# コモンセンス知識ベースを用いた推論の性能評価

## An Evaluation of Commonsense Knowledge-based Reasoning

中　原　和　洋

**要　約**　コモンセンス知識ベースの ConceptNet，推論手法の AnalogySpace や Spectral Association が開発され，コモンセンスを対象とした人工知能の研究基盤が整いつつある．筆者はこれまで日本におけるコモンセンス知識収集と知識ベース構築活動を行ってきた．本論では，上述の知識ベースと AnalogySpace を用いた概念間類似度推定および Spectral Association を用いた概念間関連度推定の性能評価手法の提案と評価結果を報告する．概念間類似度推定性能の評価として，市販の 3 ～ 4 歳児向け幼児教材の仲間外れ概念探し問題を利用する手法を提案する．また，概念間関連度推定性能の評価として，日本人の連想語調査結果を利用する手法を提案する．両提案評価手法により，コモンセンス知識収集手法や知識量に応じた推論性能の把握が可能となった．また，両評価において，コモンセンス推論を用いた手法は比較手法に対し良好な性能を示した．

**Abstract**　The basis of the commonsense AI is now ready for the further research with the development of ConceptNet, the knowledge base of commonsense, and its two techniques of reasoning, AnalogySpace and Spectral Association. The author of this paper has been involving in the acquisitions of commonsense knowledge data in Japan. There are two known techniques for commonsense reasoning: one is to infer the degree of similarity between two concepts in AnalogySpace, and the other is to infer the degree of relatedness between two concepts in Spectral Association. This paper proposes a method for each to assess the adequacy of the outputs the commonsense reasoning gives and reports the results of the assessments. The former reasoning technique was assessed using the educational materials for human child aged 3-4, which requires the child to identify the most dissimilar concept from the sets of 4-5 concepts. The latter technique was assessed using the word association survey of Japanese. The evaluation of two sets of the assessment result demonstrated the superiority of the commonsense AI over the existing methods.

### 1. は　じ　め　に

人工知能研究において，コモンセンスを対象とした知識ベースや推論は古くから重要な課題の一つとして認識されてきた．2000 年代に入り，コモンセンス知識ベース ConceptNet[1]や推論手法 AnalogySpace[2]，Spectral Association[3]が開発されるなど，コモンセンスを対象とした研究の基盤が整備されつつある．筆者らも日本におけるコモンセンス知識の獲得を進めてきた．一方で，ある時点におけるコモンセンス知識ベースを使った推論の性能を客観的，定量的に評価することは，研究到達状況の見える化や課題抽出，応用先の検討などの様々な観点から重要である．本論は，筆者らが日本で収集したコモンセンス知識と AnalogySpace を利用した概念間類似度推定および Spectral Association を利用した概念間関連度推定の客観的，定量的な性能評価手法の提案とその評価報告を目的とする．

本論の構成は以下の通りである．2 章で ConceptNet，AnalogySpace を用いた概念間類似度

推定，Spectral Association を用いた概念感関連度推定，日本におけるコモンセンス知識獲得研究について，3 章で概念間類似度推定性能の評価手法の提案，評価結果，考察，4 章で概念間関連度推定性能の評価手法の提案，評価結果，考察，5 章でまとめを述べる．

## 2.　コモンセンス知識ベースと推論

本章では，評価実験で利用したコモンセンス知識ベースの ConceptNet，およびコモンセンス推論手法の AnalogySpace を用いた概念間類似度推定，Spectral Association を用いた概念間関連度推定，最後に，筆者がこれまでに実施した日本におけるコモンセンス知識収集について記述する．

### 2. 1　ConceptNet[1]

ConceptNet は，マサチューセッツ工科大学メディアラボ（MIT メディアラボ）が開発中のコモンセンス知識ベースである．ConceptNet では，概念（Concept）をノード，概念間の関係（Relation）をアークとした表明（Assertion）の集合（意味ネットワーク）でコモンセンス知識を表現する．概念はそれを表す単語や短いフレーズで表現し，関係は IsA，HasProperty，PartOf，Desire などあらかじめ規定されたものを用いる．ConceptNet では，意味表現（Assertion）と表層表現（Sentence）を対応づけてデータを保持している（図 1）．

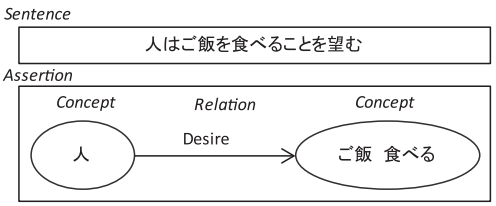


**図 1　ConceptNet の知識表現**

### 2. 2　AnalogySpace による概念間類似度計算[2]

AnalogySpace は，MIT メディアラボが開発した ConceptNet に基づいたコモンセンス推論手法である．AnalogySpace は，ConceptNet のコモンセンス知識ベースを行列（マトリクス）に変換し，線形代数における特異値分解（SVD）と次元削減を用いて主成分分析を適用する手法である．ConceptNet の Assertion を Concept-Feature マトリクス $\boldsymbol{A}$ に変換する．Feature は Concept と Relation の組である．一般に Concept-Feature マトリクスは大部分を 0 が占めるスパース行列となる．このマトリクス $\boldsymbol{A}$ に対して SVD と次元数 k による次元削減を行うことで得た $\boldsymbol{A}$ の近似 $\boldsymbol{A}_{\mathrm{k}} = \boldsymbol{U}_{\mathrm{k}}\boldsymbol{\Sigma}_{\mathrm{k}}\boldsymbol{V}_{\mathrm{k}}^{\mathrm{T}}$ によって，$\boldsymbol{A}$ の 0 成分であった ConceptNet には含まれない Concept-Feature の組（未知の Assertion）の類推値を得ることができる．また，Concept 間の類似度計算を $\boldsymbol{A}_{\mathrm{k}}\boldsymbol{A}_{\mathrm{k}}^{\mathrm{T}} = \boldsymbol{U}_{\mathrm{k}}\boldsymbol{\Sigma}_{\mathrm{k}}(\boldsymbol{U}_{\mathrm{k}}\boldsymbol{\Sigma}_{\mathrm{k}})^{\mathrm{T}}$ により算出する．AnalogySpace を提案した論文[2]では限定的なサンプリングによる推論の正確性の評価にとどまっており，人間の持つコモンセンスに比べた推論性能という観点では，客観的，定量的な評価はなされていない．

### 2. 3　Spectral Association による概念間関連度推定[3]

Spectral Association は，意味ネットワークにおける活性化拡散モデルに基づいて概念ペアの意味的関連性（連想）の強さを計算する手法である．次元圧縮により効率的に近似計算するところに特長がある．ConceptNet を用いた具体的な計算方法は，まず ConceptNet においてアークで繋がる Concept ペアの要素値は値を持つように Concept-Concept マトリクス $\boldsymbol{C}$ を生成する（アークが存在しない Concept ペアの要素値は 0）．このとき，活性化拡散モデルに基づく Concept ペアの関連度の計算は，以下の式となる．

$$\mathbf{1}+\boldsymbol{C}+\left(\frac{\boldsymbol{C}^2}{\mathbf{2!}}\right)+\left(\frac{\boldsymbol{C}^3}{\mathbf{3!}}\right)+\ldots=\boldsymbol{e}^{\boldsymbol{C}}$$

ここで $\boldsymbol{C}$ をスペクトル分解し次元圧縮した $\boldsymbol{C}_k=\boldsymbol{V}_k\boldsymbol{\Lambda}_k\boldsymbol{V}_k^T$ を上式に代入すると，$\boldsymbol{e}^{\boldsymbol{C}}\approx\boldsymbol{V}_k\boldsymbol{e}^{\boldsymbol{\Lambda}}\boldsymbol{V}_k^T$ で近似計算できる．さらに $\boldsymbol{V'}=\boldsymbol{V}_k\boldsymbol{e}^{\boldsymbol{\Lambda}/2}$ と置くと $\boldsymbol{e}^{\boldsymbol{C}}=\boldsymbol{V'}\boldsymbol{V'}^T$ となる．

### 2. 4　日本におけるコモンセンス知識収集

コモンセンス知識は当たり前すぎて Web 文書などには表明されにくい傾向にあるため，これまで人手による知識ベース化方法が取られてきた．Cyc[4]では知識専門家の手によって，OMCS[5]ではインターネット上のボランティアによって，また Game With A Purpose（GWAP）によりインターネット上のゲームによって知識ベース化をスケールさせる方法も行われている[6][7]．筆者らも，日本におけるコモンセンス知識獲得を目的とした二つのインターネットゲームサイトを立ち上げて知識ベース化を進めてきた．

#### 2. 4. 1　ナージャとなぞなぞ

2010 年に開始したインターネットゲームサイト「ナージャとなぞなぞ」[8]は，ナージャというキャラクターとインターネット上のプレイヤーの間で行われる連想ゲームで，ナージャが出す五つのヒントをもとに彼女が頭に思い浮かべることばをプレイヤーが当てるゲームである．プレイヤーの回答がコモンセンス知識として収集される．2013 年 7 月時点で，24 万件のコモンセンス知識（Sentence 数）を獲得している．

#### 2. 4. 2　日本人検定

2012 年に開始した Facebook アプリの「日本人検定」[9]は，自分の「日本人レベル」を調べることができる一回 20 問のクイズ形式のゲームである．回答者全員の回答から「日本人っぽさ」を導き出して，その得点を Facebook の友人たちと互いにシェアして競い合う．ユーザの回答内容がコモンセンス知識として収集される．2013 年 7 月時点で，約 61 万件のコモンセンス知識（Sentence 数）を獲得している．

## 3.　評価実験 1：概念間類似度推定

本章では，概念間類似度推定の性能評価手法の提案および評価結果，考察について述べる．市販の 3 ～ 4 歳児向けの幼児教材[10][11]における仲間外れ概念探し問題に取り組み，正答率を評価することとした．幼児教材の仲間外れ概念探し問題は，四つないし五つの概念を示すイラストが提示され，その中からもっとも仲間外れの概念を一つ回答する問題である．例えば，出題

概念集合 $C=\{$犬，猫，うさぎ，鳥$\}$ の各イラストが与えられ，最も仲間外れの概念である鳥を回答する問題である．様々な理由付けにより異なる概念を仲間外れと見なすことも可能であるが，幼児教材では一つの正解（常識的に判断し最も典型的で自然な仲間外れ）が割り当てられており，これを正解と見なすことで客観的で定量的な評価が可能になると考えた．

### 3. 1　評価実験 1 の手順

本節では概念間類似度推定の評価実験手順について述べる．

#### 3. 1. 1　教材問題のテキスト化

本評価では画像認識は対象外とし，筆者らが人手で上述の教材中のイラストから出題概念と正解概念のテキスト化を行い，問題の入力，評価システムの回答，正解はすべてテキストとした．人手によるテキスト化は，教材に書かれている「題意」と「正解」に従うことで客観性は担保し，イラストをテキスト表現する際の多様性に対しては以下の方針を定めて人手で対応した．

- 題意が変わらない範囲で ConceptNet に含まれる Concept のテキスト表現を用いる
- ConceptNet の複数の Concept に対応する場合は，Concept の Feature 数が最も多い Concept のテキスト表現を用いる

実際に作成した問題総数は 4 択問題が 130 問，5 択問題が 38 問の合計 168 問である．

#### 3. 1. 2　前処理

ConceptNet4.0 には一つの Concept（見出し語）に対応する複数の表層表現の統合機能が実装されているが，日本語対応は十分ではなく，例えば漢字とひらがなは統合されない．筆者らは別途これらの統合処理を行った．

#### 3. 1. 3　回答処理

$n$ 択の仲間外れ探し問題における概念集合 $C=\{c_1, c_2, \ldots, c_n\}$ から仲間外れ概念 $c_{out}$ を以下の式で導出する．

$$c_{out} = \arg\min_{c_i \in C} \left( \sum_{\substack{c_j \in C \\ c_j \neq c_i}} \mathrm{sim}(c_i, c_j) \right) \qquad (1)$$

$\mathrm{sim}(c_i, c_j)$ は，$c_i$ と $c_j$ の間の類似度である．$C$ に含まれる Concept のうち三つ以上の Concept が既知である（Concept-Feature マトリクスの行成分に存在する）場合，その問題に対して回答可能とし，既知の Concept 集合内のみで $c_{out}$ を計算しシステムの回答とした．既知の Concept が 3 未満であった場合は，その問題に対して回答不能とした．

#### 3. 1. 4　　回答の採点と評価指標の導出

問題総数を $N=168$，N のうち回答可能であった問題数を有効回答数 $N_a$，正答した問題数を $N_c$ とした時に，性能評価指標として，有効回答率 $RR=N_a/N$，有効回答正答率 $RCR=N_c/N_a$，正答率 $CR=N_c/N$ を導出した．

### 3.2　評価実験システム

評価実験システムは，コモンセンス知識ベースに ConceptNet4.0 を用い，概念間の類似度計算に AnalogySpace の実装である Divisi2 を用いた．仲間外れ概念の回答には式(1)を用いた．以降本手法をコモンセンス手法と呼ぶ．

#### 3.2.1　コモンセンス知識セット

知識量や，知識収集手法毎の性能把握のために複数の知識セットを作成し，評価を行った．All は 2013 年 7 月時点のすべての知識セット，Mid は All から 50%を，Small は All から 10%の Sentence をランダムサンプリングして生成した知識セットである．Default は ConceptNet4.0 の配布版で提供されている日本語知識セット，Nadya はナージャとなぞなぞで，Kentei は日本人検定で収集した知識のみで生成した知識セットである．それぞれの知識セットの Sentence 数，Assertion 数，Concept 数を表 1 に示す．

**表 1　評価実験で利用したコモンセンス知識セット**

| 知識セット名 | Sentence 数 | Assertion 数 | Concept 数 |
|---|---|---|---|
| All | 868,228 | 181,820 | 65,295 |
| Mid | 434,114 | 110,558 | 43,847 |
| Small | 86,823 | 32,160 | 17,590 |
| Default | 14,368 | 12,825 | 11,100 |
| Nadya | 243,010 | 102,424 | 15,462 |
| Kentei | 610,850 | 73,083 | 52,126 |

#### 3.2.2　テストパラメータ

結果に影響を与えるテストパラメータとして，SVD の圧縮次元数 $K = \{100, 200, 300, 400\}$，Concept-Feature マトリクス $\boldsymbol{A}$ の有効データとする行および列方向の最小非 0 成分数 $Cutoff = \{1, 3, 5\}$，$\boldsymbol{A}$ の正規化（各成分を行および列ベクトルのノルムで割る）の有無 $Prenorm = \{1, 0\}$，類似度計算における $\boldsymbol{U}_k\boldsymbol{\Sigma}_k$ の正規化（各成分を行ベクトルのノルムで割る）の有無 $Postnorm = \{1, 0\}$ の四つのパラメータのすべての組み合わせについて実施した．

### 3.3　比較手法

コモンセンス手法の特長は，Web 上の文書などには表明されづらい当たり前すぎるコモンセンス知識を人間から直接獲得し利用するところにある．そこで，本評価における比較手法の一つとして，Web 上の大規模文書である Wikipedia を用いた潜在意味解析（LSA）手法を選定した．もう一つの比較手法としては，概念間の類似度計算で広く用いられている WordNet を利用した手法を選定した．それぞれの比較手法で，概念間類似度 $\mathrm{sim}(c_1, c_2)$ を計算し，式(1)を用いて仲間外れ概念を回答する．

#### 3.3.1　日本語 Wikipedia 記事を利用した LSA による概念間類似度計算

日本語 Wikipedia の記事を Concept，記事内に登場する単語を Feature と捉え，Concept-Feature マトリクスを作成し，コモンセンス手法と同様のアルゴリズムを利用して，Concept

（記事）間の類似度を計算した．成分値は tf-idf とした．出題概念テキストと日本語 Wikipedia 記事の対応付けは筆者らが人手で行った．形態素解析には Mecab[*1] を用い，辞書は新語などに対応した独自辞書を利用した．対象単語は名詞，動詞，形容詞とし，“する”，“ある”などの一般語は除外した．また，マトリクスの行ベクトルおよび列ベクトルの非 0 成分数が 5 未満となるデータは除外した．さらに，Wikipedia 記事には，常識的な概念とは言い難い記事が大量に存在するため，対象記事の絞り込みを行った．絞り込み方法は，Wikipedia の記事カテゴリのリンクデータ（カテゴリグラフ）を利用し，全出題概念に対応する記事カテゴリからカテゴリグラフ上の距離 $L$ 以下のカテゴリに属する記事のみを対象記事とした．表 2 に示すカテゴリ距離 $L = \{0, 2, 4, \text{全カテゴリ}\}$ の 4 種類の記事セットを作成し，それぞれで評価した．この記事セットと，コモンセンス手法と同様のテストパラメータ，SVD 次元数 $K = \{100, 200, 300, 400\}$，$Prenorm = \{1, 0\}$，$Postnorm = \{1, 0\}$ のすべての組み合わせで実験を行った．LSA の実装には Divisi2 を用いた．上述の手法を以降 Wikipedia 手法と呼ぶ．

**表 2　日本語 Wikipedia の LSA による比較手法**

| 記事セット名 | カテゴリ距離 | カテゴリ数 | 記事数 | 単語数 |
|---|---|---|---|---|
| WLSA_A | 全カテゴリ | 116,162 | 884,583 | 390,761 |
| WLSA_4 | 4 | 4,902 | 150,823 | 224,932 |
| WLSA_2 | 2 | 2,416 | 84,029 | 157,190 |
| WLSA_0 | 0 | 566 | 34,607 | 78,764 |

#### 3. 3. 2　WordNet を用いた概念間類似度計算

WordNet では概念間の類似度計算手法がいくつか提案されており，nltk[*2] に関数として実装されている 6 種類の概念間類似度計算手法（Path Distance, Leacock Chodorow[12]，Wu-Palmer[13]，Resnik[14]，Jiang-Conrath[15]，Lin[16]）を比較手法として用いた．日本語 WordNet1.1[17] の英語版 WordNet へのリンクファイルを用いて英語版 WordNet の synset へのマッピングを行った後に，nltk を用いて英語版 WordNet 上で synset 間の類似度の計算を行った．一つの出題概念に対して複数の synset へのリンクが存在する場合は，筆者らが人手で題意として適切な synset への対応付けを行った．概念の情報量を必要とする類似度手法については，Brown コーパスを用いて算出した情報量を利用した．上述の手法を以降 WordNet 手法と呼ぶ．

### 3. 4　評価実験結果

本節では概念間類似度推定性能の評価実験結果について記述する．

#### 3. 4. 1　最大性能の比較

コモンセンス手法と比較手法の実験結果について，それぞれ最大の正答率を出したテストケースの結果を表 3 に示す．コモンセンス手法は，知識ソース = All，$Cutoff = 3$，$K = 400$，$Prenorm = 0$，$Postnorm = 1$ である．Wikipedia 手法は記事セット = WLSA_0，$K = 300$，$Prenorm = 1$，$Postnorm = 0$ である．WordNet 手法は Resnik Similarity による結果である．表 3 の p 値は，コモンセンス手法との片側検定による p 値である．

表 3　手法毎の結果

| 手法名 | コモンセンス | Wikipedia | WordNet |
|---|---|---|---|
| 有効回答数 | 159 | 160 | 153 |
| 正答数 | 103 | 91 | 83 |
| 有効回答率 | 0.95 | 0.95 | 0.91 |
| 有効回答正答率 | 0.65 | 0.57 | 0.54 |
| 正答率 | 0.61 | 0.54 | 0.49 |
| 有効回答正答率 p 値 | | 0.09153 | 0.03755 |
| 正答率 p 値 | | 0.1122 | 0.01853 |

### 3. 4. 2　知識量別の結果

コモンセンス手法における知識量別の結果を表 4 に示す．また図 2 に知識量（Assertion 数）と有効回答正答率の関係を示す．

表 4　知識量別の結果

| 知識ソース名 | Small | Mid | All |
|---|---|---|---|
| 有効回答数 | 111 | 154 | 159 |
| 正答数 | 60 | 90 | 103 |
| 有効回答率 | 0.66 | 0.92 | 0.95 |
| 有効回答正答率 | 0.54 | 0.58 | 0.65 |
| 正答率 | 0.36 | 0.54 | 0.61 |

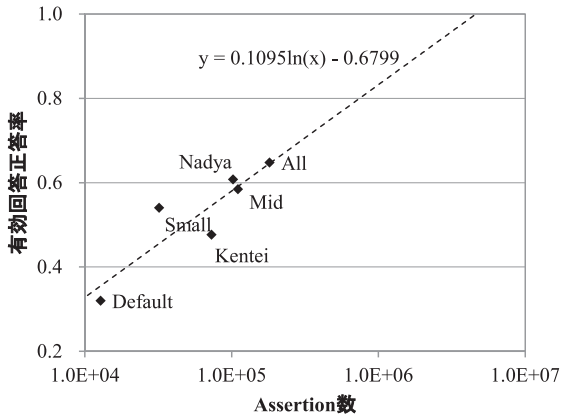


図 2　Assertion 数と有効回答正答率の関係

### 3. 4. 3　知識収集手法別の結果

コモンセンス手法における知識収集手法別の結果を表 5 に示す．

表 5　知識収集手法別の結果

| 知識ソース名 | Default | Nadya | Kentei |
|---|---|---|---|
| 有効回答数 | 25 | 148 | 86 |
| 正答数 | 8 | 90 | 41 |
| 有効回答率 | 0.15 | 0.88 | 0.51 |
| 有効回答正答率 | 0.32 | 0.61 | 0.48 |
| 正答率 | 0.05 | 0.54 | 0.24 |

### 3. 5　考察と追加実験

本節では，3. 4 節の評価実験結果に基づく考察と追加実験について述べる．

#### 3. 5. 1　性能評価

本評価により 3 ～ 4 歳児が身につけるべき概念間の類似性を評価する能力に対して，現時点のコモンセンス手法がどの程度の性能であるかを客観的，定量的に把握可能となった．表 3 に示すように，コモンセンス手法は，Wikipedia 手法や WordNet 手法に比べ有効回答正答率や正答率について良好な結果が得られた．有効回答率についての差異は少ない．コモンセンス手法の有効回答率は 0.95 であり，3 ～ 4 歳児に問われる概念の大部分をカバーできていると言える．一方で回答不能であった 5% については今後知識ベース化を進める必要がある．5% に含まれる概念（Concept）の例としては，“お年寄りに席を譲る”，“お年玉をもらう”など複数の文節で表現が必要な概念（Concept）である．このような概念を扱える知識収集手法や知識表現を検討する必要がある．

各手法の回答傾向について考察する．出題問題には，種類の違いを問う問題と用途の違いを問う問題が多く含まれる．WordNet 手法や Wikipedia 手法は，種類の違いを問う問題にはコモンセンス手法に近いレベルの正答率を示す傾向が見られたが，用途の違いを問う問題にはコモンセンス手法に比べて低い正答率を示す傾向が見られた．また，Wikipedia 手法については対象記事数を増やすほど性能が悪化する傾向にあるため，今回のような常識的な判定問題に対してはノイズとなる記事が多く存在していると推測できる．従って，良い性能を出すためにはノイズ記事の除去が必要となり，コモンセンス手法に比べると安定した性能を出すためのチューニングが難しいと言える．

#### 3. 5. 2　知識量，収集手法と性能の関係

本評価により，知識量と推論性能の関係を把握できるようになった．図 2 に示す通り，知識量の対数オーダーの増加に合わせて有効回答正答率が線形的に向上していく傾向が見られた．また本評価により，知識の収集手法の妥当性や有効性の判断も可能となった．表 5 の結果から，筆者らが行ってきたナージャとなぞなぞや日本人検定を利用して収集した知識は，Default の有効回答正答率を上回っており，有効な知識収集手法であると判断できる．ナージャとなぞなぞと日本人検定を比べると，ナージャとなぞなぞの方がすべての指標において上回っており，より良い知識収集手法であると判断できる．

### 3. 5. 3　追加実験と結果

知識数を増加させると，3. 4. 2 項の傾向に従い性能向上が見られるかを確認するために追加実験を行った．2013 年 8 月～ 2015 年 4 月までに追加で収集した知識を All に追加し，知識ソース Add を生成し，同様の実験を行った．Add において最も性能の高かったテストパラメータにおける結果を表 6 および図 3 に示す．3. 4. 2 項で推定した傾向に従い，性能が向上した．

**表 6　Add の結果（*Cutoff* = 3，*K* = 300，*Prenorm* = 0，*Postnorm* = 1）**

| 知識ソース名 | Add |
|---|---|
| Assertion 数 | 270,415 |
| 問題数 | 168 |
| 有効回答数 | 162 |
| 正答数 | 110 |
| 有効回答率 | 0.96 |
| 有効回答正答率 | 0.68 |
| 正答率 | 0.65 |

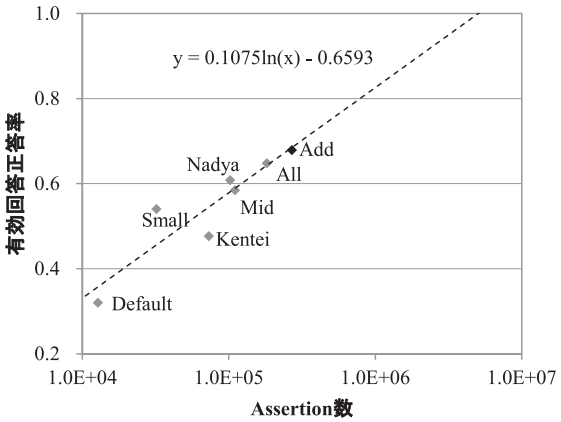


**図 3　Assertion 数と有効回答正答率の関係（Add の結果を追記）**

## 4.　評価実験 2：概念間関連度推定

本章では，概念間関連度推定の性能評価手法の提案および評価結果，考察について述べる．人間の連想語調査結果[18]を正解データと見なして，Spectral Association を用いて算出した連想結果との比較評価を行った．

### 4. 1　評価実験 2 の手順

本節では，概念間関連度推定の評価実験の手順について述べる．

#### 4. 1. 1　正解データの作成

研究[18]のことばの連想についての調査結果における日本の結果を正解データとした．この調査結果は，日本語を母語とする大学生に，ある単語（刺激語）を見せた後に大学生が連想した連想語の一覧および回答人数である．刺激語は，“青”，“家”，“新聞”などの計 30 語である．

例えば刺激語“青”に対する連想語の調査結果は，“空：127 人”，“海：51 人”，“信号：17 人”などである．本評価では，回答人数を関連度の高さと考え，回答人数の降順に順位付けた連想語リスト（1 位：空，2 位：海，3 位：信号，…）を正解データとした．ただし，連想した人数が 1 名の連想語は個人依存性が高いため除去した．

#### 4. 1. 2　前処理

3. 1. 2 項と同様の前処理を行った．

#### 4. 1. 3　Spectral Association による連想語リスト生成

Spectral Association を用いて，各刺激語に対する順位付き連想語リストを生成した．本評価では，計 18,369 件（知識セット Small に含まれる全 Concept＋刺激語）の Concept を連想対象概念とし，刺激語と全連想対象概念との間の関連度の値を Spectral Association で算出し，値の高い順に順位付けを行い，刺激語に対する順位付き連想語リストを生成した．

#### 4. 1. 4　評価指標の導出

30 の刺激語それぞれについて，4. 1. 1 項の正解データの連想語リストと，4. 1. 3 項で生成した連想語リストを比較評価した．評価指標として F 値，AP（Average Precision）を計算した．また，30 の刺激語における総合指標として，マイクロ平均 F 値，MAP（Mean Average Precision）を算出した．

### 4. 2　評価実験システム

評価実験システムの実装は，3. 2 節と同様に ConceptNet4.0 と Spectral Association の実装として Divisi2 を用いた．4. 1. 1 項～ 4. 1. 3 項の方法を用いて連想語リストを生成した．以降，本手法をコモンセンス手法と呼ぶ．

#### 4. 2. 1　コモンセンス知識セット

3. 2. 1 項で示したコモンセンス知識セットにおける All，Mid，Small を用いた．

#### 4. 2. 2　テストパラメータ

結果に影響を与えるテストパラメータとして，SVD の圧縮次元数 $K = \{100, 200, 300, 400, \ldots$ 以降 100 刻み…, 1600$\}$，Concept-Concept マトリクス $\boldsymbol{C}$ の有効データとする行および列方向の最小非 0 成分数 $Cutoff = \{1, 3, 5\}$，$\boldsymbol{C}$ の正規化（各成分を行および列ベクトルのノルムで割る）の有無 $Prenorm = \{1, 0\}$，$\boldsymbol{V'}$ の正規化（各成分を行ベクトルのノルムで割る）の有無 $Postnorm = \{1, 0\}$ の四つのパラメータのすべての組み合わせについて実施した．

### 4. 3　比較手法

比較手法として，Wikipedia 記事テキストを用いた単語共起スコア計算手法を選定した．共起の強さ（共起スコア）の計算には複数手法があるが，本評価では，一般的な cosine と，高度言語情報融合フォーラム（ALAGIN）の単語共起頻度データベース[19]で採用している dice 係数およびディスカウンティングファクター有りの相互情報量[20]（以降 Dpmi）の三つの共起

スコア計算手法を用いた．また，二つの単語の共起を見るデータ単位は，同一記事内の共起と同一文中内の共起の 2 通りで実施した．表 7 に実施した比較手法のテストケース一覧を示す．

**表 7　比較手法のテストケース一覧**

| テストケース名 | 共起度計算手法 | 共起単位 | 対象データ数 |
|---|---|---|---|
| Wikipedia_dice_doc | Dice 係数 | 記事 | 885,147 記事 |
| Wikipedia_dpmi_doc | Dpmi | | |
| Wikipedia_cosine_doc | Cosine | | |
| Wikipedia_dice_sen | Dice 係数 | 文 | 32,751,912 文 |
| Wikipedia_dpmi_sen | Dpmi | | |
| Wikipedia_cosine_sen | Cosine | | |

### 4. 4　評価結果

知識セット All において MAP 値が最大となったテストパラメータセット（$K = 1600$, *Cut-off* $= 5$, *Prenorm* $= 1$, *Postnorm* $= 1$）における結果を示す．図 4 に，全 30 の刺激語におけるマイクロ平均 F 値の結果を示す．横軸は連想語リストの上位何件分を対象とするかを示す．比較手法は，最大の MAP を得たテストケース wikipedia_cosine_sen の結果である．表 8 に MAP の結果を示す．また図 5 に，横軸に Assertion 数（対数軸），縦軸に MAP を取ったグラフを示す．

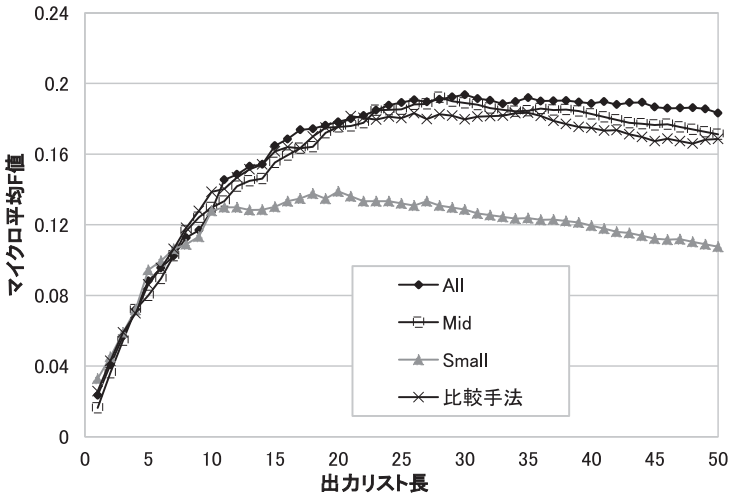


**図 4　マイクロ平均 F 値の結果**

**表 8　MAP の結果**

| 知識ソース名 | Small | Mid | All | 比較手法 |
|---|---|---|---|---|
| MAP | 0.086 | 0.108 | 0.118 | 0.113 |

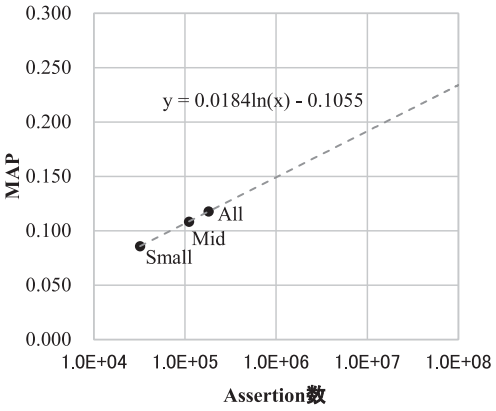


**図 5　Assertion 数と MAP の関係**

### 4. 5　考察と追加実験

本節では，4. 4 節の評価実験結果に基づく考察と追加実験について述べる．

#### 4. 5. 1　性能評価

本評価により概念間関連度推定の性能を客観的，定量的に把握可能となった．表 8 の MAP の結果から，知識セット Small，Mid においては比較手法の性能を下回ったが，知識セット All では比較手法を上回った．現時点でコモンセンス手法には大きな優位性は無いが，知識量の増加による性能向上傾向が見られる．

#### 4. 5. 2　知識量と性能の関係

本評価により，知識量と性能の関係を把握できるようになった．図 5 に示す通り，知識量の対数オーダーの増加に合わせて有効回答正答率が線形的に向上する傾向が見られた．今後，知識量を増やすことで性能の向上が期待できる．そこで，次項に示す知識量を増加させた追加実験を行った．

#### 4. 5. 3　追加実験と結果

3. 5. 3 項と同様に，追加の知識セット Add を利用した追加実験を行った．マイクロ平均 F 値および MAP の結果を図 6，図 7 に示す．図 5 の傾向に従う性能向上が得られた．

## 5.　お　わ　り　に

本論ではコモンセンス推論における概念間類似度推定と概念間関連度推定の性能評価手法の提案と評価結果を報告した．概念間類似度推定性能の評価として，市販の幼児教材の仲間外れ概念探し問題をタスク化する手法を開発した．概念間関連度推定性能の評価として，人間の連想語調査結果を正解データとする手法を開発した．これらの評価手法により，コモンセンス知識収集手法毎の性能把握や，知識量と性能の関係把握が可能となった．また，既存手法との性能を比較し，コモンセンスベースの手法は良好な結果が得られた．今後の予定としては，評価用データ数を増やして交差検定を行いたい．最後に，本研究のチームメンバー，研究や執筆のご指導を頂いた先輩方に感謝する．

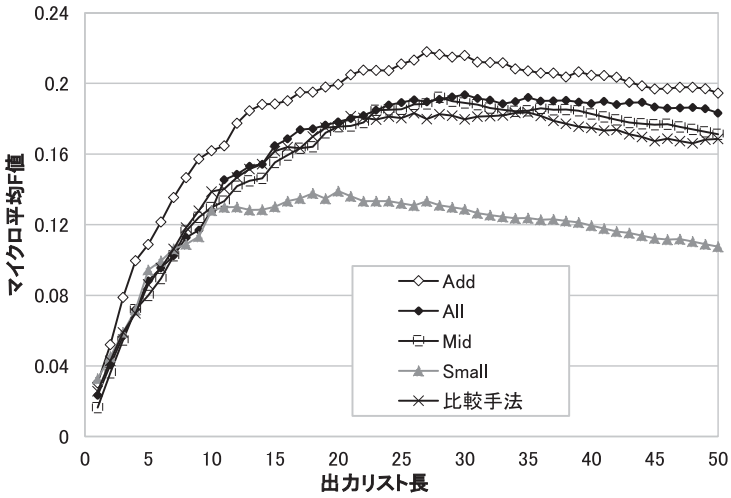


**図 6　マイクロ平均 F 値の結果（知識セット Add の結果を追加）**

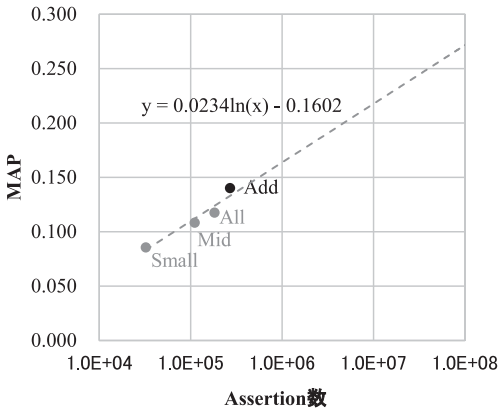


**図 7　Assertion 数と MAP の関係（知識セット Add の結果を追加）**

＊ 1　Mecab の web サイト：http://taku910.github.io/mecab/
＊ 2　NLTK（Natural Language Toolkit）の web サイト：http://www.nltk.org/

**参考文献**　［1］ Havasi, C., Speer, R., Alonso, J.: ConceptNet 3:a flexible, multilingual semantic network for common sense knowledge, In Recent Advances in Natural Language Processing , 2007
［2］ Speer, R., Havasi, C., Lieberman, H.: AnalogySpace: Reducing the dimensionality of common sense knowledge, AAAI,　AAAI Press, 2008
［3］ Havasi, C., Speer, R., Holmgren, J.: Automated Color Selection Using Semantic Knowledge, AAAI Fall Symposium: Commonsense Knowledge, AAAI, 2010
［4］ Lenat, D.: CYC: a large-scale investment in knowledge infrastructure, Communications of the ACM, ACM, Inc., 1995
［5］ Singh, P., Lin, T., Mueller, E., Lim, G. Perkins, T.,Zhu, W.: Open Mind Common Sense: Knowledge acquisition from the general public, Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, Springer Verlag , 2002

［6］ Lieberman, H., Smith, D., Teeters, A.: Common Consensus: A Web-based Game for Collecting Commonsense Goals, Intelligent User Interfaces, ACM, Inc., 2007
［7］ Kuo, Y. L., Lee, J.C., Chiang, K., Wang, R., Shen, E. Chan C., Hsu, J.: Community-based game design: experiments on social games for commonsense data collection, Proceeding KDD-HCOMP'09, ACM, Inc., 2009
［8］ 中原和洋，山田茂雄：日本でのコモンセンス知識獲得を目的としたWebゲームの開発と評価，ユニシス技報，日本ユニシス，Vol.30 No.4 通巻107号，2011年2月
［9］ 中原和洋，コモンセンス知識獲得を目的としたソーシャルゲーム“日本人検定”，ユニシス技報，日本ユニシス，Vol.32 No.4 通巻115号，2013年3月
［10］ ピグリ編集室，伊藤恭監修：能力育成問題集 28 仲間はずれ，株式会社ピグマリオン，2012
［11］ こぐま会教材開発室，ひとりでとっくん 29 仲間はずれ，株式会社 幼児教育実践研究所 こぐま会
［12］ Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense disambiguation, WordNet: An Electronic Lexical Database, MIT Press, 1998
［13］ Wu, Z., Palmer, M.: Verb Semantics and Lexical Selection, in ACL'94, 1994
［14］ Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in IJCAI'95, 1995
［15］ Jiang, J. J., Conrath, D. W.: Semantic similarity based on corpus statistics and lexical taxonomy, in 10th Intl. Conf. Research on Computational Linguistics（ROCLING）, 1997
［16］ Lin, D.: An Information-Theoretic Definition of Similarity, In Proc. of Conf. on Machine Learning, 1998
［17］ Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of the Japanese WordNet, in LREC-2008, 2008
［18］ 小川 早百合，備前 徹，佐々木 倫子，菅原 健介，ことばの意味の文化的背景を記述するための語連想研究 課題番号15520340，科学研究費助成事業データベース，2003-2005，https://kaken.nii.ac.jp/d/p/15520340.ja.html
［19］ 高度言語情報融合フォーラム（ALAGIN），単語共起頻度データベース，https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-5
［20］ Pantel, P., Ravichandran, D., Automatically Labeling Semantic Classes, In Proc. of HLT / NAACL, pp. 321-328, 2004
［21］ 中原 和洋，内田 咲，小林実央，山田茂雄，コモンセンス知識と推論を用いた幼児教材「仲間外れ概念探し」問題への取組みと評価，人工知能学会全国大会論文集，2014
※上記参考文献中のURLは，2015年7月22日時点での存在を確認.

**執筆者紹介** 中 原 和 洋（Kazuhiro Nakahara）
2004年日本ユニシス(株)入社．システム連携技術の主管部門にて各種システム開発プロジェクトに従事．2008年よりR&D部門にて，主に知識処理技術の研究開発に従事.