爬虫類〉、〈絶滅する〉という文脈においては、〈翼竜〉と〈恐竜〉の間の意味的相関のほうが、〈翼竜〉と〈鳥〉よりも大きいと予想する。

4.5.2 実験結果

実験 2-1 の MDS1 での結果を表 17, MDS2 での結果を表 18, MDS3 での結果を表 19, MDS4 での結果を表 20 に示す。同様に、実験 2-2 の結果を表 21~24 に示す。また、実験 2 の結果をまとめて表 25 に示す。表中の○は列の MDS を用いて行の文脈において実験を行った結果、予想通りの結果が示されたことを表し、× は予想に反する結果が示されたことを表す。

表 17 実験 2-1 MDS1

計量対象	文脈	
	〈車輪〉	〈背もたれ〉
〈車いす〉と〈車〉	0.000295	0.009401
〈車いす〉と〈いす〉	0.017728	0.020397

表 18 実験 2-1 MDS2

計量対象	文脈	
	〈車輪〉	〈背もたれ〉
〈車いす〉と〈車〉	0.011798	0.016893
〈車いす〉と〈いす〉	0.029586	0.024977

表 19 実験 2-1 MDS3

計量対象	文脈	
	〈車輪〉	〈背もたれ〉
〈車いす〉と〈車〉	0.006767	0.055847
〈車いす〉と〈いす〉	0.010459	0.083611

表 20 実験 2-1 MDS4

計量対象	文脈	
	〈車輪〉	〈背もたれ〉
〈車いす〉と〈車〉	0.09137	0.20811
〈車いす〉と〈いす〉	0.26068	0.13059

表 21 実験 2-2 MDS1

計量対象	文脈		
	〈くちばし〉	〈爬虫類〉	〈絶滅する〉
〈翼竜〉と〈鳥〉	0.01931	0.031421	0.003662
〈翼竜〉と〈恐竜〉	0.13036	0.008477	0.004177

表 22 実験 2-2 MDS2

計量対象	文脈		
	〈くちばし〉	〈爬虫類〉	〈絶滅する〉
〈翼竜〉と〈鳥〉	0.009631	0.12399	0.01155
〈翼竜〉と〈恐竜〉	0.019582	0.00318	0.02511

表 23 実験 2-2 MDS3

計量対象	文脈		
	〈くちばし〉	〈爬虫類〉	〈絶滅する〉
〈翼竜〉と〈鳥〉	0.01443	0.02156	0.021844
〈翼竜〉と〈恐竜〉	0.06622	0.00309	0.008146

表 24 実験 2-2 MDS4

計量対象	文脈		
	〈くちばし〉	〈爬虫類〉	〈絶滅する〉
〈翼竜〉と〈鳥〉	0.26322	0.33134	0.31298
〈翼竜〉と〈恐竜〉	0.30191	0.24544	0.10446

表 25 実験 2 まとめ

	文脈	MDS1	MDS2	MDS3	MDS4
実験 2-1	車輪	○	○	○	○
	背もたれ	×	×	×	○
実験 2-2	くちばし	○	○	○	○
	爬虫類	○	○	○	○
	絶滅する	×	×	○	○

4.5.3 実験考察

● 実験 2-1

MDS1, MDS2, MDS3 では、〈背もたれ〉という文脈において予想に反する結果が示された。表 MDS4 ではどの文脈においても予想通りの結果が示された。

● 実験 2-2

MDS1, MDS2 では、〈絶滅する〉という文脈において予想に反する結果が示された。表 MDS3, MDS4 では、どの文脈においても予想通りの結果が示された。

以上より、synset 間の文脈を考慮した意味的相関の計量が実現されていることを確認した。また、MDS4 が他の MDS に比べて予想通りの結果を一番多く示したことを確認した。

4.6 考察

実験 1 では、文脈にしたがって部分空間が選択され、メタデータが適切に射影されていることを確認した。

実験 2 では、synset 間の文脈を考慮した意味的相関の計量が実現されていることを確認した。

以上より、日本語ワードネットを用いて構築された意味の数学モデルにおいて、メタデータ間の文脈を考慮した意味的相関の計量が可能であるといえる。

また、3.2 節で示した各種操作については、各実験共に MDS4 が 4 種類の MDS の中で予想通りの結果を示した数が一番多いという結果が示されたので、本稿で提案した各種操作を組み合わせでデータ行列 M を作成することが意味的相関計量において有効であると考えられる。

5. 結論

本稿では日本語のシソーラスである日本語ワードネットを用いて MDS を作成し、意味の数学モデルを構築する方法を示した。また、逆引きや孫引きを組み合わせで 4 種類の MDS を作成し、それぞれの MDS について検証を行い、逆引きや孫引きについて有効性を示した。また、日本語ワードネットで定義されている synset 同士の文脈を考慮した意味的相関の計量について検証を行い、提案手法の有効性を示した。

今後の課題としては、意味的相関を計量する際に日本語の単語を入力できるシステムの作成が挙げられる。現状の入力は synset であり、日本語の単語を直接入力することはできない。日本語ワードネットを用いて日本語の単語に対応する synset を決定する場合、対象単語を同義語に含む synset を選択し、入力としている。しかし、対象単語を同義語に含む synset が多数存在する場合があります。どの synset を入力すべきか選択しなければならないという問題がある。現状では複数の synset から人手で最も適切な synset を選び、意味的相関を計量する際の入力としている。この際に、適切な synset を機械的に決定する手法の考案が必要である。

文献

- [1] T. Kitagawa, Y. Kiyoshi. “The Mathematical Model of Meaning and its Application to Multidatabase Systems,” Proceeding of 3rd IEEE International Workshop on Research Issue on Data Engineering-Interoperability in Multidatabase System. pp. 130-135 April 1993.
- [2] Longman Dictionary of Contemporary English. 1987.
- [3] Francis Bond, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto, “Japanese SemCor: A Sense-tagged Corpus of Japanese,” The 6th International Conference of the Global Word-Net Association (GWC-2012), Matsue.
- [4] 日本語 WordNet
<<http://nlpwww.nict.go.jp/wn-ja/>> (2013/1/24 アクセス)
- [5] Igo : a morphological analyzer
<<http://igo.sourceforge.jp/>> (2013/1/24 アクセス)