統計的意味論に基づく概念的類似度獲得手法の評価

久保田豊久† 若林 啓††

† 筑波大学情報学群知識情報・図書館学類 〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系 〒 305-8550 茨城県つくば市春日 1-2
E-mail: †s1313111@u.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし 近年，分布仮説に基づいた単語の概念的類似度の自動抽出手法が盛んに研究されているが，概念的な類似性は様々な観点から評価することができるものであり，抽出される単語類似度がどのような特性をもつのかは自明ではない．本研究では，word2vec と係り受け構造を考慮した類似度計算手法について，抽出される単語類似度と，人手で作成された既存の意味辞書との一致度を定量的に評価することにより，その特徴について新たな知見を得ることを目指す．

キーワード 係り受け構造，分布仮説，word2vec，統計的意味論

1. はじめに

情報通信技術の発達によって大量の情報を扱うことができるようになったことで，情報検索の技術はますます重要になってきている．とりわけ，単純な文字列の一致だけではなく，意味的な類似性を考慮した検索技術の発展は喫緊の課題であると言える．このことは，キーワードよりも柔軟な入力をクエリとした検索や対話インターフェース [1] の実現において，特に重要な観点になると考えられる．

現在の検索技術では，単語の共起情報に基づいた次元圧縮 [2] や機械学習 [3] ，単語のクラスタリング [4] によって，単語の話題的な類似性を利用する手法が用いられている．しかし，このアプローチでは基本的に単語とクラスタとの関係を扱うことから，直接単語間の意味的な類似性を測る方法として用いることは困難である．

近年では，分布仮説に基づいた単語の概念的類似度の自動抽出手法が盛んに研究されている．分布仮説は，単語の概念的な関係が，文書における単語同士の出現関係によって推定できるという仮説である．これまでには，辞書資源を利用したオントロジーの自動生成手法 [5] や構文情報を利用した言い換えの自動生成手法 [6] などが提案されており，分布仮説による単語の概念的関係の自動抽出の可能性が示唆されている．分布仮説に基づいた概念的な類似度の抽出手法として，word2vec [7] [8] や係り受け関係を用いた類似度計算手法 [9] が提案されている．この 2 種の手法は，文書集合を入力することで，概念的に類似した単語を出力することができるとされている．しかし，概念的な類似性は様々な観点から評価することができるものであり，分布仮説に基づいて抽出した単語類似度が，どのような特性をもつのかは自明ではない．人手で作成された意味辞書においては，単語の類義関係の集合の間の関係を記述した日本語 WordNet [10] や，単語の語彙的な類似性に基づいて単語を分類している語彙分類表 [11] があり，その基準は明確に異なるとされている [12] ．またこれらの意味辞書と分布仮説に基づいて抽出した単語類似度との対応関係は明らかにされていない．

本研究では，word2vec と係り受け構造を考慮した類似度計算手法について，抽出される単語類似度と，人手で作成された既存の意味辞書との一致率を定量的に評価することにより，その特徴について新たな知見を得ることを目的とする．また，日本語の構文的知識を利用している係り受け構造を考慮した類似度計算手法に対して，word2vec との抽出結果の比較を行うことにより，日本語における word2vec の有用性を評価する．

以下に本研究の構成を示す．第 2 章では word2vec を用いた類似度計算手法，係り受け構造を用いた類似度計算手法を示す．第 3 章では，各手法における評価手法を示す．第 4 章では，word2vec を用いた類似度計算手法と係り受け構造を用いた類似度計算手法の比較実験を行い，両者を比較，評価を行い，それぞれの手法の特徴を明らかにする．第 5 章では，本研究のまとめと今後の課題，展望を示す．

2. 分布類似度を用いた類似度計算手法

2.1 word2vec を用いた類似度計算手法

本節では word2vec を用いた類似度計算手法について述べる．word2vec とは，Mikolov ら [8] によって提案された機械学習手法であり，文書データのみを入力として単語間の類似度を計算する．分布仮説に基づいた Neural network model であり，ベクトルによって単語間の意味的な関係を表現することができる．辞書などの事前の言語的知識を利用しないため，言語や語順，固有表現の影響を受けることなく単語の関係を表現することができる．計算に際しては skip-gram-model を用いて計算を行う．skip-gram-model は式 (1) で表現される．T を対象とするコーパスの語数，c を訓練に使用する語数，w を訓練に使用される語とする．$\mathrm{p}(\mathrm{w_{t+j}}|\mathrm{w_t})$ は式 (2) のように定義し，式 (1) を最大化するように v の値を学習する．$\mathrm{v_w}$ を入力された語のベクトル，$\mathrm{v'_w}$ を出力される語のベクトルとする．

$$\frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\sum_{\mathrm{c}\leq\mathrm{j}\leq\mathrm{c},\mathrm{j}\neq 0}\log \mathrm{p}(\mathrm{w_{t+j}}|\mathrm{w_t}) \quad (1)$$

$$\mathrm{p}(\mathrm{w_O}|\mathrm{w_I}) = \frac{\exp(\mathrm{v'^{T}_{w_O}}\mathrm{v_{w_I}})}{\sum_{\mathrm{w}=1}^{\mathrm{w}}\exp(\mathrm{v'^{T}_{w}}\mathrm{v_{w_I}})} \quad (2)$$

式 (2) の計算量を抑えるために Hierarchical Softmax を用いる．Hierarchical Softmax は単語をノードに対応させた 2 分木の木構造をモデルとした式であり，ルートから単語までの各ノードまでのベクトル，内積を計算し，その積を求めるモデルである．この手法により，計算量を単語数の対数時間まで削減することができる．Hierarchical Softmax における単語確率は式 (3) によって定義される．n をルートから対象ノードまでの距離，ch(n) を n の子供ノード，L(w) を w までの距離とする．σ は式 (4) で表現される．また，頻出語の削除のために，サブサンプリングを行う．サブサンプリングとは，式 (5) の確率に従ってコーパスから単語を取り除く手法である．t を閾値とし，f(w) を単語出現頻度とする．

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma([\![n(w,j+1) = ch(n(w,j))]\!] \cdot v_w'^{T} v_{w_I}) \quad (3)$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \quad (4)$$

$$p(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \quad (5)$$

2.2 係り受け構造を用いた類似度計算手法

本節では係り受け構造を用いた類似度計算手法について述べる．日本語においては係り語と受け語は必ず 1 つの係り語に対して 1 つの受け語を持つという性質を持っている．文章を係り受け解析した例を表 1 に示す．係り受け解析では，文章を文節ごとに分け，係りと受けのペアを作成する．係り受けのペアは 1 つ文のうち，最も関係の深いにある文節をペアとして係り受け構造を構築する．

文書例：A さんはフランスパンを食べた.
↓
↓係り受け解析
↓
係り語：A さんは　　受け語：食べた.
係り語：フランスパンを　　受け語：食べた.

表 1 係り受け解析例

係り受け構造を用いた類似度計算手法は，類似した係り語や受け語をもつ単語同士が概念的に類似しているという仮説に基づいて単語間の類似度を求める [9]．この手法ではまず，コーパスを日本語係り受け解析器 CaboCha にかけ，形態素の解析と係り受け構造の解析を行う．今回は名詞の概念的類似度を計算するために，解析結果から一般名詞，固有名詞，代名詞，数詞，接尾名詞，形容動詞語幹，ナイ形容動詞語幹，サ変接続名詞を抽出する．また，係り語と受け語がどのような語で接続されているかを見るために，助格詞も併せて抽出する．抽出した語は表 2 のような形式でファイルに出力される．

経済産業省 の 安井正也官房審議官 が
安井正也官房審議官 が 務めていた
経産省資源エネルギー庁 の 原子力政策課長 を
原子力政策課長 を 務めていた
務めていた ０４年

表 2 文書抽出例

次に出力されたファイルから単語の出現頻度をカウントすると同時に共起語をリスト化する．共起語リスト化の際には，対象単語に対して共起した単語が受け語として共起したのか，係り語として共起したのかを分けてリスト化し，助詞も含めて共起頻度のリスト化を行う．共起頻度のリスト化を行った結果例が表 3 である．U が付いている場合は受け語を意味しており，K が付いている場合には係り語を意味する．

米飯缶詰 K の備蓄
サリン汚染 U 居住地区の
中心地 U ワシントンの
全般的満足度 K の要因
東南アジア歴訪 U 前回の

表 3 係り受け構造を考慮した共起リスト例

共起語をリスト化したファイルから単語同士の類似度を計算し，出力する．相澤ら [9] は類似度の計算に Jaccard 係数を用いる方法の精度が高いことを示していることから，本研究では Jaccard 係数を用いる．計算式は式 (6) で表現される．w_1,w_2 を比較する単語とし，w_1,w_2 の共起語の集合をそれぞれ V_1,V_2 とする．

$$Jaccard(w_1, w_2) = \frac{|v_1 \cap v_2|}{|v_1 \cup v_2|} \quad (6)$$

3. 評価手法

本章では実験における評価手法について述べる．本研究の目的は word2vec と係り受け構造を考慮した類似度計算手法について，人手で作成された意味辞書との一致率を定量的に評価することで，抽出される類似度の質的な特性について知見を得ることである．評価の指標としては，日本語 WordNet および分類語彙表を使用する．

3.1 日本語 WordNet

日本語 WordNet とは，日本語の単語を synset と呼ばれる単位でまとめた概念辞書である．synset とは，同義語関係にある語をセットとしてまとめたものである．それぞれの synset には他の synset との上位関係，下位関係などの関係も記述されている．さらに語は名詞，動詞，形容詞などの品詞によっても分類される．同じ語であっても品詞を名詞として見た場合，形容詞として見た場合，異なる品詞として見た場合には synset の内容は異なるものとなる．これは語が名詞の場合，語は「上位語」,「下位語」,「同義語」等の関係を持つが，形容詞の場合には「関係のある名詞」,「動詞の分詞」という関係を持っているためである．

今回は評価のために新たに,「兄弟語」という語の関係を用意する.「兄弟語」とは同一の「上位語」もしくは「下位語」を共有した語を指す．例としては，図 1 のように評価語の synset の 1 つ上位概念にあたる synset がもつ，全ての下位概念の synset に含まれる語の集合を,「兄弟語」の 1 つの集合とする．さらに，評価語の synset の 1 つ下位概念に対応する synset がもつ，全ての上位概念の synset に含まれる語の集合を，もう 1 つの

「兄弟語」の集合とする．2 つの「兄弟語」の集合を合わせたものが最終的な「兄弟語」の集合となる．

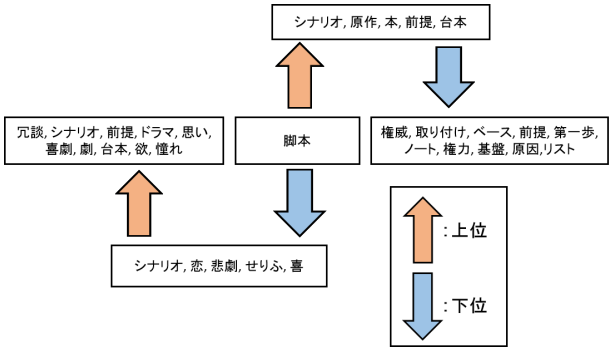

図 1　兄弟語セット例

3.2　分類語彙表

分類語彙表は言葉の意味を分類した表であり，語がそれぞれに持っている関係性を体系づけて表現している．分類語彙表はまず図 2 のように，語全体を「体の類」,「用の類」,「相の類」という，品詞の観点からみた分類に分ける．次に,「抽象的関係」,「人間活動の主体」,「人間活動-精神および行為」,「生産物および用具」,「自然物および自然現象」という意味的な観点 [12] での分類を行う．分類には特定の基準は設けられていないが，人間がこの分類を見たとき，意味的に分かりやすいことを意識して分類されている．意味的分類はさらに細目で分類され，最終的には「類」,「部門」,「中項目」,「分類項目」という 4 層構造によって体系づけがなされる．この分類によって語の意味を集合として分類することができる．

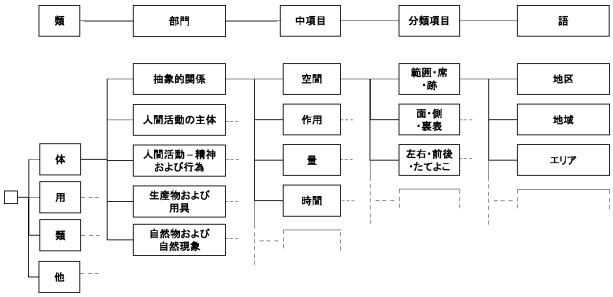

図 2　分類語彙表体系図

3.3　評価手順

今回の評価では，名詞の語を評価対象として評価を行う．日本語 WordNet では「上位語」,「下位語」,「同義語」,「兄弟語」の関係を持つ単語を評価に用い，分類語彙表では最も細分化される組み合わせである,「類」,「部門」,「中項目」,「分類項目」の 4 層全てが同じ組み合わせである単語の集合を評価に用いることとする．

まず，コーパスにおいて単語の出現頻度が 5 回以上の名詞を抽出する．これは出現頻度が極端に小さい場合，分布仮説における統計量の信頼性が損なわれてしまうことによって，正当な評価が不可能になることを防止するためである．

次に，評価する語が日本語 WordNet 及び分類語彙表において存在し，かつ両辞書における正解数が 3 以上の単語を評価候補単語とする．正解数とは，日本語 WordNet において対象単語の「同義語」,「上位語」,「下位語」がそれぞれ 3 単語以上あることとし，分類語彙表において対象単語の「類」,「部門」,「中項目」,「分類項目」の組み合わせが同じである単語が 3 単語以上あることとする．これは正答数が極端に小さい場合，1 つの正解単語の一致が全体の一致率に大きく影響することによって，適切に評価することが難しくなるためである．

最後の条件として，評価される単語の単語同士の類似度は，類似度上位の単語から評価を行うものとし，辞書に存在する単語のみを評価する．評価にあたり，辞書正解数が 3 件であったならば，類似度は上位 3 件のみを評価し，式 (7) のように一致数を辞書正解数で割った値を一致率とする．

$$\text{一致率} = \frac{\text{一致数}}{\text{辞書正解数}} \quad (7)$$

4.　実　　験

4.1　実験方法

word2vec を用いた類似度計算手法と係り受け構造を用いた類似度計算手法の比較を行い，それぞれの手法の特徴を明らかにする．本実験では対象テキストとして，CD―毎日新聞 2012 年版及び 1995 年版の記事本文データを用いる．比較を行う際には，日本語 WordNet，分類語彙表を使用し，3 章に示した一致率を参考に評価を行う．評価の項目に関して，日本語 WordNet を参照する場合,「同義語」の一致率,「同義語+兄弟語」の一致率を実験結果とする．分類語彙表を参照する場合には，評価語と類似度上位にある語の「類」,「部門」,「中項目」,「分類項目」の組み合わせが完全に一致している場合のみ，一致しているとして評価する．評価対象語は 3 章でも述べた条件に合致する単語からランダムに 100 単語を選出する．

4.2　実験結果

表 4 にランダムに選出した 100 語の平均一致率を示す．それぞれの項目を見ると，分類語彙表においては word2vec を用いた類似度計算手法が係り受け構造を用いた類似度計算手法に比べ，一致率が大きいことが確認できる．日本語 WordNet を辞書として使用した場合，同義語のみの評価では word2vec を用いた類似度計算手法，係り受け構造を用いた類似度計算手法共に同程度の一致率を示している．兄弟語と同義語を評価に使用した場合には，係り受け構造を用いた類似度計算手法では殆ど一致率の変動は無いが，word2vec を用いた類似度計算手法では 0.126 から 0.073 へと大きく一致率が落ちていることが確認できる．

使用辞書	分類語彙表		日本語 WordNet			
使用項目	分類項目		同義語		同義語＋兄弟語	
計算手法	係り受け	word2vec	係り受け	word2vec	係り受け	word2vec
平均一致率	0.124247	0.163386	0.123512	0.126252	0.125256	0.0733511

表 4 平均一致率

単語	係り受け	word2vec	差分
任務	0.545454545	0.363636364	0.181818182
速度	0.5	0.333333333	0.166666667
衣服	0.285714286	0.142857143	0.142857143
景観	0.142857143	0	0.142857143
趣旨	0.1875	0.0625	0.125
職業	0	0.125	-0.125
出来事	0.111111111	0.333333333	-0.222222222
名前	0.076923077	0.307692308	-0.230769231
迫力	0.142857143	0.428571429	-0.285714286
長所	0.25	0.625	-0.375

表 5 分類語彙表における一致率

単語	係り受け	word2vec	差分
趣旨	0.24	0.08	0.16
やり方	0.230769231	0.076923077	0.153846154
任務	0.333333333	0.19047619	0.142857143
差	0.285714286	0.142857143	0.142857143
部門	0.291666667	0.166666667	0.125
真ん中	0.125	0	0.125
事情	0.125	0.25	-0.125
文書	0.25	0.375	-0.125
曲	0.125	0.25	-0.125
理由	0.071428571	0.214285714	-0.142857143

表 6 日本語 WordNet における同義語一致率

単語	係り受け	word2vec	差分
差	0.234848485	0.034090909	0.200757576
ホール	0.210526316	0.022556391	0.187969925
仲間	0.221052632	0.047368421	0.173684211
本質	0.269592476	0.097178683	0.172413793
感情	0.298412698	0.149206349	0.149206349
裏付け	0.029411765	0.058823529	-0.029411765
規律	0.029411765	0.058823529	-0.029411765
方針	0.051546392	0.087628866	-0.036082474
書面	0.057471264	0.103448276	-0.045977011
外見	0.007168459	0.0609319	-0.053763441

表 7 日本語 WordNet における同義語＋兄弟語一致率

次に語ごとの個別一致率を見ていく．表 5 は評価辞書に分類語彙表を用いた場合の個別一致率を示した表である．表 6 は評価辞書に日本語 WordNet を用いた場合の同義語の個別一致率を示し，表 7 は同義語セットに兄弟語セットを加えた場合の個別一致率を示している．今回は 100 語評価を行っているが，ここでは主に手法によって一致率の差が大きかった語を表に示している．一致率の差はそれぞれ，(係り受けでの一致率 − word2vec での一致率）で計算している．表 5 に示した，分類語彙表を評価辞書として用いた一致率評価では，word2vec を用いた類似度計算手法の方が係り受け構造を用いた類似度計算手法の一致率を上回った語は 38 語，同じ一致率となった語は 46 語，下回った語は 16 語となった．表 6 では，word2vec を用いた類似度計算手法の方が係り受け構造を用いた類似度計算手法の一致率を上回った語は 22 語，同じ一致率だった語は 50 語，下回った語は 28 語となった．表 7 では，word2vec を用いた類似度計算手法の方が係り受け構造を用いた類似度計算手法の一致率を上回った語は 16 語，同じ一致率だった語は 2 語，下回った語は 82 語となった．

4.3 実験考察

表 5 において，最も一致率の大きかった語である「長所」と係り受け構造を用いた類似度計算手法において一致率 0 であった「職業」について詳しく見る．各手法において類似度上位と判定された語は以下の表 8，表 9 に示す．

word2vec	
語	類似度
特性	0.457273
強み	0.427378
個性	0.422949
特色	0.417858
持ち味	0.416799
利点	0.405771
弱点	0.402594
感性	0.396606
才能	0.371493
信念	0.357058
特徴	0.355441
考え方	0.341768
ユーモア	0.338416
パワー	0.335801
魅力	0.335644
資質	0.329265
底力	0.325476
スタイル	0.322571
生き方	0.321368
メリット	0.31086

係り受け	
語	類似度
功績	0.042307692
個性	0.04136253
心情	0.041237113
過ち	0.040229885
心理	0.040145985
弱点	0.039711191
生き方	0.037848606
強み	0.037647059
才能	0.03654485
使命	0.035502959
面影	0.035353535
手腕	0.034482759
思いやり	0.034090909
唇	0.034013605
感性	0.033112583
効用	0.032967033
心身	0.03286385
腕	0.032036613
キャリア	0.030769231
力量	0.030701754

分類語彙表
正解：8 件
メリット
強み
異色
弱点
特徴
特色
特性
利点

日本語 WordNet
正解：5 件
利点
値
価値
メリット
強み

(a)word2vec (b) 係り受け (c) 分類語彙表 (d) 日本語 WordNet

表 8 分類語彙表における「長所」の正解と類似度上位結果

word2vec	
語	類似度
職種	0.345404
性別	0.291723
経歴	0.249297
氏名	0.236163
教養	0.225204
年齢	0.214841
教員	0.213607
身分	0.206926
形態	0.196045
職	0.193525
里親	0.19269
教材	0.190398
人材	0.188585
キャリア	0.187696
フリーター	0.187623
職場	0.18467
役職	0.183066
部署	0.181591
趣味	0.180271
学科	0.180174

係り受け	
語	類似度
性別	0.072368421
住所	0.067625899
年齢	0.057813911
国籍	0.054325956
氏名	0.051873199
学年	0.048913043
生き方	0.034246575
境遇	0.032663317
番号	0.032200358
体調	0.029644269
名前	0.029214559
人種	0.028985507
心情	0.028735632
症状	0.027522936
口座	0.02739726
居場所	0.027088036
経歴	0.026978417
悩み	0.026392962
運命	0.026033691
性格	0.025614754

分類語彙表
正解：8 件
家事
事務
事業
職
実務
業務
公職
ビジネス

日本語 WordNet
正解：4 件
勤め
職務
職
務め

(a)word2vec (b) 係り受け (c) 分類語彙表 (d) 日本語 WordNet

表 9 分類語彙表における「職業」の正解と類似度上位結果

表 8 を見ると，分類語彙表正解データ 8 件の中で word2vec を用いた類似度計算手法では「特性」,「個性」,「特色」,「利点」,「弱点」の 5 件，係り受け構造を用いた類似度計算手法では「個

性」,「弱点」の 2 件が正解となっている．評価外となってしまうが，word2vec を用いた類似度計算手法では「特徴」,「メリット」という語が上位 20 件までに含まれている．係り受け構造を用いた類似度計算手法でも「強み」という語が評価外で存在している．両者を見比べると，辞書一致率は word2vec を用いた類似度計算手法の方が大きくなっている．分類語彙表が意味的な観点の分類において，人間にわかりやすい分類を行っていることから，word2vec を用いた類似度計算手法の計算結果はより，人間の感覚に近い計算をしていると言える．一方，係り受け構造を用いた類似度計算手法での類似度上位には,「功績」や「手腕」,「キャリア」等の仕事に関係する語が上位になっていることが確認できる．これは「長所」という語が仕事関連の記事に多く登場し，係り受け解析によって類似度上位に出現していると考えることができる．日本語 WordNet の正解データとも比較すると，正解データ 5 件の中で word2vec を用いた類似度計算手法では「強み」の 1 件のみ，係り受け構造を用いた類似度計算手法では評価内において正解は 0 件となっている．しかし，一致率では依然として word2vec を用いた類似度計算手法が優位であった．

次に表 9 を見ると，分類語彙表正解データ 8 件の中で word2vec を用いた類似度計算手法では「職」の 1 件のみ，係り受け構造を用いた類似度計算手法では正解は 0 件となっている．また，係り受け構造を用いた類似度計算手法は「性別」,「住所」,「氏名」等の関係の無い語を類似度上位に挙げている．これは新聞という媒体からコーパスを所得したことで，アンケート結果や犯罪事件の記事等から多くの係り受け関係を抽出してしまったからであると考えられる．一方，word2vec を用いた類似度計算手法は「フリーター」や「教員」といった「職業」という語に関連する語を類似度上位として挙げている．これは，サブサンプリングによって高頻度で出現するが関係の無い語を削除できたことが要因として考えられる．

次に表 6 について，係り受け構造を用いた類似度計算手法において最も一致率が大きな語であった「趣旨」について詳しく見る．各手法において類似度上位と判定された語を表 10 に示す．表 10 を見ると，日本語 WordNet 正解データ 25 件の中で係り受け構造を用いた類似度計算手法では「意義」,「中身」,「内容」,「目的」,「本質」,「役割」の 6 件が正解となり，word2vec を用いた類似度計算手法では「内容」,「目的」の 2 件が正解となっている．分類語彙表の正解データを見ると,「精神」と「理念」の 2 語が類似度上位の正解として加わっている．どちらの辞書で比較を行っても，この語では係り受け構造を用いた類似度計算手法の一致率が大きくなっている.「理念」,「見解」,「考え方」といった両手法の類似度上位に存在する語は考えや目的を表す語である．分類語彙表が日本人の分類感覚を反映した分類になっていることから,「趣旨」という語は「理念」や「見解」といった語に近い語であると言える．つまり，この両手法は語によっては人間の感覚に近い語を類似している語であると評価し，類似度上位に結果として表示することができると考えられる．

次に表 7 について，係り受け構造を用いた類似度計算手法において最も一致率が大きな語であった「差」について詳しく見る．各手法において類似度上位と判定された語を表 11 に示す．表 11 を見ると，日本語 WordNet 同義語+兄弟語正解データは 266 件あり，word2vec を用いた類似度計算手法の一致率は 0.034，係り受け構造を用いた類似度計算手法では一致率は 0.234 であった．同義語のみで一致率を評価した場合には，word2vec を用いた類似度計算手法では 0.142 となり，兄弟語を評価に含めたことで一致率は大きく低下した．これは正解データが増えたことに対して，類似度上位の正解数がほとんど増えなかったことを意味している．一方，係り受け構造を用いた類似度計算手法では同義語のみを正解データとした場合であっても一致率は 0.285 であり，一致率の低下は限定的であった．これは，正解データが増えたと同時に，類似度上位の語が正解として一致数を伸ばし，一致率の低下を抑えたためであると考えられる．

word2vec	
語	類似度
内容	0.427677
考え方	0.326624
立場	0.319737
言い方	0.303602
動機	0.300818
見解	0.298136
理念	0.296453
理由	0.295587
形式	0.289854
意向	0.289407
目的	0.285337
根拠	0.282271
心情	0.275328
観点	0.272921
信念	0.272766
やり方	0.267626
経緯	0.263896
こと	0.257035
方針	0.247879
真実	0.247135
申し入れ	0.245495
責務	0.239591
虚偽	0.234255
方法	0.232745
真相	0.232302

(a)word2vec

係り受け	
語	類似度
理念	0.045822102
意義	0.045333333
中身	0.044189853
内容	0.04345898
あり方	0.036736899
見解	0.036323202
性格	0.036290323
目的	0.035541195
精神	0.035338346
考え方	0.032258065
点	0.031415461
立場	0.030815109
方向	0.030590717
本質	0.030534351
役割	0.030176416
経緯	0.029959514
法	0.029876977
姿勢	0.0293055
ルール	0.028013582
根拠	0.027855153
在り方	0.027798648
仕組み	0.027173913
条件	0.027059774
文書	0.026475038
意向	0.026255182

(b) 係り受け

分類語彙表
正解：16 件
要領
重点
議題
話題
ストーリー
精神
主眼
争点
テーマ
理念
意義
概要
題材
含み
大筋
大綱

(c) 分類語彙表

日本語 WordNet
正解：25 件
心
本質
効用
役割
気
働き
用
中身
目的
意
狙い
メッセージ
テーマ
目当て
要領
内容
課題
物
主題
対象
題
使い方
意義
考え
役目

(d) 日本語 WordNet

表 10 日本語 WordNet における「趣旨」の同義語正解と類似度上位結果

word2vec	
語	類似度
大差	0.507882
点差	0.444878
格差	0.387074
ばらつき	0.350507
アドバンテージ	0.334557
互角	0.322745
ズレ	0.316939
僅差	0.315567
誤差	0.300305
リード	0.298296
折り合い	0.296265
オルフェーヴル	0.29455
違い	0.294232
隔たり	0.293483
引き離す	0.289107
奪え	0.286238
出遅れ	0.285195
キープ	0.283924
猛追	0.27607
劣勢	0.273951

(a)word2vec

係り受け	
語	類似度
力	0.037702504
バランス	0.035897436
格差	0.032490975
距離	0.031376065
気持ち	0.030130566
点	0.028205128
面	0.028119508
レベル	0.027624309
状況	0.02748791
動向	0.026876268
態度	0.026287554
ずれ	0.025990903
実力	0.025408348
考え方	0.025136154
価値	0.024878313
成績	0.024767802
言葉	0.024560661
性格	0.024415584
思い	0.024278392
姿勢	0.024259609

(b) 係り受け

日本語 WordNet
同義語+兄弟語
正解：266 件
遠隔
好評
域
気
活性
間隔
降り
理由
性質
行い
モチーフ
実力
香り
メッセージ
度合い
認め
外見
局面
⋮

日本語 WordNet
同義語
正解：7 件
ずれ
距離
変わり
格差
異なり
開き
溝

(c) 日本語 WordNet

表 11 日本語 WordNet における「差」の正解と類似度上位結果

実験結果全体を見比べて考察を行えば，人手によって構築された既存の意味辞書との一致率は約 1 割程度であることを確認した．

分類語彙表を用いた評価では，word2vec を用いた類似度計算手法の一致率がより大きくなった．word2vec を用いた類似度計算手法は全ての語をベクトルとして利用し，計算を行うが，係り受け構造を用いた類似度計算手法は係り受け関係以外の語は捨て，計算を行う．この違いが一致率の差として結果に現れたと考えられる．つまり，係り受け関係をもつ単語の分布のみが日本語の分類的特徴を表現するという仮定は適切ではなく，語の意味的な観点の分類には係り受け構造以外の語の関係も考慮する必要があると言える．

日本語 WordNet を用いた評価では，同義語のみの評価では両手法共に同程度，兄弟語を含めた場合には，係り受け構造を用いた類似度計算手法の一致率はほぼ変化せず，word2vec を用いた類似度計算手法のみが大きく一致率を下げた．これは辞書が評価する語の増加に対して，正解数が殆ど増加しなかったためである．一方，係り受け構造を用いた類似度計算手法は兄弟語を含めて評価した場合であっても，殆ど一致率が変化しなかったことから，評価する語の増加に比例して，正解数が増加したと言える．このことから，係り受け構造を用いた類似度計算手法では，同義語に加えて兄弟語の抽出に用いることができると言える．兄弟語は必ずしも意味的，文脈的な類似性を持つわけではなく，概念階層において同階層にある語であるという特徴をもつ．この点において，係り受け構造を考慮した類似度計算手法は，word2vec とは異なる傾向があると言える．

また，実験の評価対象では無いが，評価を行うまでの計算時間を考えると，word2vec はコーパスを分かち書きし，学習を行うのみという小さなコストで計算を行うことができる．一方で係り受け構造を考える場合には，CaboCha による係り受け解析が必要となる．よって，日本語 WordNet の同義語のみの評価であれば，計算コストの小さな word2vec を用いた類似度計算手法の方がより有効であると言える．

5. 結　　論

本研究では，word2vec と係り受け構造を考慮した類似度計算手法について，抽出される単語類似度と，人手で作成された既存の意味辞書との一致率を定量的に評価することにより，その特徴について考察を行った．また，日本語の構文的知識を利用している係り受け構造を考慮した類似度計算手法に対して，word2vec との抽出結果の比較を行うことにより，日本語における word2vec の有用性を評価した．

実験では人手によって構築された既存の意味辞書との定量的比較において，両手法共にコーパスの影響を受けたとしても，約 1 割程度を網羅することができることを確認した．word2vec を用いた類似度計算手法では，サブサンプリングによってコーパスによる影響をより軽減し，日本人の分類感覚に近い語が上位として挙がることを，分類語彙表での一致度を通して確認できた．しかし，日本語 WordNet を用いた評価によって，兄弟語の抽出が難しいという特徴も確認した．係り受け構造を用いた類似度計算手法では，日本語 WordNet において word2vec と同等の同義語抽出性能を持ち，兄弟語を上手く抽出することができることを確認した．

今後の展望として，トピックモデルを用いた類似度計算手法といった異なる計算手法との比較を行う，コーパスの量を変化させる，別コーパスを学習させといった方法によって結果にどのような差異が現れるか，検証を行っていく．また今回は名詞のみの評価であったが，動詞や形容詞も追加することで精度の向上か可能であるかを検証し，辞書自動生成やその技術の一端としての利用を検討していく．

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 25280110,25540159）の助成によって行われた．

文　　献

[1] 柴田 雅博, 冨浦 洋一, 西口 友美, “雑談自由対話を実現するための WWW 上の文章からの妥当な候補文選択手法,” 人工知能学会論文誌, Vol. 24, No. 6, pp. 507-519, 2009

[2] 北 研二, “言語モデルの応用,” 確率的言語モデル, pp. 180-185, 2004

[3] Blei, D, Ng, A, Jordan, M, “Latent dirichlet allocation,” Journal of Machine Learning Research 3, pp. 993-1022, 2003

[4] 仁科 朋也, 内海 彰, “単語グループに基づく Web 文書クラスタリング,” 自然言語処理, Vol. 17, No. 4, pp. 24-41, 2010

[5] 鈴木 敏, “辞書からの上位情報抽出とオントロジー自動生成,” 自然言語処理, Vol. 16, No. 1, pp. 102-116, 2009

[6] 乾 健太郎, 藤田 篤, “言い換え技術に関する研究動向,” 自然言語処理, Vol. 11, No. 5, pp. 152-198, 2004

[7] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space, ”arxiv preprint arxiv:1301.3781, 2013

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality, ”arxiv preprint arxiv:1310.4546, 2013

[9] 相澤 彰子, “大規模テキストコーパスを用いた語の類似度計算に関する考察,” 情報処理学会論文誌, Vol. 49, No. 3, pp. 1426-1436 2008

[10] Francis Bond, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto, “Japanese SemCor: A Sense-tagged Corpus of Japanese,” In Proceedings of the 6th International Conference of the Global WordNet Association(GWC), 2012
[11] 国立国語研究所, “分類語彙表-増補改訂版データベース,” http://www.ninjal.ac.jp/archives/goihyo/, 2004
[12] 金田一 春彦, 林 大, 柴田 武, “語と意味,” 日本語百科大辞典, pp. 373-375, 1995