The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

常識的知識を含んだ多属性記述に基づく分散表現型辞書の試作

Trial production of distributed representation dictionary based on multi attributional description including common knowledge

村井　源[*1]
Hajime Murai

[*1] はこだて未来大学
Future University Hakodate

In order to carry out detailed semantic processing for non-grammatical short sentences such as daily conversations, it is necessary to store common sense knowledge. For combining common sense to machine readable dictionary, construction method for manual distributed representation dictionary was proposed based on relationships between words that are derived from traditional ontologies. Moreover, prototype distributed representation dictionary of about 30000 words based on 37 attributes was developed. Manually described distributed representation is readable from both human and machine, and also it has both high scalability and applicability. In the future it is necessary to confirm the objectivity of description by multiple analysts.

1. 背景

自然言語処理など言語を計算機上で処理するためには機械可読な形式で記述された単語の辞書が必要となる. 機械可読な形式の単語辞書には様々な種類があり, ごく基礎的な品詞のみの情報を記したものから, 活用形や表記ゆれに対応したもの, 単語間の意味処理に利用可能な情報にまで踏み込んで記述した意味辞書などがこれまで様々に開発されてきている. 機械可読の意味辞書においては, 同義語, 類義語, 対義語,上位語, 下位語, 部分語(meronym) , 全体語(holonym)などの単語間の関係が人間可読のデータとして構造化・記号化されている場合が多い[Miller 1995]. このような意味辞書は人手によって構築されるか, 機械的に得られたデータを人手で修正して用いることが一般的である. これらの機械可読の意味辞書は存在論にあやかってオントロジーなどと呼ばれている.

一方で近年 Word2vec などのニューラルネットや機械学習を用いた意味表現の自動獲得の手法が盛んになっている[Mikolov 2013]. Word2vec などの自動的に意味表現を獲得する手法においては分散表現と呼ばれるような意味をベクトルとして表現する形式が主流である. Word2vec で得られたベクトルは意味の加法や減法の計算が可能であることから, 大きな注目を集め現在も急速に関連研究が進められている.

現在テキストマイニングなどの分野で主流になっているのは, 大規模なコーパスを対象として特定の単語を探したり数えたりするようなタスクであり, この場合には機械可読辞書は同義語や類義語がグルーピングできれば十分に実用的なレベルと言えよう. このため, 自動的に獲得された分散表現を用いて単語間の距離を計算し, マイニングを行うような実践が多数行われている.

しかし, 例えば会話文を対象として自然言語処理を行うような場合には, テキストは短文でかつ主語や述語, 目的語が頻繁に省略されるなど文法的には不整合であり, また同じ対象を指す単語が逐次言い替えによって変化する(立場による呼称の変化, 繰り返し表現を避けるための言いかえなど). 人間が会話文を理解する場合には種々の常識的知識によって省略された品詞を補ったり, 言い換えられた単語間の関係性を類推したりすることで処理を行うが, これを機械的に実現するためには, 単純な単語間の関係性のみを記した意味辞書のみでは不十分であると考えられる. なお, マイニングの対象となるようなテキストの場合にもこの種の言いかえや省略はもちろん存在しているが, 大規模なデータを収集して統計的に分析することによって, 個々のレベルの問題はノイズとして平滑化され, 全体的な傾向が抽出されるようになっている.

今後, コーパス全体ではなく特定の短いテキストの意味処理や日常会話や物語文などのくだけた口語的なテキストの意味処理を実現するためには, 従来の単語間の関係性に合わせて, 常識的な知識などの人間がテキストを理解する場合に用いる様々な知識が追加できる機械可読の辞書が必要になると考えられる. そのため, 本研究では常識的知識を含んだ分散表現型辞書を構築する手法を検討し, 辞書の試作を行った.

連絡先: h_murai@fun.ac.jp

2. オントロジーと分散表現

従来主流であった単語間の関係や構造を記号等で記述するオントロジー型の意味辞書に関しては, 生成を自動化する試みも様々になされてきているが自動生成した場合には概して精度が低く, 現段階では実用レベルで構築するためには最終的に人間による確認が必要であると言えよう. 例えば世界的に用いられている WordNet [Miller 1995]や, 日本語で頻繁に利用される分類語彙表[国立国語研究所 2004]などは基本的に人間が分類した結果に基づいて構築されている.

人手により構築されたオントロジーは, 精密に語彙の分類や関係性の抽出を行えば, 辞書編纂者によって作成された通常の類語辞書と同等の精度を持たせることが可能である. しかしその反面, 構築に多大な人的コストが必要となると考えられている. 完成したオントロジーの精度を担保するためには複数人によるチェックも必要となり, 一人で構築する場合に比べてさらに数倍のコストがかかる. なお, 複数分析者によるチェックを経ても, 語彙間の関係性や全体的構造は基本的に辞書構築者の解釈に依存するため, 類語辞典と同様に構築者によってカテゴリの種類や構成が異なりうる[山内 2010]. このように単語や意味概念の分類においては唯一絶対の分類法が存在しないという点にも留意する必要がある. つまり特定の分析者によって構築された辞書は, 辞書構築者が想定した意味分類や推論には有用性が

高いが，それ以外の様々に異なる観点から単語の意味を柔軟に比較する処理の実現にも適用可能であることは保証されない．

一方で，Word2vecなどの分散表現はベクトルで構成されるため加法や減法以外にも様々な数学的な変形や分析が可能であり，多様な類似度の計算が実現できる．

しかし，言語の性質上単に共起すれば意味が同じになるわけではなく，品詞や機能語，文法上の役割や能動態・受動態など様々な要因でテキスト中での単語が示す概念は異なりうる．実際に名詞では Word2vec から得られた分散表現は比較的高い精度を得られるものの，動詞での精度は低くなってしまっている[Schwartz 2016]．おそらく,動詞の方が能動態・受動態などの態による違いや慣用句的な複合語による意味の変化の影響を強く受けるためと推測される．また名詞の中でも文脈の類似しやすい反義語の識別に関しても脆弱である点が指摘されており，反義語識別精度の改善には単語の評価極性等を内包したオントロジー[Baccianella 2010] などが必要となるとの研究もある[Dou 2018]．

また，ニューラルネットワークを用いた分散表現で得られる類似度には多様な種類が内包されると考えられるが[Vindula 2017]，ベクトルの各次元が何を意味するかに関しては個別的にいくつかの特徴量を取り上げた考察はあるが，全体的にその構造を理解する方法は発見されていない．そのため，ある類似度の高い語彙の関係が同義語なのか上位・下位語なのかあるいは全体語・部分語なのか判別することは容易ではない．また Word2vec が出力する分散表現のベクトルは元となるコーパスだけでなく，種々のランダムなパラメータによっても構造が変化するため，既存の複数の分散表現を統合させることや，単語概念の詳細化，高度化などのための情報を事後的に追加して発展させるようなことも困難である．構造が単純で応用性が高く，コーパスさえ準備できれば人手によるコストが著しく低いというメリットの反面，ブラックボックス性が高く，追加や更新などの処理には不向きであると言えよう．

他にニューラルネットワーク等を介さずに複数文書中での単語の出現頻度や他の単語との共起頻度をベクトル化する手法も研究されてきている．この場合ベクトルの各次元が持つ意味に関しては人間可読であるが，元となるコーパスや共起単位等によって得られるベクトルが異なり，また単語間の関係性の種類の弁別も容易ではない[秋山 2010]．

3. 多属性記述に基づく分散表現

現状ではオントロジー的記述と分散表現的記述の双方にメリットデメリットがあるが，理想的には両方のメリットを合わせたような表記法が望ましい．これらの統合を図る研究には WordNet 等のカテゴリを用いて，分散表現の自動生成結果の精度を上げる試みなどが行われてきている[Vindula 2017]．また逆に分散表現の結果から語彙の上位・下位関係を抽出する試みなどもある[市瀬 2015]．

本研究では，これらに対して人間と機械双方から可読な辞書形式として，分散表現をオントロジー的知識に基づき人手で構築する方法を採用する[村井 2018]．また，常識的な知識は命題形式で，機能によって分類された動詞辞書に基づき，人手による入力を行う．

具体的にはまず，分散表現の各次元の特徴量を従来のオントロジー的な種々の属性(例えば，物質性,抽象性，生物か否か，知能を持つか否か等)に準じる形で規定してベクトルを作成する．例として鳥と飛行機の場合のオントロジー例とその分散表現への変換例をそれぞれ図 1 と表 1 に示す．図 1 では単純なオントロジーの例として is-a 関係と part-of 関係のみを示した．

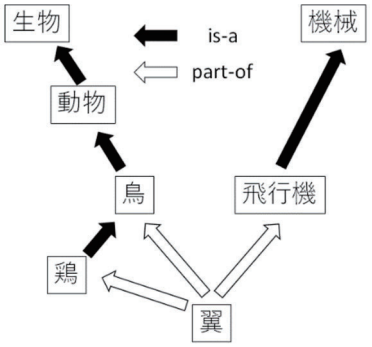

図 1　オントロジー例

表 1　分散表現例

	生物	動物	鳥	鶏	機械	飛行機
生命	1	1	1	1	0	0
能動	0	1	1	1	1	1
翼	0	0	1	1	0	1
飛行	0	0	1	0	0	1

単語の関係性から分散表現を構築する方法は無数に考えられるが，単純化のために語彙の意味的な差異の最低限の弁別ができる次元を設定する方針を採る．図 1 の場合「生物」「動物」「鳥」「鶏」と「機械」「飛行機」の群を弁別するために，「生命」のあるなしを示す次元をまず付与する．次に「生物」と「動物」「鳥」「鶏」を弁別するために「能動」性の有無を示す次元を追加する．さらに，「動物」と「鳥」「鶏」の弁別と，「機械」と「飛行機」の弁別を可能とする「翼」の有無を示す次元を追加する．最後に，「鳥」と「鶏」を弁別するために「飛行」能力を示す次元を追加すると，全単語を異なったベクトルで分散的に表現できる(表 1)．なお，「機械」と「動物」に共通して「能動」性があり，「鳥」と「飛行機」に共通して「飛行」能力があることは図 1 の関係性には含まれず，表 1 で追加されたものである．

表 1 のように分散表現を記述すると，Word2vec と同様に，

「鳥」—「生物」＋「機械」=「飛行機」

のような意味の演算処理も可能となる．

このように明示的に意味を定義された属性を次元として持つ分散表現ベクトルでは，特定の次元を複数組みあわせて(あるいは単独で)用いて距離の演算を行うことで，例えば生物全体の中での類似度や，特定の地域に存在する物の中での類似度，特定の機能の有無に合致する範囲の中での類似度など多種多様な類似度計算が実行可能である．

また各次元の意味は明示的であるため，各語彙が保持する概念の内容と，それらを用いた演算の過程及び結果は，人間の分析者にも理解可能である．そのため分析者が確認することで，誤りの発見や訂正による改善での精度向上が可能である．

表 1 のような記述は，利用者側の必要に応じて適宜新規の次元(属性)を追加して変更することが容易である．オントロジー構築で問題となる本質的な唯一の分類が存在しないという性質も，必要に応じて他の分析者が次元を併記的に追加することで実用的には回避可能である．

上記に加えて本研究ではオントロジーに基づく属性記述に合わせて，各概念に付随する一般的な常識的知識を命題的な表現形式で追加する方式を提案する．命題は分野によって異なるニュアンスを持つが，本稿では内容分析やプロトコル分析の用例に基づき，少なくとも一つの主体と言動的単語を含めた文の意味の最小単位を指すこととする[14].

例えば，「毒物」を概念として定義し記述する場合を考える．人間が道具として利用する「毒物」の機能は「殺す」ことであるが，これは動詞一つで記述可能である．これに合わせて，推理小説などで頻出する毒物を食品や飲み物に混ぜて用いるという使用方法に関する常識的知識も辞書中に含めることを考える．

命題的な表現形式を例えば，「主体:言動:対象 1:対象 2」の形式で記述するとすれば，「人が毒物を食品に入れる」は，「人:入れる:毒物:食品」のように表現できる．命題的表現は形式化された文であるため，このように 1 つの単語で表現できないような複合的な意味構造も表現することが可能である．

会話や物語文中においては特に，死などの強くネガティブなニュアンスを持つ概念は間接的に表現される傾向があり，例えば『毒物を入れられた』と書かれていれば『殺す』という単語がなくてもその可能性を類推する必要がある．単語の持つ一般的機能や常識的知識を動詞や命題の形式で辞書中に含めることで，「毒物」としか書かれていないくても「殺す」が暗示されているという情報を処理可能となる．

頻出命題の検討においては国立国語研究所によるコーパス中の係り受け関係の分析結果を公開している NINJAL-LWP for BCCWJ [国立国語研究所 2012]を参照した．また動詞の記述は機能によって分類した動詞辞書を用いて行った[村井 2017].

4. 属性記述の設計と結果

属性記述の設計にあたっては従来のオントロジー工学での比較的固定的な属性（Type 属性）と状況依存的に変化する役割的属性（Role 属性）の二種類を中心として他に文体的な属性（丁寧語，粗野語など）や各概念カテゴリに固有の属性などを必要に応じて追加して記述した．現在 37 種類の属性を規定している(表 2). Type 属性は上位下位概念や全体・部分関係，形状や色・模様の指定などは他の概念を指定する形で記述した．上位概念や全体・部分関係などは複数の概念が該当する場合があるがそれらはリストとして列挙する形で記述している．これら他の概念を参照する記述は従来のオントロジー表現に対応させることが可能である．それ以外は数値等に置き換え可能な定形表現として記述している．

辞書の作成においては，分類語彙表[国立国語研究所 2004]と日本語 WordNet[Bond 2012]所収の名詞中で現在一般的に用いられているものを中心に分類し,類似の単語をグループ化してそれぞれに属性の値を記述した．現在合計で 33013 単語を 3088 のグループに分類し，3088 グループのそれぞれに 37 種類の属性を設定している．表 2 中の「記述数」は 3088 グループ中で各属性に記述があるグループの数を示している．

実際のデータの例として，学校に所属する(part-of が「学校」)主体的な名詞（人間など言動の主体となる名詞）を抽出した場合の主な属性値記述の例を表 3 に示す．このような絞り込みは例えばテキスト中で場所が学校である場合に詳細不明の人物が出現した場合などに応用可能であると考えられる．表 3 中では分類された名詞グループが 15 種類あり，それぞれ分類名が左端に記されている．試作段階であるため各名詞のグループは比較的大きなまとまりになっているが今後より詳細に分割することも可能である．「代表的上位概念」は複数あるがいずれも人的存在か団体のいずれかの属性を有している．また年齢は「時間量」の箇所に，性別は「性」に数値ではなく文字列として記述されている．「先公」は侮蔑表現だが社会的な立場が上の対象に対する表現であるため「丁寧・粗野」が負，「社会的立場」が正の値になっている．名詞グループと関わる常識的知識は「言動」「主体的言動」に記述されている．例えば，「教える」言動を行う人物は「先生」の属性値を持ついずれかの名詞グループに属している可能性があるというような類推に利用可能である．

表 2 分散表現に用いた属性とその記述数

形式	属性名	内容	記述数
他の概念の指定	上位概念	is-aに相当する概念	3088
	被含有概念	part-ofに相当する概念	1342
	含有概念	part-ofの逆に相当する概念	1008
	構成材料	構成する物質のリスト	1344
	世界設定	特定の場所時間の指定	807
	形状	形を種類で分類	432
	色	色の指定	78
	模様	模様を種類で分類	10
数値等	全体・部分	全体か部分かの指定	480
	内部・外部	内部,境界,外部等の指定	145
	順序	前後等の指定	144
	時間前後	過去,現在,未来の指定	66
	時間指定	絶対時間か相対時間か	523
	時間量	秒,分,時等時間の単位	1477
	空間次元	1,2,3次元のいずれか	2690
	空間量	平均的なサイズの指定	1154
	三相状態	気体,液体,固体	1360
	透明	透明の場合に指定	23
	発光・光沢	発光する場合に指定	20
	位置	位置関係を示す場合に指定	80
	向き	方位等向きを示す場合に指定	74
	抽象	抽象的に用いやすい場合に指定	567
	自然	自然物の場合に指定	962
	人工	人工物の場合に指定	1019
	情報聴覚	音声を示す場合に指定	79
	情報視覚	視覚を示す場合に指定	221
	情報記号	記号を示す場合に指定	204
	情報その他	電気や熱等を示す場合に指定	83
	性	生物等の場合に性別の指定	317
	数	複数である場合の指定	583
	丁寧粗野	丁寧・粗野表現の場合に指定	158
	立場・価値	敬意や価値を含む場合に指定	195
	人称	代名詞等での人称の指定	67
命題	言動・現象	言動として用いうる場合に指定	1060
	主体言動	対象が主体の場合の頻出言動	1551
	客体言動	対象が客体の場合の頻出言動	997
	その他言動	対象がそれ以外での頻出言動	893

表 3 「学校」を被含有概念とする主体的な名詞を抽出した場合の主要な属性値の例

分類名	分類名詞	代表的上位概念						時間量	性	丁寧粗野	社会的立場	言動	主体言動
		人的存在	団体	職種	立場	先生	受講生						
教職員組合	教職員組合,日教組		1					生物_大人					A:教える
校長	校長,校長先生,学校長,学長,学部長,学科長,教頭,教頭先生,園長,学園長	1		1	1	1		生物_大人			100		A:教える,A:叱責,A:戒める
先生	先生,教授,助教授,准教授,助教,教員,教育者,教官,教諭,教師,講師	1		1	1	1		生物_大人			10	教える	A:教える,A:叱責,A:戒める
先公	先公,センコウ	1		1	1	1				-10	10	教える	A:教える,A:叱責,A:戒める
生徒会	生徒会,学級,クラス,学級会		1										A:相談
保護者会	保護者会,父母会,父兄会,PTA		1					生物_大人					A:相談
部活	部活,部活動,同好会,サークル		1										A:遊ぶ,A:スポーツ
受講者	一年生,二年生,三年生,四年生,五年生,六先生,1 年生,2 年生,3 年生,4 年生,5 年生,6 先生,塾生,予備校生,門弟,弟子,兄弟子,弟弟子,門下生,門徒,弟々子,門弟子,門人,門生,見習い,見習,学級委員,教え子,受講者,受講生	1			1		1					教わる	A:教わる,A:行く:[学校]
受験者	受験生,受験者,浪人生,浪人,一浪,二浪,三浪,仮面浪人	1			1		1	生物_青年				試験	A:合格:@[入学試験]->A:所属:C
学生	学生,大学生,学部生,院生,大学院生,修士,一回生,二回生,三回生,四回生,1 回生,2 回生,3 回生,4 回生,学徒	1			1		1	生物_青年				教わる	A:教わる,A:行く:[大学]
女学生	女学生,女子学生,女子大生	1			1		1	生物_青年	女			教わる	A:教わる,A:行く:[大学]
生徒	生徒,中学生,高校生	1			1		1	生物_小児				教わる	A:教わる,A:行く:[高校]
女子高生	女子高生	1			1		1	生物_小児	女			教わる	A:教わる,A:行く:[高校]
男子高生	男子高生	1			1		1	生物_小児	男			教わる	A:教わる,A:行く:[高校]
小学生	小学生,学童	1			1		1	生物_児童				教わる	A:教わる,A:行く:[小学校]

5. 今後の課題

本研究では常識的知識を含む，人と機械双方から可読で追加修正容易な分散表現辞書の構築手法の提案と試作を行った．

構築された辞書は種々のテキストの意味処理に利用可能と考えられるが，従来の日本語意味辞書では難しかった利用方法としては下記などがありうると思われる．

- 特定の条件下での該当単語候補の絞り込み（例：家庭にある道具，会社にいる人物）
- 常識的な知識による省略情報の類推（例：店で客にあいさつする人物がいればおそらく店員か店長）
- 特定の属性に限定した条件下での種々の類似度の計算（例：推理小説の解釈の場合に病院での犯行なら凶器に近い属性を持つものとして薬品やメスがリストアップできる）

構築された辞書の性質や有用性を確認するために今後既存のオントロジーや分散表現辞書との比較が必要であると考えている．また構築された分散表現辞書の記述は一人の分析者によるものであるため，記述の客観性を担保するためには他の分析者による確認と追記も今後の課題である．本研究の成果は内容の評価と修正後に Web 上での公開を予定している．

参考文献

[秋山 2010] 秋山哲史, 内海彰: 概念間の関係に関する単語の意味空間の性質ーコーパス，構築手法，文章単位による影響ー, 認知科学, Vol. 17, No. 1, pp. 110-128, (2010).

[Baccianella 2010] Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, Vol. 10, pp. 2200-2204, (2010).

[Bond 2012] Bond, F., Baldwin, T., Fothergill, R., and Uchimoto, K.: Japanese SemCor: A Sense-tagged Corpus of Japanese, The 6th International Conference of the Global WordNet Association (GWC-2012), (2012).

[Dou 2018] Dou, Z., Wei, W., and Wan, X.: Improving Word Embeddings for Antonym Detection Using Thesauri and SentiWordNet, NLPCC 2018, LNAI 11109, pp. 67–79, (2018).

[市瀬 2015] 市瀬龍太郎, 荒川直哉: 分散表象とオントロジーの関係, 第 29 回人工知能学会全国大会予稿集, 2I4-OS-17a-5, PDF, (2015).

[海保 1993] 海保 博之, 原田 悦子: プロトコル分析入門―発話データから何を読むか, 新曜社, (1993).

[国立国語研究所 2004] 国立国語研究所: 分類語彙表-増補改訂版, 大日本図書, (2004).

[国立国語研究所 2012] 国立国語研究所: NINJAL-LWP for BCCWJ, http://nlb.ninjal.ac.jp/search/, 2019/1/28 参照.

[Mikolov 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, (2013).

[Miller 1995] Miller, G. A.: WordNet: a lexical database for English, Communications of the ACM, Vol. 38, No. 11, pp. 39-41, (1995).

[村井 2018] 村井源: 分散表現型物語機能辞書の提案と試作, 情報処理学会シンポジウムシリーズ, Vol. 2018, No. 1, pp. 47-52, (2018).

[村井 2017] 村井 源: 言動分類による物語機能辞書の汎用化に向けて, 情報処理学会シンポジウムシリーズ, Vol. 2017, No. 2, pp. 225-230, (2017).

[Schwartz 2016] Schwartz, R., Reichart, R., Rappoport, A.: Symmetric Patterns and Coordinations:Fast and Enhanced Representations of Verbs and Adjectives, Proceedings of NAACL-HLT 2016, pp. 499–505, (2016).

[Vindula 2017] Vindula, J., et al.: Deriving a representative vector for ontology classes with instance word vector embeddings, Innovative Computing Technology (INTECH), 2017 Seventh International Conference on. IEEE, pp. 79-84, (2017).

[山内 2010] 山内隆史, 楠見孝: 概念研究からみたオントロジー工学, 認知科学, Vol. 17, No. 1, pp. 54-65, (2010).