F-052

二語関係の意味的判別
Meaning Distinction Related to Two Words

奥田 裕也‡ 吉村 枝里子† 土屋 誠司† 渡部 広一†
Yuya Okuda Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，コンピュータが急速に発展し，人間の日常生活や社会活動において必要不可欠なものとなった．しかし，一方で複雑化も進み，コンピュータに関する知識の無い人間にとっては，使い難いものとなっている．そこで，コンピュータと人間のコミュニケーションが可能であれば，誰にとっても使い易くなると考えられる．

人間とのコミュニケーションを可能とするコンピュータの実現のためには，人間と同じ判断能力をコンピュータに持たせることが，必要不可欠だと考えられる．例えば，人間は文書を理解する過程において，単語の表面的な事柄だけでなく，その単語から多くの内容を読み取り，適宜に判断しながら文書を正しく理解しようとしている．

例えば，『世界の宗教の信者数は，キリスト教の 20 億人（33%），イスラム教 13 億人（22%），ヒンドゥー教 9 億人（15%），仏教 3 億 6000 万人（6%），儒教・道教 2 億 3000 万人（4%），無宗教 8 億 5000 万人（14%），その他（6%程度）である。』という文章があればこれらの単語の関係性より，この文章は“宗教”関連の文章であると判断できる．また，“キリスト教”や“仏教”が同じ系列の単語であると理解できる．

単語間の意味の理解が可能になれば，コンピュータに人間と同じ判断能力を持たすことが可能である．

単語間の関係は，多数存在するが，本稿では特に二語関係の種類として，「同意語関係」，「反対語関係」，「包含関係」，「同列関係」，「用途関係」，「行為関係」，「原料関係」の 7 つをとりあげ，これら 7 つの関係を判別するための手法を提案する．この 7 つの関係は，一般的に社会で活動するために必要な基礎知力問題である SPI（Synthetic Personality Inventory）問題の二語関係問題に出題されており，この 7 つと「関係なし」という関係を理解できれば人間と同じぐらいの常識的な判断が行えるといえると考えられる．

2 二語関係の判別

本稿で扱う二語関係の例は，表 1 に示し，以下に定義する．

表 1 二語関係の例

種類	具体例
同意語関係	大雨：豪雨
反対語関係	収入：支出
包含関係	スポーツ：野球
同列関係	キリスト教：仏教
用途関係	ピアノ：演奏
行為関係	医師：診察
原料関係	本：紙
関係なし	袋：サッカー

- 同意語関係…二語が同じ意味を持つ関係
- 反対語関係…二語が反対の意味を持つ関係
- 包含関係…片方の語がもう片方の語に含まれる関係
- 同列関係…同じ系列に属する二語の関係
- 用途関係…片方の語の用途がもう片方の語になる関係
- 行為関係…片方の語が行う用言がもう片方の語になる関係
- 原料関係…片方の語の原材料がもう片方の語になる関係
- 関係なし…関係性がない

図 1 に二語関係判別システムの概念図を示す．

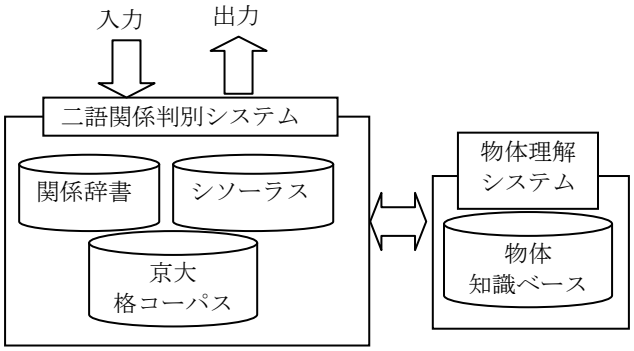

図 1 二語関係判別システムの概念図

本稿で提案するシステムは関係辞書，シソーラス[1]，京大格コーパス[2]，物体理解システム[3]で構成されている．図 1 に二語関係判別システムの概念図を示す．これらを組み合わせて用いることで，より正確な判別が可能である．以下で各技術について，詳細に説明する．

2.1 関係辞書

関係辞書とは，複数の国語辞書より，同義語，反語，上位語，類義語を格納したデータベースであり，同義語は約 20 万 6 千組，反語は約 1 万 7 千組，上位語は約 13 万 7 千組，類義語は約 9 万 2 千組が格納されている．なお同義語とは，同じ意味の単語であり，上位語とは単語の上位概念の語のことである．例えば「動物」や「植物」に対する語の上位語は「生物」という語である．また反語とは，反対の意味の単語のことであり，類義語とは，意味の似た単語である．

2.2 シソーラス

シソーラスとは，一般名詞の意味的用法を表す 2710 個の意味属性(ノード)の上位－下位関係，全体－部分関係が木構造で示されており，約 13 万語が登録されている． 縦の関係を親子関係と呼び，横の関係を兄弟関係と呼ぶ．

図 1 にシソーラスの一部を示す．図 2 より「茶」と「コーヒー」は兄弟関係，「飲物」と「茶」は親子関係である．

†同志社大学理工学部
Faculty of Science and Technology, Doshisha University

‡同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

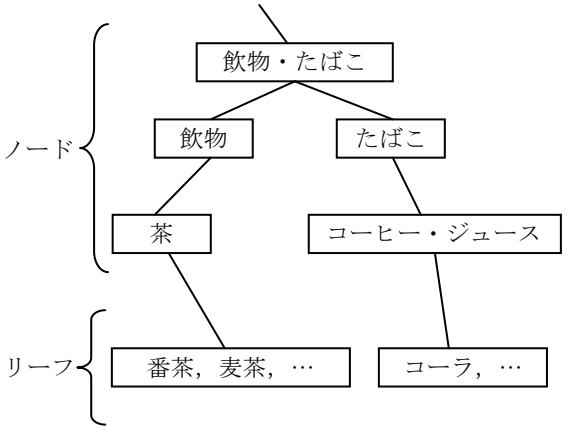

図 2　シソーラスの一部

2.3　京大格コーパス

京大格コーパスとは，用言とその用言に関係する名詞を用法ごとに整理したものである．この京大格コーパスは，Web 上の約 5 億文の日本語テキストから自動的に構築されている．また，京大格コーパスに含まれる用言の数は約 5 万語である．

この京大格コーパスを用いることで，用言からその用言に結びつく名詞，格，頻度を取得できる．また，名詞から用言を取得することも可能である．京大格コーパスの頻度とは Web 上において，その名詞と動詞が出現した回数を指す．

例として，京大格コーパスに名詞「ピアノ」，付属助詞「デ」を入力したときの結果を表 2 に示す．

表 2　「ピアノ」と付属助詞「デ」検索の結果

検索語（用言）	京大格コーパスヒット件数
弾く	263
演奏	81
歌う	47
始まる	42
参加	41

2.4　物体理解システム

物体理解システムには，日常的に使われる全 1400 語の物体を，人間にわかりやすいように物体のイメージを 9 つの観点（材質・大きさ・形状・重さ・長さ・広さ・太さ・厚さ・深さ）で示したデータが格納されている．

表 3 に物体理解システムの例を示す．以下に 9 つの観点の定義を示す．

表 3　物体「机」に対する物体理解システムの出力例

物体	材質	大きさ	形状	重さ
机	木	24	机型	40000
長さ	広さ	太さ	厚さ	深さ
80	-1	-1	5	-1

- 材質…物体の原料を表す．
- 大きさ…64 段階の数値で表され，「23」が人間と同じくらいの大きさを示す．
- 形状…物体の形を表す．形状の種類は 58 種類である．
- 重さ…物体の重さを表現したものであり，単位はグラムである．
- 長さ…物体がもつ長さ（縦，横，高さ）で最も長いものを長さとしている．単位はセンチメートルである．
- 広さ…物体を地面に置いたときに横に広がる面積を指し，主に薄く平らなものや，場所がもつ尺度である．単位は平方メートルである．
- 太さ…物体が持つ最も太い部分の太さを指し，主に丸いものが持つ尺度である．直径を指し，単位はセンチメートルである．
- 厚さ…物体の持つ厚さである．単位はセンチメートルである．
- 深さ…物体の深さを指し，主に空洞を持つものや，地形が持つ尺度である．単位はセンチメートルである．

これらのデータは，人間の常識に則って判断され格納されている．また，その項目に関して量の概念が存在しない場合は「-1」としている．

3　既存システムを用いた二語関係判別システム

3.1 分類方法

図 3 にシステムの具体的な流れを示し，以下に各関係の判別条件を記載する．

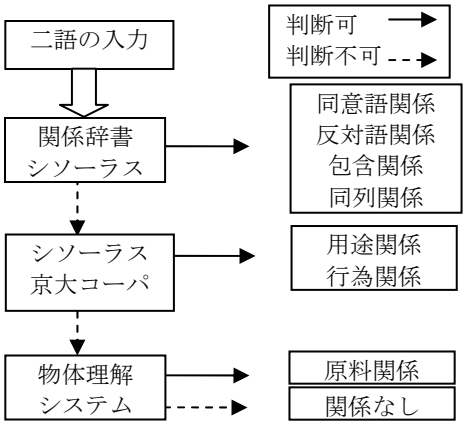

図 3　二語関係判別システムの具体的な流れ

- 同意語関係…関係辞書の同義語に記載されていること．
- 反対語関係…関係辞書の反対語に記載されていること．
- 包含関係…シソーラスで親子関係であること．
- 同列関係…シソーラスで兄弟関係であること．
- 用途関係…シソーラスにより「物」と判断し，京大格コーパスにより付属助詞を「デ」とし用言を取得し，名詞と用言が表記一致すること．
- 行為関係…シソーラスにより「人」と判断し，京大格コーパスにより付属助詞を「ガ」とし用言を取得し，名詞と用言が表記一致すること．
- 原料関係…物体理解システムの“物体”と“材質”の項目と二語が一致すること．
- 関係なし…上記のどの条件にも一致しないこと．

また，関係辞書により「同意語関係」，「反対語関係」と判別できてもシソーラスにより「包含関係」，「同列関係」の判別も行う．なぜなら「同意語関係」かつ「同列関係」という単語の組み合わせもあるからである．その例を表 4 に示す．

表 4　「同意語関係」かつ「同列関係」の具体例

同意語関係かつ同列関係
性急：短気
抹消：削除

3.2　評価結果

各二語関係の単語の組を 40 組（計 320 組）用意し，前節の判別方法を用いて，システムの精度評価を行った．単語の組み合わせは，SPI 問題集[4]から抜粋している．

評価基準として再現率と精度を用いた．それぞれ（3.1），（3.2）式で定義する．

$$\text{再現率} = \frac{\text{正解数}}{40} \qquad (3.1)$$

$$\text{精度} = \frac{\text{正解数}}{\text{ある関係に判断された 数}} \qquad (3.2)$$

表 5 に評価結果を示す．

表 5　評価結果

	再現率	精度
同意語関係	0.35	1.00
反対語関係	0.78	1.00
包含関係	0.88	0.78
同列関係	1.00	0.63
用途関係	0.43	1.00
行為関係	0.48	1.00
原料関係	0.90	0.97
関係なし	0.95	0.40
全体（平均）	0.72	0.85

表 5 より，「同意語関係」，「用途関係」，「行為関係」の再現率が低いことがわかる．再現率の低い原因に，「同意語関係」は語が関係辞書になく，「用途関係」と「行為関係」は，意味が似ているが表記が一致しないことが挙げられる．意味が似ているが表記が一致しないという点を改善するため，関連度計算方式を用いた手法を提案する．

関連度計算方式とは，語と語の意味の関連性を数値で表したものであり，表記一致でなくても語と語の意味を考慮できることから，本システムで有効に機能すると考えられる．

4　関連度計算方式による二語関係判別システム

関連度計算方式を用いた二語関係判別システムの概念図を図 4 に示す．

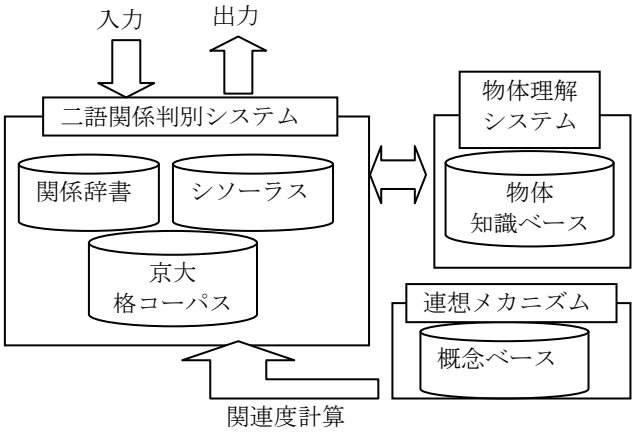

図 4　関連度計算方式を用いた概念図

4.1　関連度計算方式

関連度計算方式[5]とは，概念ベースに定義された語と語の関連の強さを，同義性，類似性に関わらず定量化する手法である．概念ベース[6]とは，複数の国語辞書や新聞等から機械的に構築した語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念ベースには，約 12 万語の概念が収録されている．なお，本稿では概念ベースに登録されていない概念を未定義語と呼ぶ．

概念は，ある語 A を属性 a_i と重み $w_i(>0)$の対の集合として式 4.1 によって定義する．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\} \qquad (4.1)$$

ここで，属性 a_i を概念 A の一次属性と呼ぶ．また，属性 a_i も概念ベースに登録されている 1 つの概念である．従って，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ．

関連度は，0 以上 1 以下の連続的な実数で表され，概念同士の関連が大きいほど関連度は高くなる．関連度は，それぞれの概念を二次属性まで展開し，その重みを利用した計算によって最適な一次属性の組み合わせを求め，それらが一致する属性の重みを評価することで算出する．

4.2　「同意語関係」判別のための関連度計算方式

3.2 節の評価結果では，「同意語関係」の二語が「関係なし」と，判断される問題が多くあった．そこで「同意語関係」では，二語の単語それぞれの同意語を同義語辞書から取得し，全ての組み合わせに対して関連度計算方式によって関連度を調べる．

例えば“作意：意図”という二語の場合，同義語辞書より“作意”と“意図”の同意語を取得する．図 5 にその例を示す．

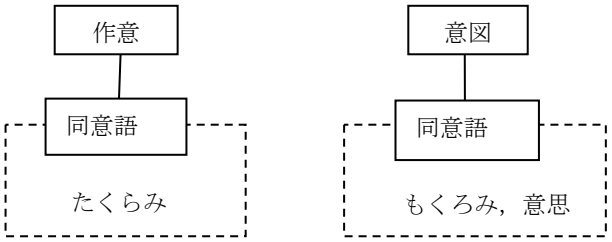

図 5　作意，意図の同意語

「作意」と「意図及び意図の同意語」すべてに関連度計算方式を用いて，語と語の関連度を取得する．例を図 6 に示す．

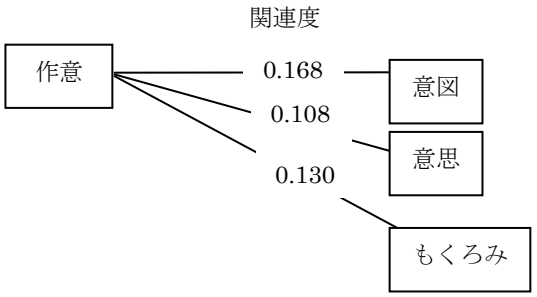

図 6　“作意”の関連度計算方式の結果

同様に「たくらみ」と「意図及び意図の同意語」すべてに関連度計算方式を用いる．そして語と語の関連度を取得する．

これら全ての関連度の平均値を求め，その平均値を“作意：意図”の関連度とする．そしてこの平均値が閾値以上の場合「同意語関係」と判別する．

なお「同意語関係」の関連度計算方式の処理は，「関係なし」と判別される直前に行う．

4.3 「用途関係」「行為関係」判別のための関連度計算方式

3.2 節の評価結果では，表記が一致しないため判別できないという点で再現率が低くなる問題が多くあった．そこで「用途関係」，「行為関係」では京大格コーパスで片方の単語より，用言を 10 語取得し，もう片方の語と 10 語の用言との関連度を調べ，関連度の最大値が閾値以上の場合「用途関係」または「行為関係」と判別する．

例えば“辞書：調査”という二語の組み合わせのとき，京大格コーパスより“辞書”と付属助詞「デ」で用言を取得する．取得した用言は，「調べる，引く，ひく，確認，検索，しらべる，見る，探す，確かめる，いい」である．次に，京大格コーパスにより取得した用言 10 個それぞれと“調査”で関連度計算方式を用いて関連度を求める．表 6 に調査と各単語との関連度を示す．

表 6 “調査”との関連度

検索語 10 語	“調査”との関連度
調べる	0.794
引く	0.029
ひく	0.009
確認	0.024
検索	0.239
しらべる	0.794
見る	0.018
探す	0.161
確かめる	0.017
いい	0.009

表 6 より“調査”と“調べる”の関連度が一番高く，0.794 である．この数値が閾値以上の場合「用途関係」と判別する．また，「行為関係」も同じように，付属助詞「ガ」で取得した検索語と関連度を求め，同様に判別する．

「用途関係」,「行為関係」の関連度計算方式の処理は，京大格コーパスによって「用途関係」,「行為関係」と判別できなかったときにのみ用いる．

4.4 改良後の評価

表 7 に改良後の評価結果，表 8 に新たに判別できるようになった二語を示す．

表 7 改良後の評価

	再現率	精度評価
同意語関係	0.68	0.52
反対語関係	0.78	1.00
包含関係	0.88	0.78
同列関係	1.00	0.80
用途関係	0.68	0.82
行為関係	0.53	1.00
原料関係	0.90	0.97
関係なし	0.74	0.51
全体（平均）	0.77	0.80

表 8 関連度計算方式を用いた成功例

同意語関係	用途関係	行為関係
寄与：貢献	ペン：筆記	役者：演技
作為：意図	ほうき：清掃	裁判官：判決
泰斗：権威	辞書：調査	ピッチャー：投球

表 5，表 7 より「同意語関係」，「用途関係」，「行為関係」の再現率を上げることに成功した．

関連度計算方式を用いることにより，全体的に再現率を上げることに成功したが，精度が下がってしまった．関連度計算方式を用いることにより，「関係なし」の再現率，精度が大きく下がったことが大きく影響した．これは，「関係なし」と判別できていた語も「用途関係」，「行為関係」と判別してしまったからである．

また「用途関係」，「行為関係」の失敗の原因の多くに，京大格コーパスでは複合名詞を扱えないという問題が挙げられる．例えば，“警察”という単語を，京大格コーパスで検索することは可能であるが，“警察官”という単語は京大格コーパスで検索できない．この問題に関しては，新たな手法を考案する必要がある．

5 おわりに

本研究では関係辞書，シソーラス，京大格コーパスなどに関連度計算方式を用いて，二語関係を判別するシ手法を構築した．本手法と会話システムなどに利用することで，コンピュータに人間と同じ判断能力を持たすことができると考えられる．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）21700241）の補助を受けて行った．

［参考文献］

［1］ NTT コミュニケーション科学研究所監修，「日本語語彙体系」，岩波書店，1997

［2］ 河原大輔，黒橋禎夫，“高性能計算環境を用いた Web からの大規模京大格コーパス構築”，情報処理学会自然言語処理研究会，171-12，pp.67-73，2006

［3］ 佐藤祐介，渡部広一，河岡司，“常識的量判断システムの構築-量に関する相対的評価の拡張”，情報科学技術フォーラム FIT2007，E-061，pp.283-286，2007

［4］ 高橋秀雄，“最頻出！SPI パーフェクト問題集’09”，高橋書店，2007

［5］ 井筒大志，渡部広一，河岡司，“概念ベースを用いた連想機能実現のための関連度計算方式”，情報科学技術フォーラム FIT2002，E-39，pp.159-160，2002

［6］ 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41–64，2007